Detecting Heterogeneity in the MPC: A Machine Learning Approach

Master's Thesis

Department of Economics

University of Mannheim

submitted to:
Prof. Krzyzstof Pytka, PhD

submitted by:
Lucas Cruz Fernandez

Student ID: 1544674
Studies: Master of Science Economics

Adress: Burgfriedenstraße 8, 60489 Frankfurt am Main
Phone: 0176/60442229
E-Mail: lcruzfer@mail.uni-mannheim.de

Mannheim, November 28, 2021

1 Introduction

How do households respond to income shocks? And how do their responses differ given their personal characteristics and economic circumstances? These questions are not only at the center of a wide academic debate but also of major importance for policymakers. The former revolves around verifying or negating the main mechanisms of the *Permanent Income Hypothesis* (PIH). Meanwhile, policymakers are interested in improving the efficiency of government transfers. These two sides have sparked many investigations using a wide array of approaches to quantify households' responses to income shocks - the *Marginal Propensity to Consume* (MPC).

At the center of macroeconomics since Keynes put it forward as a main driver in his General Economic Theory, the MPC quantifies the amount households will spend on consumption of each dollar they receive from an income shock. Research has long focused on testing whether the MPC out of income shocks is zero and, thus, in line with the PIH, but more recently the focus has shifted. Most studies still support the notion of an average zero MPC, but evidence suggests that for some groups, the response is significantly different.

Empirical research related to this heterogeneity in the MPC uses various settings to identify income shocks. Following Parker et al. (2013) and Misra and Surico (2014), we exploit the 2008 tax rebate in the USA to estimate households' MPC through data collected by the *Consumer Expenditure Survey* (CEX). Similar to these two prior studies, we are able to use the rich information on consumption the CEX provides to identify heterogeneities in the overall MPC as well as to analyze the categories of consumption goods most affected by the rebate. In contrast to previous work, our econometric approach sets us apart as it is better suited to detect heterogeneities compared to any contribution we are aware of.

It is important to highlight that the Economic Stimulus Act in 2008 was signed into law by President Bush in February 2008. The tax rebate payments, which were part of this policy, started in April of the same year and are therefore an anticipated income shock. Also, the tax rebate was disbursed to USA taxpayers during a time of national and global economic downturn. Many households receiving the stimulus might have been in economic turmoil when receiving the payment and actually spend it to cover regular expenses that they otherwise would not have been able to cover (e.g., rent, utilities, or other necessities of daily life).

The fine-grained consumption data of the CEX allows us to identify what kind of goods households consumed and what they spent their stimulus money on. As Kaplan and Violante (2014) note, the tax stimulus is anticipated and is subject to these special circumstances. Therefore, one might also speak of our estimated coefficients as a 'Propensity to consume the rebate' or 'rebate coefficient,' which is not necessarily equivalent to households' overall MPC. While a government stimulus program might not be perfectly appropriate to verify theoretical models concerned with the MPC, providing evidence on their effect on individuals is of major importance for future policymaking. When economic relief is urgent, non-targeted stimuli can present a viable option; targeted transfers can play a major role in many policy settings. Thus, understanding which households actually use government transfers for consumption and what these households spend them on is an important part of efficient policymaking. Additionally, quantifying the effectiveness of untargeted transfers is necessary to assess whether they are actually helping to boost the economy. Aggregate estimates of the MPC suggest that this is not the case, but taking a closer look and adjusting for household characteristics reveals heterogeneities and effectiveness of these transfers. Moreover, Parker et al. (2013) emphasize that some rebates were reported to be received outside of the disbursal window, which suggests that the income shocks might not have been anticipated and only noticed after their arrival.

We use the *Double Machine Learning Framework* (DML) developed by Chernozhukov et al. (2017) to estimate individual-level point estimates of the MPC as well as standard errors for each household. This enables us to run hypothesis testsfor each household whether their MPC is significantly different from zero. DML allows us to estimate the *Conditional Average Treatment Effect* (CATE) of the tax rebate on changes in consumption. Thanks to the semi-parametric nature of the DML framework, we can use reliable machine learning estimators to control for any confounding factors without having to define their relationship with the outcome. Moreover, we present an estimator that retrieves the CATE without assuming a specified relationship between variables we condition on and the CATE itself.

Our results underline the heterogeneity of the MPC out of the tax stimulus documented in Parker et al. and Misra and Surico. We find that a large mass of households shows no significant reaction upon receiving the stimulus payment, whereas a smaller fraction of households shows strong and significant reactions above an estimated MPC of 0.5. Our analysis suggests that liquidity is the main driver of MPC heterogeneities and that low liquidity households are the ones reacting the most.

Contrary to prior work, our estimated CATE does not rely on specifying subsets of the data across which we assume heterogeneity to exist. We employ modern methods to quantify the effects of single variables on the estimated MPCs to understand the role of these characteristics. Our non-linear estimators suggest that the heterogeneities presented in other work do not capture the full picture.. Next to our contribution to the MPC literature by providing an empirically more robust analysis, we also see our contribution in introducing modern and flexible estimation approaches to the macroeconomic literature. Frameworks such as the DML offer a gateway to new methods and identifications in the macroeconomic literature. We stress the importance of further research into the theoret-

ical and applied nature of these procedures and their usage in more settings.
The rest of the paper is structured as follows: Section 2 summarizes the theoretical and empirical literature on MPC heterogeneity. Section 3 discusses the data source and challenges connected with it. The empirical methodology we use is described in Section 4, while Section 5 presents the identification and estimation results of the MPC. We further investigate sources of heterogeneity in responses in Section 6. Section **??** concludes.

2 Main channels and literature

There is a vast literature on the marginal propensity to consume and the many factors that potentially play a role in its heterogeneity across households.[1] Over time, one channel has been identified as one of the key drivers, which is the access to liquid assets. In this section, we want to briefly outline how liquidity drives the MPC of households and discuss the more recent empirical literature related to our study.
The role of liquidity is linked to the nature of the income shock and borrowing constraints. Within a calibrated life-cycle model, Kaplan and Violante (2009) show that households have no access to credit markets, and can therefore not borrow, react substantially more to transitory income shocks than households without such constraints. In general, if a positive income shock is anticipated, households that are already close to or at their borrowing constraint cannot borrow new funds to smooth consumption in anticipation of higher future income. Thus, once the shock is realized, we will observe an increase in consumption. On the other hand, saving is always possible for any household, and hence we will not see a reaction once the shock is realized in case of a negative anticipated shock. Thus, more liquid households react less to a positive anticipated shock in comparison with liquidity constrained ones. In contrast, in case of an unanticipated shock, we expect the opposite. Think of an agent that is temporarily out of work and has no liquid wealth at their disposal. In case of a negative shock, the agent is forced to adjust their consumption behavior downward. Meanwhile, a positive shock will always be saved and stretched over future periods, no matter the level of households' liquidity. Bunn et al. (2018) document this asymmetry in reactions to positive and negative shocks using British data.
The applied empirical literature investigating the Marginal Propensity to Consume can be categorized into two strands. The first uses data on households' expenditures and observed income shocks; the second relies on surveys that ask respondents directly for their MPC out of hypothetical or experienced income shocks. Parker and Souleles (2019) coin these approaches the *revealed preferences* and the *reported preferences* approach, respectively. Our contribution firmly sits within the revealed preferences part of the literature.

