



UNIVERSITÀ POLITECNICA DELLE MARCHE
Repository ISTITUZIONALE

No such thing as the perfect match: Bayesian Model Averaging for treatment evaluation

This is the peer reviewed version of the following article:

Original

No such thing as the perfect match: Bayesian Model Averaging for treatment evaluation / Lucchetti, R., Pedini, L., PIGINI, C.. - In: ECONOMIC MODELLING. - ISSN 0264-9993. - ELETTRONICO. - 107:(2022).
[10.1016/j.econmod.2021.105729]

Availability:

This version is available at: 11566/294042 since: 2024-09-24T10:54:03Z

Publisher:

Published

DOI:10.1016/j.econmod.2021.105729

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. The use of copyrighted works requires the consent of the rights' holder (author or publisher). Works made available under a Creative Commons license or a Publisher's custom-made license can be used according to the terms and conditions contained therein. See editor's website for further information and terms and conditions.

This item was downloaded from IRIS Università Politecnica delle Marche (<https://iris.univpm.it>). When citing, please refer to the published version.

(Article begins on next page)

No Such Thing As The Perfect Match: Bayesian Model Averaging for Treatment Evaluation

Riccardo (Jack) Lucchetti, Luca Pedini, Claudia Pigni *

*Dipartimento di Scienze Economiche e Sociali (DiSES)
Università Politecnica delle Marche, Ancona (Italy)*

Abstract

Propensity Score Matching is a popular approach to evaluate treatment effects in observational studies. Regrettably, practitioners often overlook the issue of model uncertainty and its consequences when building the propensity score model. We tackle this problem by Bayesian Model Averaging (BMA) with an application to the 2014 Italian tax credit reform (the so-called “Renzi bonus”). Model uncertainty has a dramatic impact on the estimated treatment effects: using standard model selection procedures may lead to choosing equally defensible models that, however, produce substantially heterogeneous results. By using BMA-based estimates, we find a much more coherent picture: significant effect of the rebate on food consumption only for liquidity constrained households, in line with most recent literature.

Keywords: 2014 Italian Tax Credit Reform, Bayesian Model Averaging, Model Uncertainty, Propensity Score Matching, Reversible Jump Markov Chain Monte Carlo, Tax Rebate Policies

*Corresponding author

Email address: c.pigni@univpm.it (Claudia Pigni)

We would like to thank Giuseppe De Luca, the participants to the 2019 Gretl Conference (Naples) and the participants to the 9th ICEEE conference (Cagliari) for their useful remarks. We also thank the editor and two anonymous referees for their observations. Of course, all remaining errors are our responsibility.

1. Introduction

Propensity Score Matching (PSM) (Rosenbaum and Rubin, 1983) is a standard tool for treatment evaluation in observational studies and it has been used in a wide range of economic applications. The centrepiece of PSM is a binary model for the Propensity Score (PS), namely the conditional probability of being treated given a chosen set of characteristics; unfortunately, practitioners often build this binary model with little attention to the choice of explanatory variables, despite the fact that the selection of controls can dramatically affect the matching and, as a result, the estimate of the causal effect (Heckman et al., 1997; Caliendo and Kopeinig, 2008).

From a theoretical standpoint, the consequences of model selection can be assessed in terms of the so-called *bias-variance trade-off*. Using all the available covariates in the binary model may be tempting, to gain model flexibility and reduce the approximation bias. However, over-parametrisation may lead to violations of basic assumptions of PSM, such as the common support and balancing conditions, as well as to overfitting and to a larger variance of the estimator. Parsimonious models, on the other hand, may produce an inconsistent treatment effect estimator because of relevant omitted variables (Heckman et al., 1997).

In some cases, economic theory may provide some background information to guide the practitioner in the choice of covariates, albeit on a case-wise basis and depending on the specific treatment analysed. Therefore, more general strategies have been put forward by the literature to help overcome this impasse. Rosenbaum and Rubin (1984) suggest using the PS along with stratification matching and adjusting the set of covariates with interactions or higher order polynomials, as long as the balance conditions are met. The same recommendations are made by Dehejia and Wahba (1999). Rubin and Thomas (1996) point out that discarding covariates should be always a mindful decision motivated by the balancing conditions. Millimet and Tchernis (2009) emphasise instead the advantage of estimating redundant models over parsimonious ones whenever the bias reduction is larger than the increase in variance, arguing that overfitting has no negative consequences.

Bryson et al. (2002) warn on the opposite problem of including too many variables, as efficiency would be often affected in empirical applications. In addition, they also question the quality of the resultant matching when redundant variables are included. There is also a related stream of literature dealing with treatment effects evaluation in cases where the number of covariates is larger than the number of observations and where model selection is tackled via the *sparsity* assumption and LASSO-like estimators (Belloni et al., 2014, 2017; Chernozhukov et al., 2018). Finally, Brookhart et al. (2006) take an intermediate position between parsimonious and inclusive specifications: variables that are unrelated to the treatment assignment but related to the outcome should always be included, whereas the converse is not necessarily true.

In our view, the lack of a unique modelling strategy for the PS is a consequence of the impossibility of choosing the *true* model specification, especially in the light of the fact that empirical analysis entails, in practice, a large number of decisions that have to be made about the problem at hand, each of which could steer the results in a completely different direction, as convincingly shown by Huntington-Klein et al. (2021). As a consequence, a problem of *model uncertainty* (Chatfield, 1995) is introduced by conditioning the analysis on a single model specification, under the assumption that such specification is the true one.

However, model uncertainty can also be dealt with by averaging estimates from a range of different models, using weights that reflect the models' goodness of fit. Model averaging can be performed either in a frequentist or a Bayesian framework, where the latter provides an immediate interpretation of weights as model probabilities and a direct measure of the relative importance of covariates in the model specification. Moreover, computational techniques based on Markov Chain Monte Carlo (MCMC) samplers make Bayesian Model Averaging (BMA) computationally feasible even in high-dimensional model spaces. See Steel (2020) for a detailed discussion on the advantages of BMA over the frequentist model averaging approach.

The usage of BMA and related methods in applied economics varies widely by field: for example, in the empirical literature on economic growth BMA has become a relatively

standard tool (see *e.g.*, Brock and Durlauf, 2001; Fernandez et al., 2001b; Sala-i-Martin et al., 2004; Horvath, 2011; Eris and Ulasan, 2013; Man, 2015). In the context of policy evaluation, notable contributions include Brock et al. (2003); Brock and Durlauf (2015).

