# Detecting Heterogeneity in the MPC: A Machine Learning Approach

Last Update: 15.10.2021

Master's Thesis (Draft)

Department of Economics

University of Mannheim

submitted to:
Prof. Krzyzstof Pytka, PhD

submitted by:
Lucas Cruz Fernandez

Student ID: *********
Studies: Master of Science Economics (M.Sc.)

Adress: ******
Phone: ******
E-Mail: ******

Mannheim, ADD DATE

# 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 centre of a wide academic debate in economics but also of major importance for policy makers. While the former revolves around verifying or neglecting the main mechanisms of the Permanent Income Hypothesis (PIH), the latter are interested in improving government transfers to more efficiently use public funds. These two sides have sparked many investigations using a wide array of approaches to quantify hosueholds' responses to income shocks. This Marginal Propensity to Consume (MPC) - at the centre of macroeconomics since Keynes introduced it at the heart of his General Economic Theory - quantifies how much households will spend on consumption of each dollar they receive from an income shock. While research has long focused on quantifying whether the MPC out of income shocks is zero and thus in line with the PIH, the literature has seen a shift of focus over the last decade. On average most studies support the notion of a zero MPC, however, more recent evidence suggests that for specific groups the response is significantly different.

Empirical research related to the MPC and its heterogeneity has used several settings to identify income shocks. One of the most prominent is to use natural experiments in which households receive exogenous 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 using 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 not only identify heterogeneities in the overall MPC but also to analyse which categories of consumption goods households spend their rebate on. However, our econometric approach sets us apart as it is more sophisticated and more precise in detecting heterogeneities compared to any contribution we are aware of.

Namely, the two main channels are life-cycle dynamics and liquidity. The former is driven by a consumer's age and the associated fluctuation in income. As data consistently shows (Attanasio and Weber, 2010), consumption expenditures follow a hump shape over the life-cycle rather than being roughly constant as we would expect under the PIH. This anomaly is often referred to as the retirement-consumption puzzle and connected to an increased amount of free time which allows households to reduce the cost of their consumption. Therefore, we would expect a reduction in the measured MPC in age.

In the case of liquidity, its role is linked to the nature of the income shock and borrowing constraints. 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 a higher future income. Thus, once the shock realises, 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 realises 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 behaviour downward. Meanwhile, a positive shock will always be saved and stretched over future periods, no matter the level of households' liquidity. E.g. Bunn et al. (2018) document this asymmetry depending on the sign of the shock.
It is important to highlight that in our setting, households experience an anticipated positive income shock. The Economic Stimulus Act was signed into law by President Bush in February 2008 and payments of tax rebates started in April of the same year. Therefore, following the previous arguments, theory would expect older households to react less, but liquidity not being a major driver. However, the tax rebate was disbursed to US taxpayers during a time of national and global economic downturn. Thus, 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 or utilities). However, Parker et al. (2013) emphasise 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.
One major issue in the existing literature is the way how heterogeneity is measured. Most studies rely on either splitting their sample into smaller sub samples and estimating the MPC within each sample or use dummy variables that are defined by the authors based on some continuous variable. These approaches suffer from the severe issue that any heterogeneity that does not fall into this pre-defined pattern is not captured and will muddy the results of these investigations. In the worst case these procedures miss to pick up existing heterogeneity or missing patterns within these pre-defined subgroups. On the contrary, our Double Machine Learning (DML) approach allows us to estimate the conditional MPC out of the tax rebate of each individual household. Prior studies have to rely on looking at the correlation between their estimates and characteristics such as liquidity, but our setting enables us to calculate more sophisticated measures that capture the influence of each variable on the MPC.
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 anticiapted 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 coeffcient', which is not necessarily equivalent to households' overall MPC. We compare our estimates with the range found in the literature

using different income shock sources to get a grasp of whether this difference might play a role and for what households it does. However, 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 policy making. While in some cases when economic relief is urgent broadly defined, non-targeted stimuli might be a good option to pursue, targeted transfers can play a major role in many policy settings. **(rephrase)**

**this needs to be implemented somewhere** By exact definition, the MPC is the reaction to an unanticipated, transitory income shock. However, in out setting, the shock cannot be fully seen as unanticipated as the tax stimulus payment was signed into law and therefore known to the public several months before the payments were conducted. However, in such cases to identify only the contemporaneous reaction of households the empirical literature (e.g. Parker et al. (2013)) .

