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

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

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

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

Cross-Fitting

Algorithm 1 summarizes the Panel Double ML recipe including the cross-fitting method introduced in **?** to account for the panel structure of the data, allowing the assumption of conditional sequential exogeneity instead of strict exogeneity.

Estimation and Results

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

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

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.