In this paper, we propose to address the issue of model uncertainty in PSM by employing a BMA approach that improves upon previous proposals (see *e.g.*, Kaplan and Chen (2014); Zigler and Dominici (2014)) and is highly accurate and computationally efficient. We employ the Reversible Jump MCMC (RJMCMC) sampler (Green, 1995; Holmes and Held, 2006; Lamnisos et al., 2009), where models and their parameters are sampled simultaneously, and the PS distribution along with them. While BMA is routinely used with linear models, more solutions for more general cases, such as binary dependent variable models, have been devised only recently. To this end, we use the `gretl` package `ParMA` (see Lucchetti and Pedini, 2021), in which the RJMCMC sampler can be applied to Generalised Linear Models, thereby making it possible to perform full BMA on a proper binary model for the PS instead of relying on the approximation provided by a linear probability model. Therefore, we frame PSM as a three-step procedure, where the three steps are (a) the PS model (b) matching and (c) treatment effect evaluation. In this way, the model uncertainty issue directly affects only the PS and the related model probabilities can be attached to each treatment effect estimator.

Our proposed method is used for evaluating the economic impact of the 2014 Italian tax credit reform (Decree Law 66/2014), also known as “Renzi bonus”, which introduced a monthly wage increase of about €80 for all employees with an annual gross income between €8145 and €26000. We use data from the Survey of Household Income and Wealth (SHIW) issued by the Bank of Italy and estimate the treatment effect of the tax rebate on different types of monthly consumption, such as total, food, nondurables, car-related, and other durables consumption. In line with the results provided by Neri et al. (2017), who evaluated the effect of the Renzi bonus using the same data, we find a role of liquidity constraints, as the treatment effect is significant only on the food consumption of households who report greatest difficulty to make ends meet. In addition, we show that accounting for model uncertainty leads to different evidence than that produced by

a standard model selection procedure, which would point toward an increase in food and other nondurables consumption for all households.

The rest of the paper is organised as follows: in Section 2 we briefly introduce the mechanisms driving the consumption effects of tax rebate and illustrate the Italian 2014 policy intervention; Section 3 describes the proposed BMA for PSM; Section 4 introduces the SHIW dataset and reports the relevant descriptive statistics; Section 5 provides the estimation results; Section 6 discusses the results and contains the final remarks.

2. Background

Tax rebate policies are common stabilising instruments applied by policymakers to reduce the impact of the business cycle: these fiscal interventions are aimed at increasing household spending, thus countering the negative effect of recessions. Yet the effectiveness of tax credit policies is widely debated (Shapiro and Slemrod, 2003a,b, 2009). On the one hand, the permanent income hypothesis predicts that transitory policy interventions should not affect consumption: if individuals smooth consumption over their lifetime, a tax rate is effective only as long as is perceived as permanent. Moreover, the intervention might be anticipated, which makes it more difficult to analyse if consumers adjust their behaviour before actual implementation. On the other hand, if households face liquidity constraints, an increase in disposable income could raise consumption towards its life-cycle/desired level, thereby generating a positive effect.

A thorough review of the main theoretical issue on the subject is contained in Jappelli and Pistaferri (2010) or Jappelli and Pistaferri (2017), who also point out that, in addition to the expectation and liquidity effects, income increases may affect consumption in an asymmetric way compared to income decreases. Liquidity constraints can also be due to the asset structure of the household, as argued by Kaplan and Violante (2014), who analyse the issue of the “wealthy hand-to-mouth”.

Empirical literature, on the other hand, has relied either on dedicated surveys, where respondents are asked how they wish to use their extra funds (see Neri et al., 2017, for a review) or on expenditure survey microdata, such as the one employed in this paper. As

for the evaluations on general household consumption and expenditure surveys, Wilcox (1989), Parker (1999), and Souleles (2002) find a positive effect of tax rebates. The findings provided by Souleles (1999) suggest that tax refunds are used for non-durables by low-income/low-liquidity households, thus finding evidence of the role of liquidity constraints, whereas high-income/high-liquidity individuals seem to prefer durables expenditure. Johnson et al. (2006) find confirmation of a positive effect of tax rebate policies, whose magnitude seems to be more pronounced in the case of liquidity constraints. According Agarwal et al. (2007), consumers initially tend to save the bonus, by increasing their credit line and thereby paying off debt, while an increase in spending occurs afterwards. Finally, Parker et al. (2013) find that households spend most of the rebate in durables.

Most empirical literature that used micro data to ascertain the consumption effects of tax rebates has focused on econometric specifications in which the dependent variables is the change in consumption, controlling for a number of socio-economic characteristics of the household (see for example Johnson et al., 2006; Parker, 1999; Parker et al., 2013). Such model specification can be seen as a reduced-form formulation, which is meant to address the potential identification issue stemming from endogenous selection on observables, as households characteristics may indeed determine treatment participation. For this reason, in the empirical application we propose, we adopt a similar choice for the set of covariates in the binary equation when building PS matching variable.

Furthermore, empirical evidence from micro data has been found to be rather sensitive to methodological choices. For example, Heim (2007) warns that the insights from these empirical findings should be interpreted with care: results tend to vary depending on the selected sample, the targeted expenditure component, and the set of regressors used in the treatment evaluation model. It is worth recalling that the positive effect on consumption of tax interventions is only partly supported by the theoretical background and its evidence in empirical studies might well be the result of non-obvious modelling choices. Conclusions on the effectiveness of these policies should therefore be drawn with caution, considering how costly they may be.

According to Government estimates, the 2014 Italian Decree Law 66/2014, also known as “Renzi bonus”, entailed a total transfer of €5.9 billion to households (for about 10 million employees), which amounts to 0.4% of GDP. From a technical point of view, the tax rebate was translated into an increase in the monthly salary or pension by €80. Recipients were payroll employees or retired workers with a total annual income between €8145 and €26000. The rebate was delivered directly in the monthly check, starting in May 2014. Because the eligibility status was based on the 2014 income, which has been verified only in 2015, several cases of classifications of people near the lower threshold have occurred. In fact, it was estimated that about 1.5 million people were incorrectly classified as eligible, and had to reimburse their bonus. Finally, it is worth noting that the standard bonus amount of €80 was reduced for those people whose income was €24000-26000 and for workers who found a job in 2014, as the bonus was proportional to the number of months spent in employment during the year.