**section on results here** We show that indeed both these channels play a role in the heterogeneity of households' response to the 2008 tax stimulus. Similar to the existing literature we find... However, additionally we are able to show that the heterogeneity is not only linear/indeed linear...
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 7 concludes.

# 2 Literature Review

The literature investigating the size of the MPC and potential heterogeneity can broadly be categorized into three different strands. The first one uses quasi-experimental settings to exploit variation in income to estimate households' MPC. The second uses surveys that explicitly question participants about their MPC - be it out of actual or hyopthetical income shocks. Lastly, a vast literature focuses on building sophisticated macroeconomic models that are calibrated to match real world data and subsequently estimate the MPC agents experience in these models. Our work falls into the first of these categories.
In this section we briefly summarize the findings in all three and additionally discuss two studies - Parker et al. (2013) and Misra and Surico (2014) - in more detail as they investigate MPC heterogeneity using the same data as we do.
Quasi-experimental settings appear all the time in the real world, e.g. in case of a specific

policy being implemented or another exogenous shock happening. Researchers interested in MPC heterogeneity focus on shocks that alter the income of a household. For example, Fagereng et al. (forthcoming) use panel data from Norwegian administrative data on winners of a state lottery in which most citizens participate. Receiving a payment from the lottery can be seen as an unanticipated income shock because the chances of winning are so low. They find that households winning the lottery spend almost half of their win within one year and 90% after 5 years. Moreover, liquidity and age are the only variables correlated with the MPC, providing evidence for the existence of the liquidity and age channels. In similar vein, Golosov et al. (2021) construct a dataset of lottery winners in the USA to estimate their MPC and labor market response. They make use of tax forms provided by the lottery winners and general income tax statements. Their main goal is to estimate the labor market responses to windfall gains in unearned income but their strategy allows them to identify the MPC as well. Using a Difference-in-Difference estimator, their estimated MPC is around 60ct out of each dollar earned on average, while labor earnings are reduced by 50ct. To investigate heterogeneity in these responses, the authors split their sample based on the quartile along the liquidity distribution. Further supporting the liqudity channel, they find that households in the highest quartile spend only 49ct while the lowest quartile spends almost 80ct of each dollar won 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. (forthcoming)) or model consumption as a function of their observed variables (Golosov et al. (2021)).

Gelman et al. (2018) use the government shutdown in the U.S. as a transitory liquidity shock. Hence, contrary to other literature they only estimate how liquidity changes the consumption behavior and not the MPC directly. Still, their setup allows them to disentangle the pure effect a liquidity shock has on consumers spending as government workers receive a payback of their wage once the government shutdown is over. Hence, there are no changes in expected income. Meanwhile, studies using income shocks cannot quantify what effect stems from the liquidity channel and what stems from changes in expected income. Their findings highlight that low liquid households react more to a negative liquidity shock as they have no assets to fall back on. Low liquid government workers started postponing their credit card payments, while simultaneously increasing the amount spend using them. **probably only add very short inside of this as not so much related to raw MPC and little/bad heterogeneity investigation**

The second strand of literature uses survey data from field surveys that question households about potential or actually relaized income shocks and how their reaction looks like. Bunn et al. (2018) use (**use twice here**) data collected by the Bank of England to assess the asymmetry that we expect in households' reaction depending on the sign of

the income shock. As mentioned before, the liquidity channel suggests that an unanticipated shock calls for a stronger reaction if its negative. Indeed, the authors are able to provide ample evidence for such a reaction with their estimated MPC out of a negative shock being being between 5 to 12 times as high as the reaction to a positive shock. The balance sheet data their source provides also enables them to show that borrowers show a more pronounced asymmetry and reaction, which is also in line with the theoretical mechanisms of the liquidity channel. This also holds for households that face some kind of liquidity constraint. Additionally, Bunn et al. are capable of replicating their estimated MPCs in a model with households at the borrowing constraint. These findings are further underlined by Christelis et al. (2019) who use Dutch data for a similar study. Summarizing these studies strongly suggest the existence of the mechanisms related to households at their borrowing constraint and precautionatry saving motives (**the latter must be elaborated on in intro**).