[1]A benchmark literature review of the literature up to their publication can be found in Jappelli and Pistaferri (2010). For an elaborate overview of the more recent trends consult **XXX**

One common approach to identify income shocks is to look at lottery winners. The odds of winning the lottery are so low that a win can be interpreted as an unanticipated income shock. Studies mostly use state lotteries which have a wide range of small amounts that can be won.[2] For example, Fagereng et al. (2020) use Norwegian administrative panel data and find that households winning the lottery spend almost half of their win within one year and 90% after five years. Moreover, the authors report that liquidity and age are the only variables correlated with the MPC after controlling for confounders. However, correlations are a weak measure to assess drivers of the MPC as we cannot assess which of the variables is the driving force. In a similar vein, Golosov et al. (2020) construct a dataset of lottery winners for state lotteries in the USA. They report an average MPC of 60 cents out of each dollar won (**what timing?**). Supporting the liquidity channel, they find that the highest quartile of the liquidity distribution spends only 49 cents while the lowest quartile spends almost 80 cent of each dollar they win in the lottery. However, these two lottery-based approaches suffer from the drawback that they do not measure consumption directly. Instead, they have to either construct consumption out of households' balance sheet data (Fagereng et al., 2018) or model consumption as a function of their observed variables (Golosov et al., 2019).

Fuster et al. (2020) provide evidence that households who show strong reactions to unanticipated income shocks show no reaction to news about future gains. Additionally, their approach reveals an intensive and extensive margin of the MPC. Mixing the *revealed* and *reported preference* approaches they show that as the size of the windfalls gains increases, more households report that they would increase their spending. However, among households that reveal a significant reaction to shocks, the reaction is actually declining with the size of the shock. Overall, the extensive margin effect dominates in that more households are spending a significant amount of payments.

Recent examples of the reported preferences approach are Bunn et al. (2018) and Christelis et al. (2019). The former use data from the *Bank of England* (BoE) to estimate the MPC of British households to income shocks. In the BoE survey, participants are asked about past income shocks they experienced and how they reacted. Their results further support the liquidity channel as the important driver behind MPC heterogeneity. **here 1 (!) sentence on their heterogeneity results** In a theoretical exercise, they show that a model with occasionally binding borrowing constraints can replicate their results.Christelis et al. (2019) use Dutch data where they find an average MPC of 15% to 25%. Their results reveal strong heterogeneity in responses. Of the households in their sample, 40% react the same to positive and negative shocks, while another 40% respond asymmetrically. The latter suggests a strong role of liquidity in decision-making making

[2]Using lottery winners who win hundreds of thousands or even millions of dollars would be fruitless since the size of the shock is unreasonably large and probably changes the complete underlying choice-set of households. Additionally, sample sizes would be very small.

how to use income shocks. The remaining 20% of households reveal asymmetric behavior in the opposite direction, which the authors connect to other behavioral models and the lack of financial sophistication. This is in line with Parker (2017), who also finds a strong relationship between MPC and sophistication as well as other personal traits. He exploits the *Nielsen Consumer Panel* to investigate the 2008 tax stimulus. The *Nielsen Consumer Panel* provides a large set of questions about how households used the rebate received through the tax stimulus and on their personal preferences. Parker estimates that the MPC out of the rebate is 1.5% of the stimulus received within one week of receipt and 3.5% within the first four weeks. Based on his findings, Parker suggests that the relationship between liquidity and MPC is not situational but rather dependent on the persistence of low liquidity. I.e., households that only have low liquidity situationally do not respond differently than other households.

The already mentioned Parker and Souleles (2019) evaluate the two different approaches of the applied literature. They summarize that households that self-report a larger propensity to spend income shocks indeed have larger estimates using the revealed preferences approach. Interestingly they show that self-reported MPCs do not vary conditional on liquidity, the channel which is supported the most by existing studies.

Lastly, we want to discuss the two studies most closely related to this paper: Parker et al. (2013) and Misra and Surico (2014). The former collaborated with the Bureau of Labor Statistics (BLS) to add questions about the 2008 tax rebate to the CEX in 2008 and 2009. They estimate the average response of households using OLS and 2SLS, where they instrument the rebate amount with a dummy signaling rebate receipt. Their motivation behind the latter will be subject to discussion when we present our identification strategy in section 5.1. Estimating various specifications, Parker et al. find MPCs out of the tax rebate that range between 12% and 30% when considering changes in non-durable consumption. Taking into account all expenditure categories reported in the CEX, they even find responses between 50% to 90%. These higher estimates can be traced back to a small set of households making large purchases in the new vehicle category - a phenomenon that is also documented in Misra and Surico and our own results.

To investigate heterogeneity, Parker et al. follow an approach common to the literature. They define thresholds along the distribution of liquidity and create dummy variables that signal to which group observation i belongs. In their case, they set the cutoff points along the liquidity distribution to cut it into terciles such that each group contains the same amount of households receiving the rebate in a given quarter. Parker et al. report that indeed lower liquidity households show a stronger response to receiving the tax rebate across all their specifications. However, as Misra and Surico point out as well, they only rely on the economic significance but cannot reject the null that the point estimates across groups are significantly different from each other.

This approach is similar to another one often used in the applied literature, which is splitting the sample into subsamples and estimating the MPC within each of these. Both these approaches have a severe drawback when it comes to detecting heterogenous patterns. For example, Parker et al.'s approach will only reveal whether whether heterogeneity exists between the three terciles, but any patterns within these terciles are ruled out by construction. In that manner, our estimation approach is superior as we are able to find heterogeneous patterns without pre-defining any sub-groups.
Meanwhile, Misra and Surico (henceforth MS) use quantile regression on the same data to estimate the conditional distribution of the marginal propensity to consume. They find a distribution across all consumption categories that supports the notion by Kaplan and Violante (2014) that a large amount of households shows no significant response, while a substantial share has an MPC of around 0.5, and some households react even stronger. In fact, the lower and the upper end of the consumption change distribution are reacting significantly differently from zero, where the reaction is increasing along with the distribution. These findings are present across all consumption categories they investigate. They then check how specific variables - such as liquidity or home-ownership rates - are distributed across the distribution of consumption change. They show that in areas where households show significant reactions to tax rebate receipt - at the lower and upper end of the distribution - the median income is higher than in the center part of the distribution. They conclude that high-income households have either strong positive or zero reactions, while low-income households show a consistently positive reaction of 10% to 40%. This explains previously contradicting findings by Sahm, Shapiro, and Slemrod (2010) and others, who provide evidence on high-income households having the largest MPC, and Johnson, Parker, and Souleles (2006) and Parker et al. (2013), who find that low income is associated with higher MPCs. However, these findings are only based on correlations that do not necessarily imply a causal relationship between these factors.
Similarly, Misra and Surico's results must be taken with a grain of salt, too. They ignore a fundamental assumption of quantile regression, which is necessary to interpret their estimates as the MPC and draw the conclusions they present. The assumption in question is the *rank-invariance*, or *rank-preservation*, assumption. We will not discuss the inner workings here, but let us briefly lay out how MS interpretation relies on this restrictive assumption. The coefficient in a quantile regression when estimating the τ^{th} quantile of the outcome signals how much a one-unit change affects this quantile of the outcome's distribution - in our case, consumption change. However, individuals in this quantile before and after treatment need not be the same. Actually, we would exactly expect the opposite if some individuals react strongly to receiving the rebate and others do not. If, for example, individuals previously did not change their consumption much but now react strongly, they are part of a different quantile than before, and the coeffi-