Neri et al. (2017) provide some empirical evidence of the 2014 tax rebate based on the SHIW dataset, by extracting a panel for the years 2012 and 2014, using PSM to match treated and control groups and finally evaluating the treatment effect by difference-in-difference. On the one hand, this way of performing difference-in-difference eliminates some between-group heterogeneity, thus making the parallel trend assumption likely to be satisfied in practice. On the other, the procedure can be viewed as pure PSM where the outcome is the difference between consumption/expenditure in 2014 and in 2012. The estimated change in household expenditure amounts to 50%–60% of the sum received, with larger responses for low-liquidity/low-income households, thus confirming a role for liquidity constraints. Using the same data from the SHIW, Andini et al. (2018) argue, instead, that the effectiveness of the Renzi bonus could have been enhanced by a targeted choice of recipients: a new treated group is defined via machine learning techniques, starting from the assumption that individuals who have difficulties to make ends meet are more sensitive to income shocks. Addressing the policy to those households only would have led to a 40% larger effect on food consumption than the actual one.

3. Methods

3.1. Propensity Score Matching and Model Uncertainty

Within the frequentist framework, model averaging for PSM has been considered by Kitagawa and Muris (2016) and Xie et al. (2019). Most literature, however, relies on a Bayesian approach, that we briefly describe in this section.

Let D_i be a treatment binary indicator for the i -th individual, $i = 1, \dots, n$, with $D_i = 1$ for treated subjects and call X the $n \times k$ matrix holding their pre-treatment characteristics: in the context of our application, $D_i = 1$ for bonus recipients. The number of possible subsets of the columns of X is 2^k ; let $\mathcal{M}$ be the set of models (the model space), with M_j being its j -th member ($j = 1, \dots, 2^k$); also, define X_j as the matrix containing the corresponding subset of columns of X .

The PS, the conditional probability of being treated given a chosen set of characteristics, is usually described via a generalised linear model of the form

$$g(E[D_i | \mathbf{x}_{ij}]) = \mathbf{x}_{ij}' \boldsymbol{\beta}_j, \quad (1)$$

where $g(\cdot)$ is an appropriate link function associated with a binary choice model, $\mathbf{x}_{ij}$ is the vector of covariates for individual i from matrix X_j , and $\boldsymbol{\beta}_j$ is the parameter vector of model M_j . In other terms, j indexes the 2^k possible models for the probability of being a recipient household, given its characteristics.

PSM is traditionally performed sequentially: first, a given model M_j is estimated by Maximum-Likelihood (ML) and the estimated probability of having received the bonus (the propensity score) is computed; then, households are matched by choosing for each recipient one or more non-recipient households with similar PS.

The exact technique we employ is nearest neighbour matching with replacement: each treated unit is paired to the control unit with the closest PS, where the same control individual can be matched more than once. To enhance the quality of the matching scheme, a maximum PS distance which is acceptable matching, may be set; this is known as caliper. In order to further investigate the quality of the match, balancing conditions should be verified by testing the similarity between the two groups: a quite

popular approach (Rosenbaum and Rubin, 1983) involve testing if there are significant differences in regressor means.

Finally, the treatment effect of interest is evaluated as the mean difference in the outcome y between matched units; in our application, the outcome y corresponds to the monthly average consumption of different categories. Therefore, the treatment effect is the difference in consumption between recipient households and comparable non-recipients. In the following, we consider the average treatment effect on the treated (ATET), which is often considered the most relevant quantity in empirical applications and measures the effects on individuals for whom the treatment is intended; consequently, in our application the ATET measures the increment in spending due to the bonus. Standard unconfoundedness and overlap conditions (see *e.g.*, Rosenbaum and Rubin, 1983) needed for the identification of the ATET, are assumed to hold.

Following the notation in Caliendo and Kopeinig (2008), the PSM estimator of the ATET can be viewed as the estimator of

$$\gamma_j = E_{\Pr(\mathbf{x}_{ij})|D_i=1}\{E[y_i(1)|D_i=1, \Pr(\mathbf{x}_{ij})] - E[y_i(0)|D_i=0, \Pr(\mathbf{x}_{ij})]\}, \quad (2)$$

where $\Pr(\mathbf{x}_{ij}) = E[D_i|\mathbf{x}_{ij}]$ is the PS and $y_i(1)$ and $y_i(0)$ are the so-called potential outcomes. Since the ATET depends on the initial choice of the PS model M_j , the γ parameter also bears the subscript j .

As discussed in Section 1, the choice of using one model M_j implies an assumption on a *true* model specification, to which the parameter γ_j corresponds. If model uncertainty is taken into account, the possibility to choose among the 2^k specifications in $\mathcal{M}$ calls for a model averaging procedure. By exploiting the sequential nature of PSM, it is possible to link directly the uncertainty affecting the PS model to the subsequent treatment effect estimation. In a frequentist setting, a model averaging estimator $\hat{\gamma}_{MA}$ is given by averaging model specific treatment effect estimates $\hat{\gamma}_j$ across the model space $\mathcal{M}$. Formally, we have

$$\hat{\gamma}_{MA} = \sum_{j=1}^{2^k} \hat{\gamma}_j \omega_j, \quad (3)$$

where ω_j is the model's weight.

Equation (3) is translated into the BMA framework by considering M_j and γ_j as random variables: the model-specific weights ω_j are computed via the posterior probability for the PS model, which can be expressed as

$$\omega_j = P(M_j|\mathcal{D}) = \frac{p(\mathcal{D}|M_j)P(M_j)}{P(\mathcal{D})},$$

where $\mathcal{D}$ is the data (treatment assignment, outcome variable, covariates), $p(\mathcal{D}|M_j)$ is the marginal data density (also known as marginal likelihood) and $P(M_j)$ is the prior distribution for model M_j . Following Zigler and Dominici (2014), who define the model averaging posterior distribution for the causal effect, the treatment evaluation problem for the PSM ATET parameter γ can be expressed in terms of posterior expected values of the model-specific γ_j as

$$E(\gamma|\mathcal{D}) = \sum_{j=1}^{2^k} E(\gamma_j|\mathcal{D}, M_j)P(M_j|\mathcal{D}).$$

By exploiting the “intermediate Bayesian” methodology in An (2010), further discussed by Kaplan and Chen (2012), we can obtain the estimated counterpart of the above equation as

$$\hat{\gamma}_{BMA} = \sum_{j=1}^{2^k} \hat{\gamma}_j \hat{\omega}_j \quad (4)$$

where $\hat{\gamma}_j$ is a frequentist estimator of the ATET in (2) and $\hat{\omega}_j$ is the estimated posterior model probability.

