Detecting Heterogeneity in the MPC: A Machine Learning Approach

Lucas Cruz Fernandez

last updated: 13.08.2021

Introduction

The goal of this paper is to understand what drives differences in the marginal propensity to consume (MPC). The MPC is subject to debate at least since none other than Keynes put it at the center of his macroeconomic analysis. While Keynes declared the MPC to be meaningfully different from zero, the permanent income hypothesis developed most prominently by Friedman declares it to be irrelevant and zero as households do not react to transitory income shocks with respect to their current consumption. 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. This paper is part of this strand of literature as I document heterogeneity in the MPC that depends on the households characteristics.

The idea of investigating the heterogeneity in MPC is not new. Over the last decade the development of sophisticated heterogeneous agent models - known as HANK - also sparked an empirical investigation of heterogeneity in the MPC. While several papers have examined the heterogeneity in MPC, they lack the use of sophisticated methods to detect heterogeneous patterns that are not somehow pre-determined. For example, one of the earlier contributions is Parker et al. (2013) who use data from the 2008 Consumer Expenditure Survey (CEX) to analyse the effect of receiving a tax stimulus paid by the government on the consumption change. 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 inter-

pretation. 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.
In this paper, I use the CEX dataset provided by Parker et al. to detect heterogenous patterns that are not predefined and actually uncover the MPC for each individual in the survey. Applying a Double/Debiased Machine Learning estimator developed by Chernozhukov et al. (2016) I am able to estimate individual level treatment effects conditional on a household's characteristics. Moreover, I can quantify whether this estimated MPC is significantly different from zero for each individual to uncover which households actually experience a temporary increase in consumption when receiving an income shock.
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

To detect heterogeneity in the MPC I apply the Double/Debiased Machine Learning method (DML) developed by Chernozhukov et al. (2016). This approach is also known as Orthogonalized Machine Learning as it relies on the idea of orthogonalizing the dependent variable and the treatment/variable of interest to establish the causal relationship between the two. More precisely, I model the relationship between treatment and outcome with a Partially Linear Model (PLM):

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

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

This semi-parametric model allows to define the following equation:

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

where we know that

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

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

These two conditional expected values can be estimated using any machine learning method of choice and then used to calculate the respective residuals

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

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

This orthogonalization removes any variation in the treatment and outcome that stems from any of the observed confounders. Hence, regressing $\tilde{Y}_{it}$ on $\tilde{R}_{it}$ shows the effect of the treatment on the outcome without any interference from confounders, which allows to identify the causal effect from the treatment on the outcome.
In my setting the treatment - i.e. the amount of rebate received - depends heavily on observables such as the number of children because they determined the rebate amount directly. However, what is not observed is the individual's net-tax liability which determined the base size of the rebate a household received. This varies on the individual level and generates exogenous variation in the treatment. Additionally, the timing of the rebate was naturally randomized as checks were sent out based on the ending digits of individuals social security number, which is randomly generated. These two sources of exogenous variation allow to identify the causal relationship between rebate amount and change in consumption.

Panel Structure

The original DML approach is only valid for estimating cross-sectional data, i.e. under the assumption that (X_i, W_i, Y_i) are iid. However, in the setting at hand this is not case as I have a panel structure. In the traditional fixed effects setting all observations are demeaned by the individual level mean to remove the fixed effect and then OLS is run on the residuals, which are iid. Hence, in theory when data is pre-processed in the correct manner, the observations in my setting are iid as well and the DML approach will work fine.
One advantage is that the dependent variable is the change in consumption, hence any individual level fixed effects that influence consumption are removed and do not have to be accounted for. Additionally, the data is driven by seasonality. Adjusting for that gets more tricky.