Detecting Heterogeneity in the MPC: A Machine Learning Approach

Last Update: 21.08.2021

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

Introduction

The Marginal Propensity to Consume (MPC) is at the centre of the macroeconomic model introduced by John Maynard Keynes in his General Economic Theory. Eversince its introduction, the role and size of the MPC has been subject to debate. While Keynes declared the MPC to be meaningfully different from zero, the permanent income hypothesis developed by Milton Friedman and a corner-stone of modern macroeconomics declares it to be irrelevant to current consumption decisions and, thus, irrelevant to economic policy making. However, both are wrong and right at the same time. More recently, the focus of research concerned with undertanding the MPC to guide policies - such as stimulus payments - has shifted to painting a more diverse picture of households' willingness to spend out of a transitory income shock. New, sophisticated models formalize the heterogeneity of agents in the macroeconomy, including their MPC. Additionally, empirical work has shifted from trying to prove an MPC of zero - or the opposite - to understanding the difference across households and allowing for heterogeneity in the MPC. **add two examples of channels - liquidity constraint and ...**

Using the 2008 tax stimulus as an exogenous income shock, my contribution to the empirical literature is twofold: First, I use a new and highly flexible estimation approach, that allows me to identify a wider range of heterogenous effects. The so-called Double Machine Learning Approach allows for a semi-parametric setup in which the functional form of any confounding factor does not have to be specified. Second, the literature using the data from the 2008 stimulus (or the general literature? **check!**) so far has investigated its effect (poor) methods that lack the advantage of the DML approach while at the same time not allowing to account for the panel data setting of the data. These approaches implicitly impose a strict exogeneity condition, while the Panel DML model is capable of accounting for possible effects of past characteristics on the change in current income. Therefore, I am able to identify the causal effect of the tax rebate on consumption change more clearly (or: actually identify it, but maybe to harsh).

I rely on data collected by the Consumer Expenditure Survey (CEX), which included a special part in the 2008 and 2009 surveys dedicated to the tax stimulus. This effort was promoted by Johnson, Parker and Souleles (2013; henceforth JPS) who quantify the effect of the tax stimulus on consumption changes. The data they use is publicly available and also used by Misra and Surico (2014; henceforth MS). Hence, to improve comparability with two of the more recent and prominent contributions, I use the data provided publicly by JPS as well. While both document some heterogeneity in the MPC, there are several drawbacks in their respective analysis. Meanwhile, the DML estimation allows me to identify household level point estimates and standard errors, allowing me quantify whether the estimated MPC is significantly different from zero for each individual to un-

cover which households actually experience a temporary increase in consumption due to a temporary income shock (**rephrase**).

rewrite and put this somewhere else Understanding which underlying factors drive heterogeneity in the MPC is crucial for policy makers. While short-term untargeted tax-stimuli such as the one in 2008 are reasonable in times of economic crisis when time is short, targeted stimuli can improve the payoff of each dollar invested into an economic stimulus.

add a brief summary/overview of what I find

The rest of the paper is structured as follows: Section 2 summarizes the theoretical and empirical literature on MPC heterogeneity putting a focus on the issues concering JPS and MS analysis. Section 3 discusses the data source and challenges connected with it. The empirical methodoogy I use is described in Section 4, while Section 5 presents the results. Section 6 concludes.

Literature

The literature investigating the size of the MPC and potential heterogeneity can be summarized in three different strands.

The first one uses quasi-experimental settings to identify income shocks and the resulting reaction of consumption. Settings considered are US tax stimulus progrmas during the times of economic crisis in 2001 and 2008 or lottery wins by individuals. Johnson, Parker and Souleles (2006) estimate the size of the MPC out of the 2001 tax stimulus and in a more recent contribution also take a look at the 2008 tax rebate program (Johnson, Parker and Souleles, 2013). The latter is closely related to our procedure and is hence discussed in more detail further below. They find XXX. Meanwhile, Fagereng et al. (2020) estimate the heterogeneous MPC out of lottery winners in Norway. Golosov et al. (2021) do the same using bla data. The second strand of literature uses self-reported MPC from household surveys. However, studies based on self-reported data are prone for measurement error - specifically the so-called self-report bias which leads respondents to misreport their data. In the case of the Marginal Propensity to Consume we expect this to be even larger than in the survey data exploited in quasi-experimental settings since respondents do not only have to document their raw spending behaviour (e.g. indicating how much money was spent in total) but assess their MPC on their own. Such calculations are likely to increase the risk of measurement error, especially the more abstract the concept becomes. There is also a more theoretical side to the discussion focusing on calibrating heterogenous agent new keynesian models (HANK) to uncover general equilibrium effects of single agents' MPCs on the aggregate MPC out of income shocks. Finally, there are two contributions that are by default most closely related to our setting since we make use of