In practice, Equation (4) is problematic to compute. First, it is analytically intractable because in general there is no closed formula for model weights. Second, for realistic values of k the cardinality of the model space is too large to make direct computation feasible. Finally, posterior probabilities must be estimated not only for the PS models $P(M_j|\mathcal{D})$ but also for the parameters β_j in (1): the complication arises because closed-form posterior probabilities are generally not available.

The solutions put forward in the literature range from simple approximations (such as using the Bayesian Information Criterion, BIC, for the model posterior probabilities and the ML estimate as the mean of parameters posteriors), to more sophisticated approaches

such as the conjugate prior method (Chen et al., 2008), the Stochastic Search Variable Selection (SSVS) by George and McCulloch (1993), and the RJMCMC (Green, 1995), which is what we adopt here.

In standard BMA applications (Madigan et al., 1995; Hoeting et al., 1999) the model and related parameters can only be retrieved via two-step procedures: model posteriors are computed first, and in a second step the related parameters are obtained via an additional MCMC from each sampled model. In the RJMCMC approach, instead, the model and related parameters are sampled simultaneously at each MCMC iteration and the model averaging posterior distribution of the parameter vector is immediately available, as well as the estimated posterior model probability, which is simply the empirical frequency of a model being selected over the MCMC iterations. As will be argued in the next subsection, the exploration of the full model space reduces to those specifications with a higher posterior probability. Moreover, since parameter posteriors are available, the sample of parameters from RJMCMC can be used to compute a sample of PSs, whose distribution encompasses the variability within each single model and between different models.

The RJMCMC approach overcomes many limitations of other BMA approaches applied to PSM: Kaplan and Chen (2014), for example, adopted a two-step strategy in which PS model probabilities are first approximated by the BIC and used to sample model specific parameters by MCMC, that are then employed to compute the PS and the resulting treatment effect. This approach, however, is not based on simultaneous sampling of models and parameters, which leads to computational inefficiency and low estimation accuracy. Zigler and Dominici (2014), on the other hand, overcome these limitations by means of the SSVS, which allows for the simultaneous sampling of model and parameters, but PSM is framed within the control-function approach: the outcome equation is specified as a function of both the treatment and the PS, along with a set of controls. However, this gives rise to the so-called “feedback effect”: since some covariates may appear in both equations, for the PS and outcome, posterior distributions for the outcome models can affect the ones for the PS model. This issue has been extensively

debated, as it can severely undermine the reliability of the whole PS apparatus (see also Cefalu et al., 2017). Our RJMCMC methodology is fully compatible instead with the standard sequential framework for PSM estimation, which shields our procedure from feedback effects, that may hamper estimation of the ATET. Finally, compared to the SSVS approach, RJMCMC does not require strong parametric assumptions on the the priors and link function specification. This makes SSVS more binding but, at the same time, computationally more convenient; in this application, however, we mitigate the greater computational burden of RJMCMC by CPU parallelisation granted by the **ParMA** package.

3.2. The RJMCMC technique

Formally, the RJMCMC technique can be seen as a variation of the standard Metropolis-Hastings scheme (Hastings, 1970), where computer-generated random draws are used to analyse via simulation the posterior distributions of interest. The main difference between Metropolis-Hastings and RJMCMC lies in the fact that in the latter the size of the vector of parameters of interest β is allowed to change across simulations.

The algorithm can be briefly described as follows: given an initial model $M^{(0)}$, a model $M^{(1)}$ is proposed by adding or dropping one or more variables from $M^{(0)}$ according to a proposal distribution $q(M^{(1)}|M^{(0)})$. The vector $\beta^{(1)}$ (the parameter vector of $M^{(1)}$) is generated via the following rule: first, random entries $u^{(1)}$ are put in the appropriate places for the new explanatory variables that are in $M^{(1)}$ but not in $M^{(0)}$; then the resulting vector is transformed via a differentiable function $f(\cdot)$, so that $\beta^{(1)} = f(\beta^{(0)})$. This process is then iterated so as to obtain $M^{(2)}$ from $M^{(1)}$ and so on.

Suppose for example to have $k = 3$ covariates $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ with the related set of parameters $\beta_1, \beta_2, \beta_3$ and assume to start, at iteration $h - 1$, from $M^{(h-1)} = (\mathbf{x}_1, \mathbf{x}_2)$. If the proposal distribution draws a new specification by adding $\mathbf{x}_3$, then $M^{(h)} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$. Since the dimension of the new model is larger, in order to compute $\beta^{(h)} = (\beta_1^{(h)}, \beta_2^{(h)}, \beta_3^{(h)})$ an additional term u is needed: $\beta^{(h-1)} = (\beta_1^{(h-1)}, \beta_2^{(h-1)}) \rightarrow \beta^{(h)} = f(\beta_1^{(h-1)}, \beta_2^{(h-1)}, u^{(h)})$.

The probability of accepting a movement from $(M^{(h-1)}, \beta^{(h-1)})$ to $(M^{(h)}, \beta^{(h)})$ is

$$\alpha = \min \left[1, \frac{P(M^{(h)}, \beta^{(h)} | \mathcal{D}) q(M^{(h-1)} | M^{(h)})}{P(M^{(h-1)}, \beta^{(h-1)} | \mathcal{D}) q(M^{(h)} | M^{(h-1)})} \cdot \left| \frac{\partial f(\beta^{(h)}, M^{(h)}, u^{(h)})}{\partial(\beta^{(h)}, M^{(h)}, u^{(h)})} \right| \cdot z(u^{(h)}) \right]$$

where $P(M^{(h)}, \beta^{(h)} | \mathcal{D})$ is the target distribution, i.e. the joint posterior for the couple $(M^{(h)}, \beta^{(h)})$, $\left| \frac{\partial f(\beta^{(h)}, M^{(h)}, u^{(h)})}{\partial(\beta^{(h)}, M^{(h)}, u^{(h)})} \right|$ is the Jacobian term, and $z(u^{(h)})$ is given by the ratio of the density distributions of u referring respectively to $M^{(h)}$ and $M^{(h-1)}$. In the spirit of the Metropolis-Hastings technique, this scheme has the effect of extracting the highest number of samples during the MCMC iterations from the models with the highest posterior probability.