cient only reveals how much the τ^{th} quantile changes - e.g., because households reacting strongly move out of it. Therefore, their coefficients do not reveal the MPC as we do not look at individuals but only at the distribution. Assuming rank-invariance implies that the rank in the distribution before and after treatment stays the same, and we thus compare the same individuals within quantiles. As we have laid out already, this is not reasonable to assume in our setting and is actually quite counterintuitive. Keeping this issue in mind, Misra and Surico's results can still be helpful when we are interested to understand the distributional effects of the tax rebate. In that light, their results provide evidence for a wider dispersion of the distribution of consumption change after the tax stimulus program.

3 Data

We use data collected by the Consumer Expenditure Survey (CEX) that is administered by the Bureau of Labor Statistics. Its main purpose is to provide information on the consumption preferences of US households to adjust the goods basket used to calculate the *Consumer Price Index* and other inflation measures (BLS, 2021). In an effort to understand the effects of the 2008 tax stimulus, Parker et al. (2013) expanded the questionnaire between June 2008 and March 2009. Due to its original purpose, the CEX provides a finely-grained set of information on the type of goods households consume. This enables us to analyze the kind of goods households with a non-zero MPC spend their rebate on. In the following, we briefly outline the stimulus program and describe the CEX data.

3.1 The 2008 Tax Stimulus Program

Due to the global financial crisis and the subsequent recession, the United States government passed the Economic Stimulus Act (ESA) in February 2008. With projected costs of more than 150 billion USD, it was the largest relief program passed in the history of the USA up to this point. While not a focus of this study, the EAS also enacted other steps, taking up roughly one third of the program to provide economic relief.
The stimulus was paid out to any household that filed for income taxes. Households that fell beneath the minimum amount of income required to have to file for federal income taxes had to file for taxes anyway ss long as their annual income was at least 3,000 USD they were eligible for the minimum stimulus payment of 300 USD. Eligible households received their net tax liability as their rebate; however, the payments were bounded by a minimum of 300 and a maximum of 600 USD. For couples filing jointly, the limits were 600 and 1,200 USD, respectively. Parents of children under the age of 17 received an additional 300 USD per child. Additionally, the rebate was capped for high-income households. The

Figure 1: CEX quarterly rotation procedure

Interview year and month		Interview set 1	2	3	4
2015	APR	a			
	MAY	b			
	JUN	c			
	JUL	d	a		
	AUG	e	b		
	SEPT	f	c		
	OCT		d	a	
	NOV		e	b	
	DEC		f	c	
2016	JAN			d	a
	FEB			e	b
	MAR			f	c
	APR				d
	MAY				e
	JUN				f
	JUL				
	AUG				
	SEPT				

Columns show number of interview and a letter signals a specific household. Source: `https://www.bls.gov/opub/hom/cex/data.htm`

rebate was reduced by 5% of the amount that the reported income exceeded 75,000 USD (150,000 USD for couples).

3.2 Consumer Expenditure Survey

The CEX is a representative survey in the USA interviewing households about their consumption patterns on a quarterly basis. Once a household is selected to participate, they are interviewed a total of five times. The first interview is a baseline interview during which general household characteristics, employment-related variables, and their stock of non-durable goods are documented. Over the course of one year, households are then interviewed every quarter and asked about their expenditures over the period since the last interview. After this interview, the household is rotated out of the CEX and replaced with a new one one rendering our dataset a rotating panel. Figure 1 is taken from the CEX website and illustrates this procedure. Note that a household is defined as a Consumer Unit (CU), which can represent either the number of blood or legally related persons living in one household (e.g., foster children), a single individual - even if living with other people as long as the individual is financially independent - or unrelated people who are pooling their income. All information about a Consumer Units' members is collected regarding their relationship to the reference person. This person is defined as the one named when asked who rents or owns the home. For personal traits such as age, we follow the convention by Parker et al. (2013) and take the average of the characteristic of all CU members.

It is important to highlight the limitations set by the usage of CEX data. As mentioned, the main objective of the CEX is to assess what goods the average household consumes to create the goods basket for inflation measurements. This focus results in a lack of interest in dense documentation of household characteristics and income-related variables. For example, the lack of asking for liquidity-related measures in each quarter prevents us from controlling for changes in liquidity, but we can only control for households' overall self-reported levels of liquidity and the variables collected are only crude measures for liquidity (Parker et al., 2013).

While this is a disadvantage in comparison with other data sources, the CEX's richness in information on consumption behavior is unmatched. Keeping in mind the risk of measurement error through the self-reported consumption measurement, the CEX enables us to analyze not only the MPC for overall consumption but to dissect it and see which goods drive responses and heterogeneity seen in higher level estimates.

4 Methodology

To estimate the causal effect of tax rebate receipt on changes in consumption, we use the *Double Machine Learning* framework developed by Chernozhukov et al. (2017). This new kind of estimation approach allows to efficiently estimate semi-parametric models of treatment effects using Machine Learning methods. The semi-parametric approach we follow has the major advantage that it does not restrict the effect of confounders on the outcome to a specific functional form. Moreover, specific DML estimators enable us to estimate heterogeneity given observables without defining in which form the observable affects the treatment effect. Past contributions that were looking into heterogeneity had to rely on choosing the correct interactions with observables. DML estimators open the door to combining powerful machine learning algorithms with causal inference, which can detect interactions and non-linearities without having to define them beforehand. Meanwhile, its implementation procedure deals with common biases arising in more naive estimation procedures that employ Machine Learning methods.

From a more theoretical perspective, the DML estimator yields very efficient properties when it comes to its asymptotic behavior. Chernozhukov et al. (2017) are able to prove $\sqrt{n}$ consistency of the estimator, a rate of convergence not achieved in other nonparametric approaches. However, we will not further elaborate on these details and refer the reader to Chernozhukov et al. (2017) for a more technical discussion. Instead, we focus on the general idea behind the DML framework and the different estimation methods we use in our analysis.

4.1 Setup

We start with considering a *Partially Linear Model* of treatment and outcome

$$Y_{it} = \theta(X_{it})D_{it} + g(X_{it}, W_{it}) + \epsilon_{it} \tag{1}$$

$$D_{it} = h(X_{it}, W_{it}) + u_{it}, \tag{2}$$

where Y_{it} is the outcome, D_{it} is the treatment and X_{it} and W_{it} are observable variables. We distinct between simple confounders W_{it} which affect the outcome and also potentially the treatment and X_{it}, which additionally are considered to impact the treatment effect of D_{it} on Y_{it}. Further, we assume that $E[\epsilon_{it}|X_{it}, W_{it}] = 0$ and $E[u_{it}|X_{it}, W_{it}] = 0$.
We are interested in $\theta(X)$, the *Conditional Average Treatment Effect* (CATE). In Rubin's (XXXX) potential outcomes framework (**citation missing**) it is defined as

$$\theta(X) = E[Y_1 - Y_0|X = x]$$

where Y_d is the outcome when treatment is $D = d$. In our setting, where treatment is not binary, $\theta(X)$ represents the marginal CATE

$$\theta(X) = E\left[\frac{\delta Y(d)}{\delta d}\middle| X = x\right].$$

The marginal CATE measures how much a marginal increase in the continuous treatment changes the outcome for individuals that have a set of characteristics $X = x$. Note that in our setting, we assume that the CATE is linear in treatment, i.e., the treatment effect is independent of the size of the treatment. The task is now to find an appropriate estimator $\theta(X_{it})$.