the same data. Namely, these are Johnson, Parker and Souleles (2013) and Misra and Surico (2014). They estimate a simple fixed-effects regression in which they interact their income schock variable with pre-defined dummies. Those dummies are based on continuous variables and created by choosing discrete cut-off points. However, this prohibits the detection of heterogeneous patterns that are not captured by the variables considered or are not inside the defined thresholds. Using the Parker et al. data, Misra and Surico (2014) replicate their approach but use quantile regression to analyse the heterogeneity in the MPC distribution. While quantile regression can be of service to detect heterogeneity in coefficients, it does not allow for the correct interpretation. The treatment effects they uncover are the effect of the income shock on the difference in consumption before and after for a respective quantile. However, this quantile does not need to include the same individuals. Hence, the quantile regression only uncovers shifts in the overall distribution but is silent on how specific individuals changed their consumption pattern - and hence the actual MPC.

Lastly, as Kaplan and Violante **(or who exaclty was it?)** point out, empirical analysis that use stimulus payments as a temporary income shock to identify the MPC might actually estimate another coefficient, which they coin the coefficient of rebate. They argue that the conditions of a stimulus payment as well as the overall economic conditions that lead to such a payment are too specific (**rephrase**) to

Methodology

rewrite this intro and specify that this is a two-stage approach

To identify the causal effect of receiving the tax rebate on households' consumption changes, I use the Double Machine Learning approach (DML) developed by **?** and extended for Panel Data settings by **?**. More precisely, this approach estimates a Partially Linear Model (PLM) of the form

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

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

where Y_{it} is the outcome and the goal is to estimate the conditional treatment effect $\theta(X)$ of treatment T_{it}. The functions $g(X_{it}, W_{it})$ and $h(X_{it}, W_{it})$ are some non-parametric functions. Hence, the DML approach has the advantage that the effects of the confounders on treatment and outcome do not have to be formalized into a specific functional form. To remove these effects the DML estimator suggests a two-stage approach to orthogonalize treatment and outcome with respect to the confounders and uncover the causal effect of the treatment on outcome. Orthogonalization removes any variation in the two variables that

is due to the confounders (X_{it}, W_{it}) by removing the conditional mean of the respective variable. The orthogonalized version of (**??**) is then

$$\Delta C_{it} - E[\Delta C_{it}|X_{it}, W_{it}] = \theta(X_{it})(R_{it} - E[R_{it}|X_{it}, W_{it}]) + \epsilon_{it} \tag{3}$$

and I denote

$$E[R_{it}|X_{it}, W_{it}] = h(X_{it}, W_{it}) \tag{4}$$
$$E[\Delta C_{it}|X_{it}, W_{it}] = f(X_{it}, W_{it}). \tag{5}$$

The advantage of the DML approach is that the two conditional means can be estimated by any machine learning method, hence guaranteeing strong flexibility of the estimation. At the same time, the DML's asymptotic properties are outperfoming other non-parametric methods in terms of the rate of consistency making it less data-hungry than standard econometric approaches. In my case, I use a random forest to predict the first stage functions $\hat{f}(X_{it}, W_{it})$ and $\hat{g}(X_{it}, W_{it})$. The random forest has proven reliable and efficient in a wide variety of prediction tasks without making any assumptions on the functional form. Hence, contrary to the existing literature, I capture any interactions and power series of the confounders that affect the treatment or outcome variable.
Once the first stage estimation

$$\tilde{Y}_{it} = Y_{it} - \hat{f}(X_{it}, W_{it}) \tag{6}$$
$$\tilde{R}_{it} = R_{it} - \hat{h}(X_{it}, W_{it}) \tag{7}$$

Panel DML Recipe

Algorithm 1 Panel DML Recipe

[1] Partition the data into K-folds based on their time index. An observation is added to partition I_k if:

$$I_k = \{(i,t) : T(k-1)/K + 1 \leq t \leq Tk/K\} \tag{8}$$

For each partition k use a first stage estimator to estimate $\hat{d}_k$ and $\hat{l}_k$ using data of all folds except k (cross-fitting). Orthogonalize treatment and outcome of observation (i,t) using the predictions of the corresponding fold to get the residuals. Use all residuals to estimate the coefficient θ using a suitable estimator.

Estimation and Results

$$\Delta C_{it+1} = \theta(X_{it})R_{it+1} + g(X_{it}, W_{it}) + \epsilon_{it} \tag{9}$$

$$R_{it} = h(X_{it}, W_{it}) + u_{it} \tag{10}$$

where ΔC_{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.