Detecting Heterogeneity in the MPC: A Machine Learning Approach

Last Update: 05.09.2021

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

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 - a corner-stone of modern macroeconomics - declares it to be irrelevant to current consumption decisions and, thus, irrelevant to short-term economic policy making. However, both are wrong and right at the same time. Both are wrong and right at the same time.

The more recent literature has shifted the focus from assessing the reaction to income shocks across all households to a more nuanced view that takes into account heterogeneity across households. This follows the overall trend in macroeneomic theory focusing on heterogeneous agents that has taken place over the course of the last two decades. The theoretical as well as empirical literature is now working on improving the understanding what drives a households reaction to an income shock and whether there are substantial differences across these reactions.

Similar to previous quasi-experimental contributions to the empirical literature I use the 2008 U.S. tax stimulus program to quantify the MPC. The data is taken from the Consumer Expenditure Survey (CEX), which also collected information on timing and size of the rebate households received. This effort was made thanks to **add correct paper here (Parker et al 2013 or something)** who added those question to the regular questionnaire of the CEX. We use the same cleaned dataset as Parker et al. (2013) and Misra and Surico (2014) to compare our results to two of the most prominent papers investigating the heterogeneity of the MPC using the tax stimulus. However, compared to these two already existing works, our estimation procedure is less restrictive and far more rigorous. More precisely, we use a Double Machine Learning (DML) approach capable of detecting non-linear heterogeneities and controlling for confounders without any assumptions on their relationship with consumption and the rebate. Hence, 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.

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. 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. It is important to hihglight that in our setting, households experience an anticipated - at least for most - positive income shock. 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 citizens (**maybe citizens sounds like only those got the rebate, which is not true 100%**) 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). 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.

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 4 discusses the data source and challenges connected with it. The empirical methodology we use is described in Section 3,

while Section 5 presents the identification and estimation results. Section 6 concludes.

2 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.

However, what becomes evident in all strands of the existing literature is the important role liquidity and the size of the shock plays in households' response.

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 et al. (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

In the JPS and MS specifications, they also consider simple linear estimators (OLS and Quantile Regression) that imply the assumption of strict exogeneity. Since we are looking at quarterly data and JPS/MS only consider age and change in the size of family as confounders, one could argue that there is little to no variation in these variables between quarters. In that case,

3 Methodology

3.1 Notes on Methodology for writing

- see: `https://econml.azurewebsites.net/_autosummary/econml.dml.DML.html#econml.dml.DML`
- how is the CATE achieved in second stage?
  - second stage in Linear DML is OLS regression
  - CATE is achieved through interaction of some mapping $\phi(X)$ of the confounders X with the 'base' treatment effect Θ.
  - effectively running an OLS regression with interaction terms $\tilde{D} \otimes \phi(X)$
  - hence, assume that CATE is linear in X unless using polynomial mapping $\phi(X)$
  - nonparametric DML such as Causal Forest DML circumvents this assumption and detects non-linear heterogeneity in case there is any (run this as kind of robustness check/better specification)
- causal forests/generalized random forests:
  - random forests only allow for prediction, no inference possible
  - Athey and Wager (2016) and Athey, Tibshirani and Wager (2019) develop causal forests/grf
  - these allow to estimate any desired local moment equation - e.g. those of a treatment effect analysis
  - more specifically, in the DML case the moment equation estimated is

$$E\big[\left(Y - E[Y|X,W] - < \theta(X), T - E[T|X,W] > -\beta(x)\right)(T;1)|X = x\big] = 0$$

  - I should push this explanation into the Appendx though and just explain the GRF in more general terms (it detects heterogeneity that is not specified before)

To identify the causal effect of the income shock on households' change in consumption, we use the Double Machine Learning (DML) estimator developed by Chernozhukov et al. (2017). This rather new kind of estimator allows to efficiently estimate semi- or non-parametric models of treatment effects. The DML estimator 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 identifies the pure effect of treatment on the outcome. Meanwhile, its implementation procedure deals with common biases arising in more naive estimation procedures that employ Machine Learning methods, opening the door to making use of these spohisticated methods in causal inference studies. Even in settings in which the raw predictions of the ML alogrithms used are not of high quality, the estimator yields desirable results and properties (**rephrase this last sentence again**).
From a more theoretical perspective the DML estimator also yields very efficient properties when it comes to its asymptotic analysis, especially a rate of convergence that is faster than other nonparametric estimators. Under certain assumption, Chernozhukov et al. (2017) are able to prove root-n consistency of the estimator. However, we will not further elaborate on these lattertechnical details but rather focus on how the estimator works in general. For a more technical discussion the reader is referred to Chernozhukov et al. (2017) and Semenova et al. (????). Instead, this section introduces the general idea behind the DML estimator as well as the different variants we will use in Section 5.