4.2 Regularization bias and how to get rid of it - alternative title: A quest to avoid biases

As Chernozhukov et al. (2017) point out, we could come up with some seemingly straightforward approach to estimate the PLM using machine learning methods. For example, approximating the function $g(X, W)$ with a high polynomial and using a Lasso regression for regularization or using a combination of random forests for predicting $g(X, W)$ and then an OLS regression to find $\theta(X)$. However, any machine learning-based approach that follows this notion will suffer from a bias. To avoid overfitting and the resulting large variance of the estimator, machine learning methods deliberately induce a bias into their predictions. This bias does not vanish asymptotically, leading to inconsistent results.[3]

[3] See Appendix X.X (or only the paper?).

However, we can deal with this regularization bias. For this, we define

$$E[Y_{it}|X_{it}, W_{it}] \equiv f(X_{it}, W_{it}) \tag{3}$$

$$E[D_{it}|X_{it}, W_{it}] \equiv h(X_{it}, W_{it}) \tag{4}$$

where (4) follows from (2).[4]. It is straightforward to estimate these conditional means using any machine learning method of choice, which is the first stage of the DML framework. Using (3), (4) and the PLM defined above, we can find

$$Y_{it} - f(X_{it}, W_{it}) = \theta(X_{it})(D_{it} - h(X_{it}, W_{it})) + \epsilon_{it}. \tag{5}$$

Subtracting the conditional means from Y and D is known as orthogonalization and removes the impact of X and W, respectively. The residuals then only contain variation that does not stem from any of the confounders. In Section 5.1 we discuss what this means in our setting in more detail. The estimate of $\theta(X)$ retrieved from estimating the orthogonalized PLM in (5) is no longer suffering from the regularization bias. Excitingly, Chernozhukov et al. (2017) are able to prove that even in case that the first stage estimators - $\hat{f}$ and $\hat{h}$ - are converging at slower rates than root-n to the true parameter value, in the final estimator the regularization bias converges and the estimation error converges to zero at a potential rate of root-n.

In practice, the first stage of the estimation process consists of choosing an appropriate Machine Learning method, predicting the conditional expectation functions f and h and calculating residuals

$$\tilde{Y}_{it} = Y_{it} - \hat{f}(X_{it}, W_{it})$$
$$\tilde{D}_{it} = D_{it} - \hat{h}(X_{it}, W_{it}).$$

A welcome property of the DML estimation is its agnostic to the first stage estimator. To retrieve $\hat{f}$ and $\hat{h}$ we can choose the appropriate prediction method for the given setting.

4.3 Cross- against Overfitting

While orthogonalization takes care of the regularization bias plaguing more naive ML-based estimators, it implicitly induces a new bias. Machine Learning estimators are prone to overfitting models. Instead of picking up signals in features to predict the outcome, they start interpreting noise in the training data we feed them. To avoid this behavior, one can tune hyperparameters of the algorithm of choice to minimize this issue. Still, it is not unlikely that noise in the data is interpreted as a signal.

[4]Derivations are shown in Appendix A.A.

This same individual level noise is contained in the structural error terms of the PLM, ϵ_{it} and u_{it}. Thus, our predictions of f and h are not independent of these. **Using the orthogonalized outcomes and treatments to estimate $\theta(X)$ then leads to terms such as $u_{it}(\hat{f}(X_{it}, W_{it}) - f(X_{it}, W_{it})$ to show up in the estimation error $\hat{\theta}(X) - \theta(X)$. - UGLY** The dependence of the structural errors and the prediction errors - both driven by the individual level noise - are then not vanishing asymptotically. Similar to the regularization bias, this lets the asymptotic variance of the estimator explode and prohibits any convergence. However, it is rather easy to resolve this issue using sample splitting - a procedure called "crossfitting."
Instead of using all observations to find the estimates of f and h and then estimate $\theta(X)$ using the whole sample, consider the case in which we split the sample into two. The first sample is used to retrieve the first stage predictions. Those are used to predict the conditional means of the second sample, which are then subsequently used for orthogonalization and the second stage estimation. Since noise is independent across individuals, the noise affecting the first stage prediction error and the structural errors coming into play in the second stage estimation are independent as well. It is then easy to show that terms leading to problems when using the whole sample are vanishing asymptotically. In case we are interested in the unconditional average treatment effect (ATE), this procedure is repeated with the role of the samples reversed and the resulting estimators averaged. However, in the CATE case, we are interested in individual-level point estimates. Therefore, while the role of both samples is switched, we do not average any results but keep the individual-level estimates of all observations. The cross-fitting procedure for splitting up the sample into any K folds is described in Algorithm 1, which summarizes the whole DML estimation procedure.[5]

4.4 Retrieving the CATE

After retrieving the residualized outcome and treatment, the second stage estimates the conditional average treatment effect as defined in (6). It takes the following form

$$\theta(X) = \phi(X) \times \Theta, \tag{6}$$

where Θ is the baseline treatment effect of each individual and $\phi(X)$ is a mapping of our controls X. The form of the latter depends on the estimator chosen for the second stage. In Chernozhukov et al. (2017) estimators are proposed which have a linear second stage, either using a standard OLS estimator or Lasso to regress $\tilde{Y}_{it}$ on $\tilde{D}_{it}$. In these cases, the second stage boils down to a linear regression in which the residualized outcome is

[5] Note that Chernozhukov et al. (2017) argue that K=4 or K=5 performs reasonably well, even for smaller samples.

regressed on interactions $\tilde{D}_{it}$ and each element of X_{it}. This implies that the treatment effect we estimate is linear in the covariates X. It is also possible to include polynomials of or interactions between different elements of X_{it}. However, we choose a simple linear mapping of X for our linear DML approach presented in Section 5.

To identify nonlinearities in the CATE, we use a nonparametric approach that allows us to uncover these without defining them beforehand. Namely, we use a *Generalized Random Forest* estimator introduced by Athey et al. (2018). It has been developed to take advantage of the powerful random forest predictor for causal inference. Similar to DML, the GRF is an estimation framework. The GRF replaces the original objective function of the random forest algorithm (Breiman, 2001) with a moment condition containing some loss function that can be defined by the researcher. When using it for moment conditions such as (7) to identify conditional average treatment effects, the GRF is also known as a *Causal Forest*, which is presented in earlier work by Athey and Wager (2016). It allows for causal inference as Athey et al. (2018) develop the theory that allows retrieving standard errors of the estimated coefficients. Appendix A elaborates in more detail how the Causal Forest algorithm works and how it identifies the treatment effect. In our case, the moment condition is defined as

$$E\left[\left(\tilde{Y} - \theta(X) \times \tilde{D}_{it} - \beta(x)\right) \times (\tilde{D}_{it}; 1)\right] = 0 \tag{7}$$

where we choose the CATE $\theta(X)$ and constants $\beta(x)$ to solve it. The causal forest non-parametrically estimates $\theta(X)$ and therefore puts no assumption on the form of the mapping $\phi(X)$. The term $(\tilde{D}_{it}; 1)$ represents a matrix consisting of the vector of orthogonalized treatments and ones to capture the constant effects.