Note that the RJMCMC scheme is perfectly suited to data-rich experiments where the number of covariates k is large: as shown by Lamnisos et al. (2009, 2013), RJMCMC achieves good performance even when the cardinality of the model space becomes very large especially if combined with suitable prior choices or multi-parallel chains. For further details on RJMCMC and its application to generalised linear models, we refer the reader to the documentation of the **ParMA** package (Lucchetti and Pedini, 2021).

For the PSM estimation procedure at hand, the issue of model uncertainty for the PS model is handled by using the fact that RJMCMC delivers a sample of $(\beta^{(h)}, M^{(h)})$, $h = 1, \dots, H$. Each parameter is used to compute a PS and the related treatment effect $\hat{\gamma}^{(h)}$. To obtain the counterpart of Equation (4), we simply compute the sample mean as

$$\hat{\gamma}_{BMA} = \frac{\sum_{h=1}^H \hat{\gamma}^{(h)}}{H}. \quad (5)$$

For the posterior variance $V(\gamma | \mathcal{D})$, we can use

$$\hat{V}(\gamma | \mathcal{D}) = \frac{\sum_{h=1}^H \hat{\sigma}^{2(h)}}{H} + \frac{\sum_{h=1}^H (\hat{\gamma}^{(h)} - \hat{\gamma}_{BMA})^2}{H - 1}, \quad (6)$$

where the first element on the right hand-side is the sample mean of the variances $\hat{\sigma}^{2(h)}$ of the treatment effects, and the second one is the sample variance of the treatment estimates. In this way, we account not only for the variability of $\hat{\gamma}$ due to the model uncertainty and the PS, but also for the intrinsic variability of the parameter γ itself.

4. The data

For our empirical analysis, we use Bank of Italy’s SHIW dataset. The SHIW collects micro-data on the economic and financial behaviour of Italian households with biannual periodicity. It contains a sample of individuals belonging to about 8000 households across 300 municipalities, about half of which are panel households who were also interviewed in earlier waves. Households are selected randomly in two stages: first, municipalities are chosen from a stratified sample by region and population guaranteeing that panel units are included, then households are drawn from each municipality. Two datasets are available, that can be matched at the individual and household level. One dataset is cross-sectional, which includes additional information specific to certain waves, such as special modules for research purposes or inquiries about the effects of new reforms; the other one is a historical dataset, that includes harmonised waves for a subset of households since 1989.

Here, we use the 2014 wave and the historical panel data for the years 2012 and 2014. The cross-section dataset includes information about the 2014 tax credit: whether the interviewed household received the bonus, how many components were benefited, the amount received and spent. As pointed out by Neri et al. (2017), this particular part of the survey mirrors the one in Shapiro and Slemrod (2003a), who introduced the use of survey responses to analyse tax rebate impacts. Pre- and post-treatment information is taken from the panel datasets. For a group of panel households (some of which eligible for the bonus) the consumption pattern can be observed across the two years and the impact of the bonus can thus be inferred. By joining the two data sources, we build a sample of 4459 households observed in 2012 and 2014, of which 864 were eligible for the bonus (19.4%). Individual observations refer to households, so the assignment status indicates whether at least one person is eligible.

The outcome variables we consider are: total consumption, food consumption, non-durables expenditure, car-related consumption, and other durables (furniture, appliances). All variables are expressed as € per month; Table 1 reports the related summary statistics for the selected sample. Total consumption decreased from 2012 to 2014

in both treated and control units. As for specific categories, it is interesting to note that food consumption, car consumption and other durables increased in 2014 for the treated group, whereas non durable consumption decreased.

The covariates used to build the PS model are described in Table 2, where the main summary statistics are also reported. These variables pertain to the pre-treatment period 2012 and reflect individual socio-cultural and economic condition. It is important to note that these variables are not used as factors that explain consumption choices *per se*, but rather as observable features which determine the probability for a household to include one or more bonus recipients.

As it can be seen from Table 2, most tax credit recipients belong to the blue-collar and office worker category. However, the high share of “not currently employed” individuals (either unemployed, or retired, or otherwise outside the labour force) in the treated group deserves some clarification. An explanation comes from the definition of QUALP10: the variable refers only to the declared head of household, which may differ from the actual recipient. From the 2014 wave, it is impossible to discern which component inside the household is the true receiver since the binary treatment assignment refers to the household as a whole.

Bonus receivers seem to hold mainly intermediate levels of educational attainment, while the income class is rather heterogeneous. As for the geographical area, about 50% of the units live in the North. The CONDGEN variable is particularly important here: it contains an indicator of perceived economic hardship. Perhaps surprisingly, the share of household who reportedly have the greatest difficulties in making ends meet is higher among non-recipients, with treated units belonging mostly to the middle category (moderate or little economic problems). This is a consequence of the tax rebate design, that was mainly aimed at individuals with a steady job.

Table 1: Outcome variables: main summary statistics

	2014			2012		
	Full sample	treated	control	Full sample	treated	control
c_tot	Total consumption					
mean	1948.3	2129.0	1904.9	2149.2	2326.7	2106.5
std.dev.	1174.3	1031.3	1202.3	1276.3	1107.8	1310.2
median	1683.3	1875.0	1600.0	1858.3	2107.5	1791.7
min	133.33	300.00	133.33	-583.33	-416.67	-583.33
max	15617	10742	15617	15867	11937	15867
c_food	Food consumption					
mean	545.08	616.61	528.37	544.95	597.85	532.23
std.dev.	295.72	281.05	296.75	301.64	283.55	304.50
median	500.00	600.00	500.00	500.00	550.00	500.00
min	50.000	100.00	50.000	50.000	50.000	50.000
max	3500.0	2500.0	3500.0	3000.0	2700.0	3000.0
c_notd	Non durable goods					
mean	1866.1	1989.8	1836.4	2065.8	2224.4	2027.7
std.dev.	1056.1	853.25	1097.3	1175.7	1021.1	1206.9
median	1600.0	1800.0	1600.0	1800.0	2045.0	1745.0
min	150.00	550.00	150.00	150.00	500.00	150.00
max	14200	6750.0	14200	15867	11012	15867
c_car	Cars/means of transportation					
mean	51.473	94.736	41.076	49.378	58.688	47.141
std.dev.	283.72	392.48	249.60	274.60	271.86	275.24
median	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
min	-1000	-666.67	-1000	-2416.7	-2416.7	-1583.3
max	5166.7	5166.7	4166.7	4333.3	2083.3	4333.3
c_otherd	Other durable goods					
mean	30.745	44.421	27.458	34.031	43.662	31.717
std.dev.	141.04	159.27	136.11	117.06	136.67	111.73
median	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
max	4166.7	2500.0	4166.7	2750.0	1666.7	2750.0

Note: Negative values in car consumption are due the definition of the variable, which includes depreciation.