## 2.1 2008 Tax Stimulus

Lastly, we want to elaborate in more detail two already mentioned quasi-experimental studies: Parker et al. (2013) and Misra and Surico (2014). These two studies have been quite influential when it comes to studying MPC heterogeneity and use the same dataset as we do. Therefore, we want to lay out their approache sin detecting heterogeneity in greater detail here and highlight the major advantages our estimation approach offers. Compared to the general literature, we are the first to focus on the form of the heterogeneities reported in the literature so far (**check whether this isn't too bold of a claim**).

Thanks to Parker et al. (2013), the 2008 and 2009 BLS added questions about the tax stimulus to the CEX survey. Parker et al. (2013) use this data to first estimate a simple homogenous model controlling for age and the change in the family size using Ordinary Least Squares (OLS). Meanwhile, they set off to detect heterogeneity using interaction terms. More precisely, similar to Golosov et al. (2021) they create dummies that signal to which quartile a household belongs, where the quartiles are defined based on the distribution of various variables. The interaction terms then capture systematic differences between households across these quartiles. However, as we have already pointed out, this procedure can be quite problematic. It increases the potential to miss substantial heterogeneity across households. For example, consider the case where the largest heterogeneity is between the top 10% and the top 20% of households along some distribution. The approach using sample splitting or dummy variables is not capable of detecting these heterogeneities as those households belong to the same category. Additionally to this problem, using an OLS estimator, Parker et al. (2013) completely ignore any panel dy-

namics that might take place. Meanwhile, individual level fixed effects are identified as a strong potential channel that drives heterogeneity in MPC. The results by Parker et al. (2013) are therfore potentially biased (**two times potential**).
To adress the heterogeneity issue, Misra and Surico (2014) follow a new approach: they employ a quantile regression estimator to estimate the heterogenous response of households to receiving the rebate. The quantile regression estimator is similar to the OLS but estimating conditional quantiles instead of the conditional mean. The authors therefore claim to estimate the conditional MPC out of the rebate for any quantile of consumption change. However, there are severe dissimilarities with what they are actually estimating and what they interpret. QR estimates how much a variable affects the outcome at a given quantile, e.g. the 10% quantile, of the conditional distribution of the outcome.

**after mentioning all studies, here notes on why sample splitting or dummies are way worse than my approach** While their procedure of sample cutting is common in the literature, it is an inconvenient procedure to detect heterogeneity as it only allows the authors to analyse differences between their pre-defined groups. Hence, in case heterogeneity is strongest between other groups, their findings underestimate or even completely miss patterns in the data.

## 2.2 Channel description

**This will be pushed/worked into another section: decide whether at beginning of literature review or in introduction; for now here only to formulate something in a more consistent manner than what's in the intro right now**
The theoretical literature has identified several channels which drive MPC heterogeneity. The two most prominent ones are life-cycle dynamics and liquidity. The former is driven by a consumer's age and the associated fluctuation in income. As data consistently shows (sources), consumption follows a hump shape over the life-cycle. In the case of liquiditiy, its role is linked to the nature of the income shock and completeness of the credit market. 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 a higher future income. Thus, once the shock realizes, we will observe an increase in consumption - although if we follow the PIH, this increase is rather small as the additional income is spread out over all future periods. In case of a negative anticipated shock, saving is always possible for any household and hence we will not see a reaction once the shock realizes. E.g. Bunn et al. (2018) document this asymmetry depending on the sign of the shock (**they document this for unanticipated shocks, shift back**). Thus, more liquid households react less to a positive anticipated shock in comparison with liquiditiy 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. However, these theoretical predictions are made within a permanent income framework in which households try to smooth consumption over time.