As part of our analysis we will compare the results to check whether the relationship is indeed linear or whether we discover non-linear heterogeneities that the linear DML approach does not account for and have not been considered in the literature yet. However, note that when using a nonparametric second stage the convergence rate of the estimator declines. This implies that the Causal Forest based approach is more demanding when it comes to the number of observations.

This has to look better and be more 'algorithmic'.

Algorithm 1 Double Machine Learning Estimator

1: Split up sample into K folds.
2: To estimate $\widehat{h}$ and $\widehat{f}$ for for the k^{th} fold use observations $j \notin k$.
3: To get residuals for observations in k, calculate $\widehat{h}(X_i)$ and $\widehat{f}(X_i, W_i)$ for $i \in k$ and use to retrieve residuals.
4: Once residuals of each fold retrieved, estimate $\theta(X_i)$.

5 Estimation and Results

We investigate the heterogeneity of the Marginal Propensity to Consume by estimating the following partially linear model

$$\Delta C_{it+1} = \theta(X_{it})R_{it+1} + g(X_{it}, W_{it}) + \epsilon_{it} \tag{8}$$

$$R_{it+1} = h(X_{it}, W_{it}) + u_{it} \tag{9}$$

where our outcome of interest is the change in consumption between two quarters, ΔC_{it+1} and our treatment is the rebate amount, R, household i receives. The choice of confounders X_{it} and W_{it} depends on the specification we estimate, as does which variables we consider to be part of X_{it} and thus have an effect on the treatment effect. Which variables are included in each specification is listed in Table X.X. We follow Parker et al. by including monthly dummies in W_{it} to account for seasonality and to capture any unobserved effects that might appear in periods in which households learn about the upcoming rebate. By canceling these effects stemming from the anticipation of the treatment, our estimate represents the effect of actually receiving the rebate.

In total, we distinguish between between three different layers in our estimations: we investigate different outcomes ΔC_{it+1} by using the rich information on expenditure categories included in the CEX. With the term 'specifications,' we distinguish between the different sets of confounders X_{it} and W_{it} we use. Lastly, we estimate each outcome-specification pair twice: once using the linear and once using the causal forest second stage. Since our estimation procedure predicts MPCs and we can retrieve standard errors for each household, we can run hypothesis tests on whether their response to the tax rebate is statistically significant.

However, one drawback in our specification including liquidity, salary, and income - which we also refer to as the *financial status* - is the already mentioned lack of detailed documentation of household characteristics in the CEX. Our sample size shrinks because they are not consistently documented for each interviewed household. This sample reduction can induce a sample selection bias because it is possible that households that answer questions on their liquidity are systematically different from households that do not provide such information. We have to keep this drawback in mind although the DML framework achieves fast convergence rates even in cases in which the first stage predictions do not converge rapidly.

5.1 Identifying the Income Shock

Since we use the same event and data source to estimate the MPC, our baseline identification approach is similar to Parker et al. (2013). The key factor is the design of the stimulus rollout, which we exploit to identify the income shock. The tax stimulus was paid out to households over several weeks as administrative and technological restrictions made it impossible to pay out all rebates at once. Instead, the date of rebate receipt depended on the last two digits of tax filers' social security number. These digits are randomly distributed, and therefore the timing of the treatment is random, rendering it exogenous from any household characteristics. Therefore, we observe rebate receipts at different points in time, which allows us to use all other households that received their rebate in a different quarter as the control group.

Parker et al. identify a potential source of bias in this identification approach. They argue that using the actual amount of tax rebate received can lead to an omitted variable bias. This concern arises because of how households' stimulus payments are determined. Remember that the tax rebate directly depends on the number of children, which certainly affects the absolute level of households' expenditures, as each dependent child adds 300 USD to the stimulus received. However, this is not a problem because we can control for the number of children in each specification.

The stimulus excluding the child bonuses equals the household's net income tax liability (NTL; in the following, also referred to as the net tax liability) as long as it is within the exogenously defined boundaries we discussed in Section 3.1. Parker et al. argue that the NTL might also drive changes in consumption, rendering the treatment endogenous. Their solution is to instrument the amount received with a dummy variable that only signals whether the stimulus was received or not in the given quarter. While their results and the authors themselves suggest that this is not much of a concern, we decisively disagree with their identification approach in general.

Parker et al. do not control for any variable related to households income or salary. These variables are without a doubt directly connected to our treatment because the NTL - i.e., how much a household owes in income taxes - is a function of the household's income. Excluding these variables leads to omitted variable bias causing inconsistent estimates. However, other than through the channel of income, we deem it highly unlikely that the net tax liability itself is driving changes in consumption. It might be possible that in other years the NTL plays a role for households' income as it can be perceived as an anticipated income - or liquidity - shock.[6]However, in 2008, the NTL affected households via their tax rebate, i.e., it does not affect the consumption change through other channels than what is captured by the tax rebate. Therefore, we argue in favor of using the actual rebate amount since it has two advantages: for one, we have an additional source of variation, and second, it allows us to estimate the continuous treatment effect and interpret it as the actual MPC. Moreover, this is in line with Misra and Surico (2014) whose results also

[6]Households usually should know that they will have to pay this/receive this because of past experience and because the NTL also depends on how much income tax was already paid during the previous year.

suggest that the endogeneity of the NTL is not a concern.

5.2 Main Results

We analyze our results in several steps and begin by looking at the empirical distribution of the estimated MPCs. Figure X.X shows the distribution of MPCs for the four main expenditure categories considered by Parker et al.: Food (FD), Strictly Non-Durables (SND) as defined by Lusardi (1996), Non-Durables (ND), and Total (TOT) expenditure. These categories are increasing in their level of aggregation, e.g., SND includes expenditures on food. A detailed list of all sub-components of each of these categories is listed in Appendix X.X. Here, we only want to highlight that the difference between SND and ND consumption categories are so-called 'semi-durables,' such as health expenditures, which are not included in the SND category.

A single plot of the empirical distribution in Figure X.X is retrieved as follows. We slice the range between the minimum and maximum of the point estimates into 20 equidistant bins and calculate the share of estimated MPCs that fall into each bin. The x-axis signals the borders of the different bins, and the y-axis shows the respective frequency. The blue part of a bar signals the total frequency of this bin. To illustrate how many of these estimates are actually rejecting the null of a zero MPC, we calculate the share of point estimates that reject the null at the 10% level within each bin. This is depicted by the red part of the frequency bars. I.e., a completely red bar implies that all observations within this bin are statistically significant, whereas a bar that is only red up to half of its height signalsthat this only holds for half of the point estimates within this bin. The vertical dashed line marks the average CATE - the average treatment effect across all households - as a benchmark. Each plot's description notes whether this ATE is significant or not.