Table 2: Explanatory variables: main summary statistics (pre-treatment period)

	Full sample	treated	control
NEQU	Number of components of the household		
mean	1.700	1.973	1.635
std.dev.	0.570	0.548	0.555
median	1.500	2.000	1.500
min	1.000	1.000	1.000
max	4.400	3.800	4.400
CLY2	Income class (quintiles)		
category 1	0.172	0.075	0.195
category 2	0.184	0.155	0.191
category 3	0.188	0.206	0.184
category 4	0.215	0.267	0.202
category 5	0.241	0.297	0.228
AREA3	Geographical area		
North	0.416	0.471	0.403
Centre	0.196	0.206	0.194
South/Islands	0.388	0.323	0.403
ETA5	Head of household age class		
≤ 30	0.025	0.051	0.019
31 – 40	0.095	0.198	0.070
41 – 50	0.195	0.339	0.161
51 – 65	0.323	0.337	0.320
> 65	0.362	0.075	0.430
STUDIO	Head of household educational level		
None	0.038	0.009	0.045
Elementary school	0.216	0.077	0.249
Middle school	0.353	0.455	0.328
High school	0.270	0.344	0.253
Bachelor degree	0.111	0.108	0.112
Post-graduate	0.012	0.007	0.013
QUALP10	Head of household job type		
Blue-collar worker	0.149	0.359	0.099
Office worker	0.136	0.292	0.098
Junior manager	0.023	0.032	0.021
Manager	0.014	0.010	0.014
Member arts/professions	0.022	0.008	0.025
Sole proprietor	0.008	0.009	0.008
Freelance	0.041	0.027	0.044
Family business	0.008	0.003	0.009
Active shareholder	0.012	0.012	0.012
Not employed	0.589	0.248	0.670
CONDGEN	Perceived economic hardship		
Great difficulty	0.189	0.176	0.191
Difficulty	0.159	0.174	0.156
Some difficulty	0.300	0.340	0.291
Fairly easy	0.265	0.257	0.267
Easy	0.064	0.046	0.068
Very easy	0.023	0.007	0.026

5. Estimation results

5.1. PS models

For the RJMCMC application presented here we use a probit model for the PS, with a uniform prior distribution for models, $P(M_j)$, and a Zellner-g type (Zellner, 1986) for the PS model parameter vector β_j . These prior choices are standard in the BMA framework, and can be briefly illustrated as follows: the prior probability for a model $P(M_j)$ is given by the binomial formula

$$P(M_j) = \prod_{l=1}^k \phi_l^{\delta_{jl}} (1 - \phi_l)^{1-\delta_{jl}} \quad (7)$$

where k is the total number of covariates considered, $0 \leq \phi_l \leq 1$ is the prior probability that the l -th variable is significant and δ_{jl} is a 0-1 indicator of variable inclusion. In this exercise, we adopt the choice $\phi_l = 0.5$ for all variables, which leads to a uniform distribution; the implied assumption here is that all variables are treated alike and have a 50% chance of being significant in model M_j . This choice takes each PS model as equally likely *a priori* and is considered a conservative choice that does not favour any particular specification.

The choice for the prior distribution of the parameters of a given model $P(\beta|M_j)$ adopts a normality assumption, as in most BMA literature:

$$\beta|M_j \sim N(0, V_{0,j}); \quad (8)$$

for the prior variance $V_{0,j}$ we adopted the so-called Zellner-g type prior, which assumes $V_{0,j} = n(X_j'X_j)^{-1}$ (except for the constant, for which a diffuse prior is generally used); this choice takes the correlation among covariates into account, differently from the often-used ridge prior which involves $V_{0,j} = cI_j$, with I_j as the identity matrix and $c > 0$ a shrinkage parameter. The intercept, moreover, is included in every model specification so, as a consequence, all other covariates are centred (Fernandez et al., 2001a).

Our choices for the priors are fairly standard in the literature: we explore several alternatives in section 5.3 below.

The model proposal distribution follows the one by Madigan et al. (1995), who propose to update the model specification at each step by simply adding or deleting one covariate. A more complex alternative which allows for more general movement patterns between models is also possible, but explorations in this direction made no significant difference. As for the differential function $f(\cdot)$ requested for the definition of new model parameters β_j , we refer to Lucchetti and Pedini (2021), where the automatic RJMCMC sampler by Green (2003) and Lamnisos et al. (2009) is used. To sample models and parameters, we use 500000 MCMC iterations with a preliminary burn-in stage of 50000. Finally, the whole algorithm was performed using eight parallel Markov chains.

In Table 3 we report, for each variable, the posterior inclusion probability (PIP), the frequency with which a covariate is included in the PS specification, giving an indication of the variable’s relevance. For the corresponding β , we report the sample mean and standard deviation, in two varieties: unconditional (labelled “Mean” and “St. Err.”) and conditional on variable inclusion (labelled “C. Mean” and “C. St. Err.”); the latter are computed only on model specifications where the variable appears with a nonzero coefficient.

According to the PIP indicator, the number of household components and all the income classes seem to be extremely relevant, together with the South/Islands geographical dummy. As for the age categories, the PIP for all but the 31–40 interval are close to 1. For the education variable STUDIO, only the middle school and post-graduate categories report a PIP greater than 0.50. For the job type variable QUALP10, we have instead that most dummies exhibit a large inclusion frequency (with different intensity) except for office workers and sole proprietors. Among the economic deprivation dummies, the category with no perceived difficulties seems to show a significant effect.