**Literature Notes**

Parker et al. (2013)

- three sources of variation exist:
  1. timingand type of payment
  2. amount
  3. type of payment
- result: on avg. hosueholds spent 12-30% of rebate on nondurable consumption goods → significant
- life-cycle/PIH model (LCPIH) is rejected by these findings
- LCPIH: no response to anticipated shocks at timing of arrival → borrowing/liquidity constraints may be main driver
- prior research: larger payments may skew consumption/usage of rebate towards durables (Souleles (1999) finds significant increase in ND and D goods in response to larger payments (federal tax refund in springtime))
- problem data: assets are not measured in detail and frequently
- find no significant response (interpret it anyway)
- keep in mind that tax rebate was disbursed during time of major economic downturn/turmoil
- look at relationship to owning house compared to renters → homeowners spend more than renters
- closest literature: Agarwal, Liu and Souleles (2007); Broda and Parker (2008); Bertrand and Morse (2009)
- Tax rebate
  - at least 300 (couples 600), at most 600 (1200) if 300+tax liability are above 600

  - at least 3000$ qualifying income
  - phased out with income starting at 75000$: 5% reduction of amount gross income exceeds 75k (couples 150k)
  - hh received a notice in advance of payment → anticipated shock!

- CEX
  - questions were added from june 2008 to march 2009
  - use 2007 and 2008 waves of CEX (2008 data includes first quarter of 2009)

- identification
  - use

# 3 Data

We use data collected by the Consumer Expenditure Survey (CEX) that is administered by the Bureau of Labor Statistics. Additionally to the main questionnaire, the CEX added questions about the stimulus payments to their surveys conducted between June 2008 and March 2009 (**isthis correct time in dataset?**). The main advantage of the CEX is that it creates a unique representative sample that contains finely grained information on the type of goods households consume. While its main purpose is to serve as the benchmark to determine the goods basket used to measure inflation in the USA, it enables a detailed analysis on what households spent their rebate on. In the following, we briefly outline the stimulus program and describe the CEX data.

## 3.1 The 2008 Tax Stimulus Program

**work following fact in as well: People that did not have to file for taxes because their income was too low had to file anyway to be eligible for rebate; at least have 3000 dollars of income though (wages, self-employed, social security)** 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. Next to the stimulus payments, which made up roughly two thirds of the program, the ESA also enacted other steps meant to provide economic relief such as enabling government owned entities (Fannie Mae and Freddie Mac) to buy up more mortgages. However, we only focus on the effects of the stimulus payments.

The rebate was paid out to any household that filed for income taxes (**unsure about the wording of 'to file for taxes'**) and reported a minimum annual income of 3,000 USD in 2007. Eligible households received their net tax liability as their rebate, however, the payment 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 additional 300 USD per child. Also, the stimulus was designed to fade out in gross income. Households reporting an income above 75,000 USD - 150,000 USD for couples - the rebate was reduced by 5% of the amount the income exceeded this upper limit. (**the last sentence is close to wikipedia source; also nee a source fot his paragraph**)

## 3.2 Consumer Expenditure Survey

The CEX is a representative survey of households in the USA interviewing households about their consumption patterns on a quarterly basis. Once a household is selected to