In general, we find strong support for heterogeneity in the Marginal Propensity to Consume. The plots in Figure X.X show a large variation of MPCs across all consumption categories and estimation approaches. This underlines the importance of accounting for heterogeneous responses to income shocks. The heterogeneity is similar to Misra and Surico's findings. Our results show a large mass of households having a Marginal Propensity to Consume close to 0 and for many households, we cannot reject the null hypothesis of a zero MPC. On the other hand, there is a smaller share of households that show strong, significant responses. Table X.X depicts the shares of significant MPCs we estimate for each specification and model when we look at changes in non-durable consumption.[7]

Meanwhile, the ATE is very close to zero in all expenditure categories (except TOT) and across all estimations the bin it falls into never contains more than 15% of all point estimates. This highlights the weak representativeness of the ATE and its inability to

[7]Significance shares for the other three main outcomes discussed in this section are reported in Appendix A.A.

reliably assess the success of programs such as the 2008 tax stimulus. Accounting for heterogeneity reveals positive effects of the stimulus program on household spending that far exceed the average responses.

Taking a look at the measure of interest, the conditional average treatment effect, the similarity of our results to Misra and Surico's depends on our specification and choice of estimator. Using the linear DML, we find similar results across all expenditure categories in Specification 1, which includes the same set of controls as their estimation. The most significant MPCs for ND, SND, and FD goods are between 0.5 and 1; however, our point estimates for changes in total expenditure are largely exceeding the estimates presented by Misra and Surico. In general, the TOT responses are very strong and, in part, exceeding 1. The causal forest adapts quite more strongly to individual level heterogeneities, potentially driven by subtle variations in non-linear interactions between households' characteristics.

With respect to total consumption, we find a very wide distribution of MPCs with ranges varying across estimators and specifications. In general, it becomes obvious that the causal forest picks up even more heterogeneous patterns as we observe a more narrow distribution of point estimates using the linear DML. **move to another part: Turning our attention to specification 3, which includes liquidity, salary and income, any negative MPCs we find in the other two specifications are no longer existing.** However, we observe the significance shares reduce drastically comparing the linear DML to the causal forest based approach. This might be attributed to the amount of observations we are capable of using. Non-parametric regressors such as the Causal Forest converge slower and are more data-intensive. Still, the significant estimates we find point to the causal forest estimator picking up non-linearities in the CATE $\theta(X)$ that are ignored by the linear approach.

In case of Non-Durables - and as becomes visible in Figure X.X also for Strictly Non-Durables - we observe similar effects when it comes to controlling for liquidity and the changes in significance between the linear and causal forest estimator. However, the estimates we find across all specifiations are more reasonable than for the TOT consumption category. As Parker et al. and Misra and Surico argue, this is due to a small subset of households consuming large amounts in the new vehicle category in the same quarter they receive their rebate check. These households therefore experience an extreme change in consumption and we find these large estimates of MPCs. Exlcuding these consumption categories reduces the amount of outliers we observe.

In case of specification 2 we see that also controlling for the marital status of households, the significant MPCs are slightly wider disbursed and especially in the linear CATE we see the MPC showing an almost bi-modal distribution centred around a positive and negative amount. Still, including liquidity relaxes this effect a lot, which highlights that marital

status as well captures effects that are driven by salary, income or liquid assets.
An important observation is the fact that once we control for liquidity, the linear DML picks up no significant responses in the ND and SND categories of consumption. Moreover, any negative MPCs we find in the previous two specifications vanish. This hints to a negative bias induced by leaving out the *financial status* variables in Specifications 1 and 2. We address the occurrence of negative MPCs further below and our more detailed analysis of the drivers of the estimated responses in Section 6 yields more insights. This further underlines that the causal forest estimator detects non-linear heterogeneities that are not uncovered by the linear CATE. We observe that the described dynamics hold for the SND category as well although we see slighlty smaller ranges of MPCs.
Briefly touching upon reactions in the food consumption category, we see a pattern that is distinct from the ones we discuss above. Contrary to a wide distribution of MPCs in the more aggregate categories, food consumption reveals a bi-modal looking pattern of MPCs where some households show small positive and another set shows small negative reactions. The negative spike vanishes once we control for the *financial status*, a phenomenon that we observe across all consumption categories and we address further below.
Summarizing, our baseline results suggest that the 2008 tax rebate was used to increase consumption by a specific set of households. Contrary to the estimated average effect, these MPCs are statistically significant and are, in part, also exceeding ranges present in the literature so far. Again, we want to underline that these estimates provide evidence for a heterogeneous MPC that can be driven by a multitude of factors.
Lastly, we want to address the negative MPCs we find throughout almost all estimations. The MPC is bounded by zero at its lower bound by the definition of the concept. However, in our setting, negative estimates are still possible, as Kaplan and Violante point out. This relates back to their argument that when using the tax stimulus, we estimate a 'rebate coefficient' and not necessarily the MPC. However, the rebate coefficient can very well be negative, as they show in estimations using their calibrated two-asset model. In their model, they explain the heterogeneous response to the 2001 US tax stimulus by distinciting between hosueholds that are welathy but only hold illiquid assets (wealthy hand-to-mouth) and households that have no liquidity and hold no illiquid assets (poor hand-to-mouth). In this two-asset model, the households holding illiquid assets have to pay transaction costs to increase their holdings of the illiquid asset. Kaplan and Violante show that when these transaction costs are relatively low compared to the size of the income shock, households will choose to pay the costs and make a deposit once they receive the payment resulting in a negative effect on contemporaneous consumption. In our case, once we control for liquidity, we observe a reduction in the significantly negative MPCs and in part see them completely vanish. Staying in Kaplan and Violante's framework, it is possible that their described effect is captured by the liquidity related variables as

a proxy. We have no measure of illiquid holdings but the variables additionally included in Specification 3 might act as kind of a proxy, accounting for this illiquid asset effect on consumption change.
Overall, our results hint to a small amount of households showing a significant consumption response to the tax stimulus payments. Next, we investigate what these responses are driven by.

6 Understanding the roots of heterogeneity

In the previous section, we discussed the conditional average treatment effect of each individual given their specific set of characteristics. Prior contributions have looked at the correlations between the significance of the estimated MPC and households characteristics to get a glimpse into which factors play a role in the MPC. However, this approach does not reliably tell us which variables really drive the response. The correlation might very well be spurious or driven by other factors that are correlated with the characteristic we are looking at. Therefore, it is more fruitful to look at measures that can help us identify what role a variable plays in our predicted MPCs. In the case of specifications using the linear DML estimator, we know that this relationship is linear by construction since the CATE is defined as a linear combination of the single effects of interactions between treatment and the respective variable (see equation (6)). However, the causal forest-based approach will help us reveal whether there are any non-linear patterns underlying the effect of characteristics on the MPC without assuming any functional form of these patterns.
For this, we turn to the Machine Learning literature, which has developed a number of tools to analyze the relationship between prediction and feature. Feature is a different term for control variables used in the Machine Learning literature. In our setting, these are the variables we condition on to find the CATE, i.e., variables contained in X_{it}. Since variables in W_{it} are assumed to not impact the CATE, they are not contained in the second stage and therefore play no role in predicting individuals' MPC.