In order to give the reader an idea of the impact of model uncertainty in the PS model specification, we report in Table 4 the estimation results for a probit model where all the covariates are used (Full) and for the 10 top-ranked probit models, from M1 to M10, that are the 10 models with the largest $\hat{P}(M_j|\mathcal{D})$ from the RJMCMC sampler. We use ordinary frequentist estimation methods so as to adhere as closely as possible

Table 3: PS model: RJMCMC sampling statistics for the parameter vector β

	Mean	St. Err.	PIP	C. Mean	C. St. Err.
Intercept	-1.114	0.029	1.000	-1.114	0.0291
NEQU	0.322	0.056	1.000	0.322	0.056
CLY2 (Ref: category 1)					
Category 2	0.398	0.107	0.986	0.404	0.096
Category 3	0.521	0.109	0.992	0.525	0.100
Category 4	0.553	0.117	0.992	0.557	0.106
Category 5	0.585	0.135	0.991	0.590	0.124
AREA3 (Ref: North)					
Centre	0.000	0.0113	0.024	0.003	0.074
South/Islands	-0.130	0.091	0.739	-0.175	0.0573
ETA5 (Ref: ≤ 30)					
31 – 40	-0.014	0.072	0.075	-0.185	0.196
41 – 50	-0.299	0.111	0.964	-0.310	0.096
51 – 65	-0.544	0.104	1.000	-0.544	0.104
> 65	-0.999	0.126	1.000	-0.999	0.126
STUDIO (Ref: None)					
Elementary school	0.004	0.034	0.054	0.0796	0.123
Middle school	0.127	0.096	0.705	0.181	0.059
High school	-0.032	0.078	0.192	-0.165	0.099
Bachelor degree	-0.085	0.146	0.305	-0.279	0.126
Post-graduate	-0.379	0.425	0.512	-0.740	0.292
QUALP10 (Ref: Blue collar)					
Office worker	-0.002	0.022	0.033	-0.049	0.108
Junior manager	-0.252	0.236	0.597	-0.423	0.146
Manager	-0.573	0.343	0.811	-0.707	0.224
Member arts/professions	-1.169	0.213	1.000	-1.169	0.213
Sole proprietor	-0.051	0.168	0.110	-0.465	0.253
Freelance	-0.889	0.129	1.000	-0.889	0.129
Family business	-1.294	0.330	0.999	-1.295	0.328
Active shareholder	-0.535	0.327	0.803	-0.667	0.214
Not employed	-0.764	0.067	1.000	-0.764	0.067
CONDGEN (Ref: Great difficulty)					
Difficulty	0.002	0.017	0.034	0.054	0.072
Some difficulty	0.007	0.030	0.085	0.088	0.061
Fairly easy	-0.038	0.076	0.238	-0.158	0.072
Easy	-0.090	0.150	0.313	-0.289	0.125
Very easy	-0.789	0.267	0.975	-0.809	0.238

Boldface is used to highlight those variables with $pip > 0.5$ following Raftery (1995).

to the common practice in PS modelling. Although BMA should not be regarded as a model selection procedure, it is interesting to examine which models emerged with the highest posterior probability using, for example, information criteria, such as the BIC. Note that a full enumeration of the model space for the lowest BIC would imply, in our case, analysing 2^{30} models, which is of course computationally infeasible.

Alternatively, a stepwise model selection procedure to identify the best models could be employed. However, this procedure may not be optimal, as some specifications may be skipped simply as a result of the sequential deletion/addition of variables. In this respect, the model space exploration provided by the RJMCMC grants a more widespread and computationally efficient analysis of the model set.

According to the posterior model probabilities, M1 appears to be slightly more probable than the others, with $\hat{P}(M_j|\mathcal{D}) \approx 0.05$. Interestingly, this model is also associated with the lowest BIC value among the top 10 selected.

Finally, it is important to clarify that the values for the estimated coefficients reported in Tables 3 and 4 refer to the treatment probability estimated by the PS. Therefore, they cannot be compared to those estimated in most previous empirical literature on the effects of tax rebates, where the dependent variable of the model is the change in consumption rather than the treatment binary indicator: the effects of households socio-economic characteristics on consumption are not identified in our empirical setting.

5.2. ATET estimation

In this section, we compare the ATET estimates for our proposed BMA-based procedure with those yielded by the set of 10 top-ranked models selected according to the sampling frequency of the RJMCMC. By focusing on the differences in the ATET estimates between the two approaches we show that model uncertainty, which arises from the application of standard model selection procedures, is likely to produce final results that are largely dissimilar, even if based on PS model specifications with a comparable fit. As anticipated in Section 3.1, matching is performed using pairwise nearest neighbour with replacement, using 0.01 as caliper.

Table 4: PS model: full and 10 top-ranked probit models (std. errors in parenthesis)