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`

participate, they are interviewed a total of five times. The first interview is a baseline interview in which some general household characteristics, the financial circumstances and their stock of nondurable goods are documented. The next four interviews are administered every three months in which households are asked to document their expenditures over the period since the last interview. After that, the household is rotated out of the CEX and replaced with a new one. Hence, each month of data documented in the CEX contains a different set of households as new ones are added and others are rotated out of the survey. Figure 1 is taken from the CEX website and illustrates this procedure. Note that here a household is defined as a Consumer Unit (CU), which can represent either a number of blood or legally related persons (e.g. foster children count as well), 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 CUs members are collected with respect to their relationship to the reference person. This person is defined as the one named when asked "Start with the name of the person or one of the persons who owns or rents 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 a dense documentation of household characteristics. For example, financial variables such as liquidity are only asked for in the initial screening interview, which for

example prevents us form demeaning all controls in a panel like setting. Instead we have to rely on the inclusion of individual level fixed effects, which we explain in more detail in section 4. 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 analyse 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 (DML) framework developed by Chernozhukov et al. (2017). This new kind of estimation approach allows to efficiently estimate semi- or non-parametric models of treatment effects. It has the major advantage that it does not restrict the effect of confounders on the outcome to a specific functional form. Instead it uses Machine Learning methods to freely estimate this relationship. Through the orthogonalization step discussed below it takes care of any confounding effects and cleanly estimates the pure effect of treatment on the outcome. 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. On the other hand, sophisticated DML estimators can detect these interactions without knowing them beforehand. Meanwhile, its implementation procedure deals with common biases arising in more naive estimation procedures that employ Machine Learning methods. This opens the door to combine powerful machine learning algorithms - proven to perform well in detecting patterns in data - with causal inference.

From a more theoretical perspective the DML estimator yields very efficient properties when it comes to its asymptotic behaviour. Under certain assumptions, Chernozhukov et al. (2017) are able to prove root-n consistency of the estimator, a rate of convergence not achieved in other nonparametric estimators. 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 approach 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 average treatment effect of $D_{it}$ on $Y_{it}$. The choice of these variables is left to the researcher. We are interested in $\theta(X)$, the conditional average treatment effect (CATE). In Rubin's potential outcomes framework 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, treatment is not binary but continuous, hence $\theta(X)$ 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 $X = x$. The task is now to find an appropriate estimator to find $\hat{\theta}(X_{it})$.

## 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 use 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 due to regularization. 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.[1] However, we can deal with this regularization bias using orthogonalization.

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

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). It is straightforward to estimate these conditional means using any ML method of choice. Using these 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 $W$ on them, respectively. The residuals only contain variation that does not stem from any of the confounders. In Section 5 (**make reference to subsection in which we discuss identification**) we discuss what this means in our setting in more detail. Indeed, the estimate of $\theta(X)$ retrieved from estimating the orthogonalized PLM in (5) is no longer suffering from the regularization bias. Excitingly, the authors are able to prove that even in case that the first stage estimators of $\hat{f}$ and $\hat{h}$ are converging at slower rates than $\sqrt{n}$ to the true parameter value, in the final estimator, the regularization bias clearly converges to zero and can achieve root-n consistency.
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. Thus, it allows choosing 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 impliclty 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 data. To avoid this behaviour, one can tune hyperparameters of the algorithm of choice to minimize this issue known as overfitting. Still, it is not unlikely that noise in the data is interpreted as a signal. While there is a vast literature on this issue on its own, this also leads to problems when following the procedure described so far.

The noise picked up in the first stage predictions stems from the error terms $\epsilon_{it}$ and $u_{it}$ in (1) and (2). Therefore, the predictions $\hat{f}$ and $\hat{g}$ are driven by these error terms. When we now calculate the respective residuals of outcome and treatment, these are by construction correlated with the error terms, which also appear in the seocnd stage estimation of **here reference of equation of estimation of second stage - right now this is nowhere but I have to show it**. Similarly to the regularization bias this lets the asymptotic variance of the estimator explode and prohibit 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 the errors are assumed to be i.i.d., noise in one sample is not correlated with the noise in the other **don't like this sentence yet**. 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 are averaged. However, in the CATE case we are interested in individual-level point estimates. Therefore, while the role of both samples are 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 summarizes the whole DML framework.[2]

## 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). We assume that it takes the following form

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

where $\Theta$ 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 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 polyno-