6.1 Marginal and Partial Dependence Plots

this section is quite long for something I do not show, right?
Two popular approaches are marginal plots (M-Plots) and Partial Dependence Plots (PDPs; Friedman, 2001). Both use the same general idea to quantify the impact of some feature x_S on our predictions: we replace the value of x_S of each observation with some value v_1. Then we fit our trained prediction model to this "counterfactual" dataset and take the average over all these predictions. For example, we predict for each individ-

ual what their MPC looks like if they had a certain age and average the predicted MPC. Then we continue with $x_S = v_2$ and so on, where the values v_j are chosen from a grid along the distribution of x_S. The difference between M-Plots and PDPs is the distribution of all other features $X_C = X \setminus x_S$ we average over. In the case of Marginal Plots, contrary to what the name might suggest, we use the conditional distribution of X_C given x_S, $p_{X_C|x_S}$, to obtain the impact of x_S on our prediction. On the other hand, PDPs use the marginal distribution of X_C, p_{X_C}. Partial Dependence Plots are more common in the Machine Learning literature as M-Plots suffer from a severe weakness when features in X_C are correlated with x_S. However, PDPs also fail to reliably uncover the effect of x_S in such a setting.

To illustrate the issues arising in M-Plots and PDPs when features are correlated, let us consider a simple example. Let's say we have some predictive model m that only depends on two predictors x_1 and x_2, which are positively correlated. To now calculate the M-Plot of x_1 at v_1 we use the conditional distribution $p_{x_2|x_1}$. In practice, we plug in $x_1 = v_1$ for each observation that is within a specified neighborhood of $x_1 = v_1$ (e.g. observations in the same quantile). Then we predict and average to obtain the M-Plot value at $x_1 = v_1$. Repeating this procedure for other values v_j then results in the M-Plot of x_1. However, because the two variables are correlated, we do not know which variable drives the observed effect - if x_1 is increased, the values of x_2 we use for our predictions also increase because we only use x_2 of observations that are close to having $x_1 = v_j$. This problem is known as 'conflation.'

On the other hand, Partial Dependence Plots do not suffer from this problem because they use the marginal distribution of x_2. We use all observations of x_2 instead of only looking at a neighborhood in which $x_1 = v_1$ and, therefore, average out the effect of x_2 on our predictions. Since we use the same set of x_2 values at each point v_j, we know that changes in our predictions must stem from x_1. Still, the PDPs are not a good tool when features are correlated, and this is connected to the machine learning estimators we apply them to. These are nonparametric estimators that are usually very weak in predicting outcomes based on observations they have never seen before. This extrapolation, however, becomes necessary when we create the "counterfactual" dataset by setting $x_1 = v_j$. By doing so, we effectively create observations that are extremely unlikely or even impossible to observe in the real world because of the correlation between the features. For example, in our data, age and salary are strongly correlated, which is quite intuitive because once retired, households do not receive a salary anymore. When creating PDPs, we ignore this fact and create households that have a high salary and are very old. The weakness in extrapolation leads the model to create imprecise predictions, which then severly bias the Partial Dependence Plots. (Apley and Zhu, 2020)

Therefore, while PDPs do not suffer from theoretical drawbacks like M-Plots, in practice, they are unable to uncover the effects of x_1 on our predictions in a stable manner because of the underlying predictive estimator. If the true model is indeed linear and we use a linear prediction method with the correct specifications of any interaction terms etc., then this extrapolation issue is unlikely to occur. Moreover, by construction, a linear predictor will result in linear Partial Dependence Plots.

6.2 Accumulated Local Effects

To circumvent issues arising in M-Plots and PDPs from correlated features, Apley and Zhu (2019) propose Accumulated Local Effects (ALE). The extrapolation issue PDPs suffer from is bypassed by using the conditional distribution $p_{X_C|x_S}$ as we do in M-Plots. The 'conflation' effect that results from this is tackled by not using average predictions at $x_S = v_j$ but rather the average marginal change in predictions at this point. In other words, we use the partial derivative of our predictor in question m with respect to x_S at the point v_j. Although tree-based machine leanring methods such as the causal forest have no concept of a gradient, Apley and Zhu are able to derive proofs for non-differentiable predictors m. Further does the gradient play no role when actually estimating the ALE as it is approximated by a step-function. The ALE is defined as

$$\hat{m}_{S,ALE}(x_S) = \int_{z_{0,S}}^{x_S} E_{X_C|X_S=x_S}[\hat{m}^S(X_S, X_C)|X_S = z_S]dz_S - constant. \tag{10}$$

Looking at this equation step-by-step reveals how the ALE recovers the effect of x_S on our predictions even when features are correlated. As already mentioned, the ALE avoids 'conflation' by using the partial derivative of m, where we have $m^S = \frac{\partial m}{\partial x_S}|_{x_S=v_j}$ as the partial derivative of m evaluated at the point we want to find the ALE for. Since we only look at an infinitesimally small change, this change in x_S will not affect the features that are correlated with it in X_C unless the correlation is extremely high. In our analysis, we would want to avoid this case anyways to avoid problems in the estimation itself (e.g., multicollinearity). Once the changes in prediction are obtained for each observation, we average them over the conditional distribution, i.e., only using observations that are within a neighborhood of $x_S = v_j$ and actually exist. Now we have the average local effect of x_S on our prediction. To better visualize the global role of this feature, Apley and Zhu argue that this can be achieved by accumulating all local effects up to $x_S = v_j$. Thus, we simply integrate overall local effects up to $x_S = v_j$, where $z_{0,S}$ is the lower bound of the distribution of x_S.[8]

To estimate the ALE we use the following estimator, which illustrates the procedure in

[8] For more on this, see Section 5.2 of Apley and Zhu (2019) where they demonstrate how accumulation helps to improve the interpretability of ALE plots.

more intuitive terms:

$$\hat{\hat{m}}_{j,ALE}(x_S = v_j) = \sum_{k=1}^{k_j(x_S=v_j)} \frac{1}{n_j(k)} \sum_{i:x_{i,j} \in N_j(k)} [m(z_{k,j}, x_{i,\setminus j}) - m(z_{k-1,j}, x_{i,\setminus j})] \tag{11}$$