3.2 Idea behind DML

We start with considering a Partially Linear Model of treatment and outcome. Note that the DML estimator is capable of estimating various models, but we will also use a PLM specification later on. Additionally, it helps to intuitively understand the main idea of the DML estimator and its mechanics. Section 3.5 will briefly present what a fully non-parametric approach looks like.
The Partially Linear Model (PLM) is given by

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

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

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), which in Rubin's **missing citation here** potential outcomes framework 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 constant, hence $\theta(X)$ represents the marginal CATE

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

The marginal CATE measures how much a marginal increase in the continuous treatment changes the outcome, conditional on the individual having a set of characteristics $X = x$. The DML now follows a two step procedure to identify $\theta(X)$. 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). We can use these two conditional expectations to show that the PLM can be boiled down to

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

This orthogonalization of treatment and outcome guarantees that the regression coefficient $\theta(X)$ only captures variation that the treatment invokes in the outcome and is free of any confounders acting through treatment D. Moreover, as Chernozhukov et al. (2017) show, simply plugging in estimates of f and h into the PLM results in a regularization bias when estimating $\theta(X)$ that does not vanish asymptotically, effectively prohibiting consistency of the estimator. It is circumvented by the orthogonalization of treatment and outcome, while there is no difference in orthogonalization and controlling for confounders.
(Frisch-Waugh-Lovell Theorem kind of, also potentially missing details on orthogonalization)
The first stage of the estimation process consists of choosing an appropriate Machine Learning method and finding estimates of the conditional expectation functions f and h. 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.
Once we obtain the first stage predictions $\hat{f}$ and $\hat{h}$, we use them to orthogonalize treatment and outcome to retrieve the residuals

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

The second stage then only consists of a linear regression of $\tilde{Y}_{it}$ on $\tilde{D}_{it}$ that yields $\hat{\theta}(X)$. More precisely, the partially linear model we consider here implicitly assumes a parametric

form of the CATE

$$\theta(X) = \phi(X) \times \Theta,$$

where Θ is the base treatment effect and $\phi(X)$ is a mapping of confounders X. In practice, the estimator boils down to a linear regression which includes interaction terms $\tilde{D} \otimes \phi(X)$. The mapping $\phi(X)$ can take any parametric form we might think of. In Section 3.5 we discuss how this assumption can be avoided, allowing to detect any heterogeneity without pre-defining its functional form.

3.3 Cross-Fitting

The DML estimator achieves its desirable consistency results by avoiding two common biases arising in settings that employ Machine Learning estimators: overfitting and regularization bias. We have already briefly discussed that the regularization bias is avoided through orthogonalization. However, the overfitting problem needs to be adressed on its own.

Overfitting is the result of an estimator adjusting too much to the given data such that when it is exposed to new data it performs very badly resulting in a high variance of this estimator. Again, Chernozhukov et al. (2017) show how this variance term does not vanish in the asymptotic analysis leading to an inconsistent estimator. However, this issue can be avoided using a technique they coin *cross-fitting*. 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 estimate the first stage estimators. Those are used to predict the values in the second sample, which are then subsequently used for orthogonalization and the second stage estimation. 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 the baseline DML estimator overall.[1]

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

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 $\theta(X_i)$.

3.4 Panel DML

So far we have considered the original DML estimator that relies on the assumption of strict exogeneity. While even in the panel setting at hand this assumption might be reasonable, Semenova et al. (????) propose another DML estimation method that relaxes it. More precisely, for their estimator to be consistent we only have to assume conditional sequential exogeneity, which enables us to control for panel dynamics in a more precise manner. We lay out why the assumption of strict exogeneity might be reasonable in Section 2 and will compare these arguments to assuming conditional sequential exogeneity in Section 5.
More formally, in the Panel DML setting we assume

$$E[\epsilon_{it}|X_{it}, W_{it}, \Phi_t] = 0$$
$$E[u_{it}|X_{it}, W_{it}, \Phi_t] = 0,$$

where Φ_t is the information set in period t. Semenova et al. (????) show in a setting with low-dimensional treatment, this assumption still results in the same second-stage estimator as the original DML. The only difference in the estimation procedure is the cross-fitting algorithm in the first stage. We form its folds based on the time index instead of simply randomly splitting up the sample. Moreover, we can include lagged values of treatment and outcome to account for the information set Φ_t. As discussed in Section 2 including lagged treatment actually improves the proper identification of the MPC.

3.5 Nonparametric DML

In the PLM setting, the functional relationship how the CATE is influenced by confounders X via the mapping $\phi(X)$. However, the DML estimator also enables us to use a nonparametric approach that can detect any interaction between treatment and confounders to uncover heteroegeneity. It has the same first stage, but estimates the second stage using the Causal Forest estimator proposed by Athey, Tibshirani and Wager (????).

The Causal Forest is a generalization of the Random Forest prediction method developed by Breimann (2001), which has found application in a wide array of predictive tasks. However, the original algorithm - as most Machine Learning methods focusing on prediction - does not allow for any causal inference. The Causal Forest solves this problem by generalizing the objective function to fit the potential outcomes framework and developing 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. Using the CF as a second stage enables us to estimate the model

$$Y_{it} = g(D_{it}, X_{it}, W_{it}) + \epsilon_{it}$$
$$D_{it} = m(D_{it}, X_{it}, W_{it}) + u_{it}.$$

As part of our analysis we will compare the results to check whether the relationship is indeed linear or whether we discover non-linear heterogeinities that the PLM 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.

4 Data

5 Estimation and Results

$$\Delta C_{it+1} = \theta(X_{it}) R_{it+1} + g(X_{it}, W_{it}) + \epsilon_{it} \tag{5}$$
$$R_{it} = h(X_{it}, W_{it}) + u_{it} \tag{6}$$

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.

6 Conclusion

References

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.

Semenova, V., M. Goldman, M. Ai, V. Chernozhukov, M. Taddy, and M. Ai (????): "Orthogonal Machine Learning for Demand Estimation: High Dimensional Causal Inference in Dynamic Panels," 66.