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

mials 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. When using it for moment conditions such as REF TO Moment Condition it is similar/the same (?) as the Causal Forest presented in Athey and Wager (2016) and is often referred to it with the same name. The Causal Forest 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. Morevoer, they 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 (EconML, 2020)

$$E\left[Y - < \phi(X), Y > -\beta(x)\right].$$

**This has to look better and be more 'algorithmic'.** As part of our analysis we

**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 $\theta(X_i)$.

will compare the results to check whether the relationship is indeed linear or whether we discover non-linear heterogeinities that the linear DML approach does not account for and have not been considered in literature yet. However, note that when using a nonparametric second stage the convergence rate of the estimator declines. While still achieving faster rates than most other nonparametric estimators, this implies that the Causal Forest based approach is more demanding when it comes to the number of observations.

# 5 Estimation and Results

$$\Delta C_{it+1} = \theta(X_{it}) R_{it+1} + g(X_{it}, W_{it}) + \epsilon_{it} \tag{7}$$

$$R_{it} = h(X_{it}, W_{it}) + u_{it} \tag{8}$$

where $\Delta C_{it+1}$ is change in consumption, $R_{it+1}$ is the amount of rebate received by the household and $g(X_{it}, W_{it})$ and $h(g(X_{it}, W_{it})$ are non-parametric functions of confounders. $X_{it}$ and $W_{it}$ are distinct by the assumption that only $X_{it}$ influences the marginal effect of the rebate, $\theta(X)$, while $W_{it}$ denotes the set of confounders that play no role in the effect. **or make the following a completely new section?**

# 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. Similarly to prior contributions, we also looked at 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 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 in 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 analyse the relationship between prediction and feature. Feature is a different term for control variables. In our setting these are the variables we condition on to find the CATE, i.e. variables contained in $X$. Since variables in $W$ are assumed to not impact the CATE they are not contained in the second stage and therefore play no role in predicting individuals' MPC. Machine Learning estimators such as random forests are blackboxes as they only provide predictions but stay quiet on which variables are important to arrive at this prediction. The literature has proposed multiple approaches that help quantify the role of a single feature, some of which we look at in the following.

## 6.1 Marginal and Partial Dependence Plots

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 individual 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 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

$$\hat{f}_M(x_S = v_j) = \int p_{X_C|x_S=v_j} m(x_S = v_j, X_C) dX_C, \tag{9}$$

where $m$ is our predictor and $\hat{m}_M(v_j)$ is the effect at $x_S = v_j$. On the other hand, PDPs use the marginal distribution of $X_C$, $p_{X_C}$,

$$\hat{f}_{PDP}(x_S) = \int p_{X_C} f(x_S, X_C) dX_C \tag{10}$$

where $\hat{m}_{PDP}(v_j)$ is the Partial Dependence of our predictor on $x_S$ at $v_j$. Using the marginal distribution of $X_C$ effectively "marginalizes out" the effect of any other variables than $x_S$ at some point $v$ and therefore reveals what impact $x_S$ has on our prediction at this point. 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 simply example. Lets 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, do 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 effectviely create observations that are extremely unlikely or even impossible to observed in the real world because of the correlation 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 weak 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 results in linear Partial Dependence Plots. **Remember that in our linear DML approach, we assume that the CATE we estimate is linear in features $X$ and the second stage - the fitted model we actually investigate here - is simply a linear regression, which results in a linear PDP by construction as our predicted MPC is simply the sum of all coefficients for individual $i$ given their characteristics. $\rightarrow$ I am not so sure about this part yet**

Indeed, results of the partial dependence plots are rather spurious. They are reported in more detail in Appendix X.X, where we look at the PDPs for non-durable consumption. The effects have a high variation and point estimates out of line of the existing literature. **this is not a good reasoning of why I don't show them because they are too close to the ALEs - I guess**