First, we bin our data into n_b bins based on quantiles of the distribution of x_S. To mimic the marginal change represented by the partial derivative m^S in 10 we make two predictions for each individual. For an observation i that falls in bin k, we predict its outcome with $x_S = z_k$ and $x_S = z_{k-1}$, where z_k represents the upper bound value of bin k. We then average over all individuals that fall within this bin k denoted in equation (11) by the neighborhood $N_j(k)$. Finally, we accumulate all predicted preferences over all bins up to the bin at which point x falls, denoted by k(x). Only looking at individuals within the neighborhood $N(k)$ accounts for the conditional distribution used in (10).
Additionally, Apley and Zhu propose to center the effect around the average of all ALEs such that the mean effect is zero. Then, the ALE has to be interpreted relative to the average prediction, and it shows whether for a given $x_S = v$ the effect of x_S is above or below the average prediction. I.e., whether x_S affects our predictions at $x_S = v$ more than it does on average. In practice, the *constant* in (10) is replaced by

$$\frac{1}{n} \sum_{i=1}^{n} \hat{m}_{j,ALE}(x_{i,j}).$$

Note that we yet cannot say something meaningful about the statistical significance of these results. Most fields are only interested in the predictive power of machine learning methods and in understanding how these predictions are achieved, but there is no notion of statistical significance in these settings. Therefore, a specific approach to quantify the uncertainty of these measures has not yet been developed. While a deeper look is out of scope of this paper, we use bootstrapping to briefly dive into the topic of statistical significance. We simulate the ALEs for $n_{bootstrap}$ samples. These create an empirical distribution on which basis we can calculate a pseudo confidence interval using the reverse percentile approach (Davison and Hinkley, 1997, p. 194 eq 5.6). We report these as the red lines surrounding the ALE in Figures X.X. We see that these bands are very wide in certain parts - especially in areas where there is a small number of observations. We strongly encourage a deeper investigation of the statistical properties of ALEs and a potential way of quantifying their uncertainty in a more rigorous manner. While the ALEs show us the relationship between a specific variable and our predictions, having a sound theoretical foundation would be desirable to better assess the statistical significance of these relationships.

6.3 Results

To investigate what role single features play in the MPC we mainly focus on the response with respect to changes in non-durable consumption. We will briefly draw comparisons to other consumption categories where we gain important insights from doing so. Moreover, we only discuss the features relevant within the context of the literature. As we will see these are proving to be drivers of the MPC as expected. The ALE figures for all specification-estimator pairs of each of the main consumption categories are presented in Appendix A.A. As pointed out in Section 6.1, by the construction of the estimator and the ALE, it will always depict a linear relationship when looking at the linear DML setting. However, we can still infer in what direction the relationship is going - e.g., whether predictions are above or below average for young people. More importantly, it is useful as a benchmark to compare the ALE of our estimates using the causal forest as the second stage estimator.

Comparing the ALE of the causal forest estimator to the linear model helps us understand where non-linearities play an important role in understanding MPC heterogeneity. Still, we have established in Section 5.3 that the causal forest estimator reveals more significant MPCs in specification 3, where we control for liquidity, which is likely to occur because of nonlinearities not picked up in the linear CATE model. This notion becomes evident right away when we compare the ALE of the linear CATE and the non-linear CATE we retrieve with the causal forest estimator. The clear positive relationship between age and MPC we find in the linear estimator immediately breaks down in the causal forest based model. Instead, we see a rather unstable relationship with wider pseudo confidence bands. Younger and middle-aged households (45 to 55) seem to be experiencing lower MPCs than older households. However, only for very narrow settings do our bootstrapped CIs completely lie below zero. This signals a very unstable relationship between age and the MPC overall, which is a strong contrast to the linear DML results. These dynamics are observed throughout all consumption categories.

Interestingly, the role of age in the linear model is reversed once we introduce liquid assets, salary and income in Specification 3 (Figure X.X Panel B). The deviations decrease substantiallyand the ALE shows that older households seem to have a lower MPC. Given the correlation between age and each of these variables this underlines their importance since age seems to act as a proxy for them. This bias vanishes by including the *financial status* variables. This dynamic is also upheld in the causal forest estimator, where we see that the effect of age to be more stable and more closely fluctuating around zero. It supports the notion that younger households show a stronger response to the income shock.

The main channel identified in the literature is liquidity. Our discussion of the underlying theory of binding borrowing constraints and lacking access to liquid assets provides the intuition for the following analysis. Note that the pattern described here is also present across all consumption categories although the ALE magnitude is varying. The picture painted by the liquidity channel is only in part supported by our results. First, the linear estimator shows a negative relationship between MPC and liquidity only for changes in total consumption. When we turn to non-durables and strictly non-durables, the ALE actually suggests that households with higher liquidity show a stronger positive reaction to the income shock. While these effects are very small, this still suggests that the liquidity channel does not necessarily act as expected in creating MPC heterogeneity. One possible explanation is that for high liquidity households the tax rebate represents a relatively small income shock. Therefore, they do not depend on it financially, are less informed and are taken more by surprise. However, this phenomenon is quickly turned around by the causal forest results. It fully supports the liquidity channel as we find that low liqudity leads to a high MPC, while once a certain threshold in liquid asset holdings is exceeded the MPC is decreased. Hence, the liquidity channel in cases of anticipated income shocks such as the tax rebate is underlined by the causal forest results. Low liquidity households are not able to react at the announcement to the shock and thus show a strong reaction once the tex rebate is received. On the other hand, a higher level of liquidity enables households to increase consumption prior to receiving the rebate and smooth consumption level overall.

Interestingly, the causal forest reveals that at the very bottom of the liquidity distribution, households show a negative response to rebate receipt. These negative responses might explain why the linear DML shows no role of liquidity at low levels of liquidity as it does not fully capture non-linearities and interactions, leaving the estimator biased. On the other hand, they are contrary to the main idea of the liquidity channel. This result shows that the clear relationship the liquidity channel paints in theory does not hold up in the data. Instead, the negative effect of very low levels of liquidity on the predicted MPC highlights a new role. A possible explanation might again be related to the circumstances under which the stimulus is received. Households with no liquidity might save up the stimulus to build up a buffer during the times of economic downturn as were taking place in 2008.

here tran-sition missing The sharp drop after roughly 2,000 USD suggests that households that were capable of increasing their consumption at announcement by tapping into their liquid assets consequently have a lower MPC out of the rebate as they save it to further smooth consumption by saving it. The drift to a positive role of liquidity once households have high liquid asset holdings is very unstable as evidenced by the very wide CIs that also include negative values and zero across all levels of liquidity. Most likely, this is due to the small amount of households in our sample that report such high levels of liquidity leaving the causal forest performing imprecisely.

Lastly, we turn to the other two financial variables, salary and income, which are presented in Figure X.X. In both we see highly non-linear dynamics that appear to be quite similar once a threshold of 50,000 USD is passed. We see that a very low salary leads to a lower MPC, which is reasonable to expect considering the economic circumstances in which the tax rebate took place. Low income households struggling during the recession might show a smaller response because they use the tax rebate for precautionary savings (maybe a citation here of a precautionary savings paper). On the other hand, the MPC is higher for households that have higher salaries in range between 50,000 and 75,000 USD before the effect vanishes at higher levels. This is also in line with what we expect as the tax rebate was phased out at reported incomes above 75,000 USD. Thus, high income households for which the tax rebate plays a minor role in their income anyway also received even less.

References

Bunn, P., J. Le Roux, K. Reinold, and P. Surico (2018): "The consumption response to positive and negative income shocks," *Journal of Monetary Economics*, 96, 1–15.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2017): "Double/Debiased Machine Learning for Treatment and Causal Parameters," *arXiv:1608.00060 [econ, stat]*, arXiv: 1608.00060.

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

Parker, J. A., N. S. Souleles, D. S. Johnson, and R. McClelland (2013): "Consumer Spending and the Economic Stimulus Payments of 2008," *American Economic Review*, 103, 2530–2553.