## 6.2 Accumulated Local Effects

To circumvent issues arising in M-Plots and PDPs from correlated features, Apley and Zhu (XXX) 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. As with M-Plots we use the conditional distribution to bypass the extrapolation issues that PDPs suffer from. 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. We apply this by using the partial derivative of our predictor $m$ with respect to $x_S$ at the point $v_j$. Although many machine learning methods such as tree based learners have no concept of a gradient, Apley and Zhu are able to derive proofs for non-differentiable

functions $m$ (see Section X.X) and further does this not play a role when it comes to the ALE estimation. The ALE is then defined as

$$\hat{f}_{S,ALE}(x_S) = \int_{z_{0,S}}^{x_S} E_{X_C|X_S=x_S}[\hat{f}^S(X_S, X_C)|X_S = z_S]dz_S - constant, . \tag{11}$$

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, but we are interested in how $x_S$ affects our predictions and not how it affects changes in predictions. Thus, we simply integrate over all local effects up to $x_S = v_j$, where $z_{0,S}$ is the lower bound of the distribution of $x_S$.
To estimate the ALE we use the following estimator, which illustrates the procedure in more intuitive terms:

$$\hat{\tilde{f}}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_{i,j} \in N_j(k)} [f(z_{k,j}, x_{i,\setminus j}) - f(z_{k-1,j}, x_{i,\setminus j})] \tag{12}$$

The intuition behind the estimator is straightforward. 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 11 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 quantile $k$. We then averages over all individuals that fall within this bin $k$ and finally accumulate all predicted differences from the lowest bin up to bin $k$. Only looking at individuals within a neighborhood $N(k)$ - effectively observations in the same quantile - accounts for the conditional distribution used in 11.
As a last step, Apley and Zhu propose to center the effect around the average of all ALEs such that the mean effect is zero. Thus, 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 (11) is replaced by

$$averagetermhere$$

.

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 to understand how these predictions are achieved, but there is no notion of statistical significance in these settings. Therefore, a specific approach to quantify uncertainty of these measures has not yet been developed. A deeper look into this topic is, however, out of the scope of this work. To briefly dive into the topic of statistic significance, we use a bootstrapping based approach. We simulate the ALEs for $n_{bootstrap}$ samples. These create an empirical distribution (need at least e.g. 500-1000) on which basis we calculate pseudo-standard errors. Figure X.X reports the Confidence Intervals using the reverse percentile approach (see Appendix X.X) and the mean point estimates in each bin. 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 way. While the ALEs show us the relationship between a specific variable and our predictions, understanding whether this relationship is statistically significantly different from playing no role would be a major improvement.

One weakness of ALE is further that they are a global measure, i.e. they do not help to uncover heterogeneity in the reaction of the prediction to a single variable. One method that can unbeil such heterogeneities are Individual Conditional Expectations, but they suffer from the same conceptual problems as Partial Dependence Plots. Next to investigating the role of uncertainty in the measures presented in this section we therefore also urge the development of theoretical foundations of such a measure. For now, we account for heterogeneity by plotting the unaveraged ALEs of each individual. To avoid overplotting, we only plot the households at the quintiles of the ALE distribution. (**delete last part of this if not doing it!!**)

### 6.3 Results

Figure X.X plots the Accumulated Local Effects.

## 7 Conclusion

(from intro 2 draft but doesn't fit that much anymore) Our contribution is twofold: for one, we estimate the conditional MPC out of the tax stimulus in the most precise and rigorous manner thus far. Second, we use an estimator that exploits the power of machine learning methods for causal inference and contribute to the wider understanding and promoition of this method among applied researchers. Machine Learning predictors are powerful tools

when it comes to handling large data and/or complex relationships between variables without any specification of those.

# 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.

Fagereng, A., M. B. Holm, and G. J. Natvik (forthcoming): "MPC Heterogeneity and Household Balance Sheets," *American Economic Journal: Macroeconomics*.

Golosov, M., M. Graber, M. Mogstad, and D. Novgorodsky (2021): "How Americans Respond to Idiosyncratic and Exogenous Changes in Household Wealth and Unearned Income," Working Paper 29000, National Bureau of Economic Research, series: Working Paper Series.

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.