	Full	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Intercept	-0.275 (0.254)	-0.963 (0.116)	-0.950 (0.116)	-0.981 (0.116)	-0.984 (0.115)	-0.883 (0.118)	-0.946 (0.116)	-0.895 (0.118)	-0.967 (0.116)	-0.751 (0.120)	-0.970 (0.115)
NEQU	0.291 (0.053)	0.345 (0.049)	0.341 (0.049)	0.347 (0.049)	0.301 (0.047)	0.307 (0.051)	0.336 (0.050)	0.306 (0.051)	0.343 (0.049)	0.326 (0.050)	0.298 (0.047)
CLY2 (Ref: Category 1)											
Category 2	0.433 (0.099)	0.398 (0.096)	0.401 (0.096)	0.396 (0.096)	0.407 (0.095)	0.412 (0.096)	0.401 (0.096)	0.411 (0.096)	0.398 (0.096)	0.405 (0.095)	0.410 (0.095)
Category 3	0.606 (0.102)	0.507 (0.094)	0.512 (0.094)	0.499 (0.094)	0.531 (0.093)	0.549 (0.095)	0.513 (0.094)	0.544 (0.095)	0.505 (0.094)	0.540 (0.094)	0.535 (0.093)
Category 4	0.670 (0.108)	0.518 (0.094)	0.524 (0.094)	0.509 (0.094)	0.579 (0.091)	0.594 (0.097)	0.531 (0.094)	0.591 (0.096)	0.515 (0.094)	0.569 (0.095)	0.583 (0.091)
Category 5	0.790 (0.122)	0.532 (0.099)	0.547 (0.099)	0.503 (0.099)	0.613 (0.095)	0.675 (0.108)	0.568 (0.100)	0.659 (0.107)	0.520 (0.099)	0.617 (0.102)	0.627 (0.095)
AREA3 (Ref: North)											
Centre	-0.021 (0.067)	-	-	-	-	-	-	-	-	-	-
South/Islands	-0.183 (0.061)	-0.171 (0.056)	-0.166 (0.056)	-0.172 (0.055)	-	-0.193 (0.056)	-0.176 (0.056)	-0.195 (0.056)	-0.167 (0.056)	-0.163 (0.056)	-
ETA (Ref: ≤ 30)											
31 – 40	-0.232 (0.144)	-	-	-	-	-	-	-	-	-	-
41 – 50	-0.498 (0.139)	-0.280 (0.075)	-0.285 (0.075)	-0.294 (0.075)	-0.278 (0.075)	-0.288 (0.075)	-0.282 (0.075)	-0.301 (0.075)	-0.300 (0.075)	-0.306 (0.076)	-0.284 (0.075)
51 – 65	-0.742 (0.138)	-0.513 (0.074)	-0.522 (0.074)	-0.521 (0.074)	-0.524 (0.073)	-0.517 (0.074)	-0.514 (0.074)	-0.525 (0.074)	-0.531 (0.074)	-0.562 (0.075)	-0.533 (0.074)
> 65	-1.222 (0.155)	-0.954 (0.097)	-0.965 (0.097)	-0.956 (0.097)	-0.962 (0.097)	-0.959 (0.097)	-0.958 (0.097)	-0.961 (0.097)	-0.967 (0.097)	-1.059 (0.099)	-0.973 (0.097)
STUDIO (Ref: None)											
Elementary	-0.268 (0.195)	-	-	-	-	-	-	-	-	-	-
Middle school	-0.194 (0.195)	0.174 (0.052)	0.168 (0.052)	0.191 (0.052)	0.187 (0.052)	0.171 (0.052)	0.173 (0.052)	0.186 (0.052)	0.185 (0.052)	-	0.181 (0.052)
High school	-0.362 (0.199)	-	-	-	-	-	-	-	-	-0.180 (0.059)	-
Bachelor	-0.491 (0.211)	-	-	-	-	-	-	-	-	-0.320 (0.090)	-
Post-graduate	-0.954 (0.327)	-	-0.579 (0.258)	-	-	-	-	-	-0.611 (0.259)	-0.807 (0.263)	-0.603 (0.258)
QUALP10 (Ref: Blue collar)											
Office worker	-0.058 (0.082)	-	-	-	-	-	-	-	-	-	-
Junior manager	-0.413 (0.152)	-0.435 (0.141)	-0.422 (0.141)	-	-0.439 (0.141)	-0.393 (0.141)	-0.418 (0.141)	-	-	-0.403 (0.141)	-0.424 (0.141)
Manager	-0.668 (0.228)	-0.748 (0.210)	-0.678 (0.215)	-0.697 (0.210)	-0.770 (0.210)	-0.733 (0.212)	-0.743 (0.211)	-0.686 (0.212)	-0.627 (0.215)	-0.619 (0.217)	-0.697 (0.215)
Member arts/prof.	-1.141 (0.211)	-1.218 (0.201)	-1.181 (0.201)	-1.174 (0.200)	-1.218 (0.201)	-1.181 (0.202)	-1.192 (0.201)	-1.139 (0.201)	-1.137 (0.200)	-1.127 (0.202)	-1.181 (0.201)
Sole proprietor	-0.461 (0.249)	-	-	-	-	-	-	-	-	-	-
Freelance	-0.937 (0.134)	-0.898 (0.127)	-0.902 (0.127)	-0.869 (0.127)	-0.900 (0.127)	-0.891 (0.127)	-0.897 (0.127)	-0.865 (0.127)	-0.874 (0.127)	-0.907 (0.128)	-0.903 (0.127)
Family business	-1.346 (0.321)	-1.250 (0.313)	-1.261 (0.314)	-1.210 (0.313)	-1.257 (0.314)	-1.261 (0.317)	-1.257 (0.315)	-1.226 (0.317)	-1.223 (0.313)	-1.301 (0.315)	-1.269 (0.315)
Act. shareholder	-0.733 (0.216)	-0.672 (0.210)	-0.679 (0.210)	-0.633 (0.210)	-0.630 (0.210)	-0.680 (0.211)	-0.674 (0.211)	-0.645 (0.210)	-0.642 (0.210)	-0.679 (0.211)	-0.639 (0.210)
Not employed	-0.822 (0.073)	-0.773 (0.063)	-0.775 (0.063)	-0.748 (0.063)	-0.781 (0.063)	-0.763 (0.063)	-0.766 (0.063)	-0.740 (0.063)	-0.751 (0.063)	-0.793 (0.064)	-0.782 (0.063)
CONDGEN (Ref: Great difficulty)											
Difficulty	-0.013 (0.087)	-	-	-	-	-	-	-	-	-	-
Some difficulty	-0.023 (0.082)	-	-	-	-	-	-	-	-	-	-
Fairly easy	-0.188 (0.094)	-	-	-	-	-0.180 (0.065)	-	-0.192 (0.065)	-	-	-
Easy	-0.328 (0.136)	-	-	-	-	-0.330 (0.115)	-0.232 (0.109)	-0.355 (0.115)	-	-	-
Very easy	-0.859 (0.240)	-0.741 (0.221)	-0.753 (0.222)	-0.758 (0.222)	-0.739 (0.221)	-0.886 (0.225)	-0.776 (0.222)	-0.910 (0.226)	-0.770 (0.223)	-0.706 (0.224)	-0.751 (0.222)
BIC	3633.0	3563.4	3566.2	3564.8	3564.5	3567.7	3567.1	3567.3	3567.0	3568.6	3566.8
$\hat{P}(M \mathcal{D})$	-	0.0487	0.0231	0.0188	0.0170	0.0165	0.0143	0.0138	0.0118	0.0105	0.01039

Table 5 reports the balancing properties resulting from the matching in the RJMCMC PS models. In order to assess the quality of the matching procedure, we perform paired and unpaired t -tests on the variables mean difference between treated and control units. As a summary measure of the test rejection rate, we use here the average p -value of the above tests, conditional on variable inclusion, over the 500000 RJMCMC samples; in other words, we average the p -values over all models that include that particular variable.

The PIP associated with each variable is reported in Table 3. As Table 5 shows, the null hypotheses of no mean differences in the covariates between the treated and control group, in both the paired and unpaired case, cannot be rejected.

The ATET estimator is computed in the same way as Neri et al. (2017): we take the difference between the 2014 and 2012 consumption ATET values, using the same matching group for both. This strategy produces an outcome equivalent to a difference-in-difference, where the comparison groups are adjusted for PS (sometimes known as propensity score diff-in-diff). Similar results could be obtained by taking as outcome variable the consumption difference between 2014 and 2012: even if the final ATET estimate will be the same, a different standard error will be produced. These results are not presented, but available upon requests.

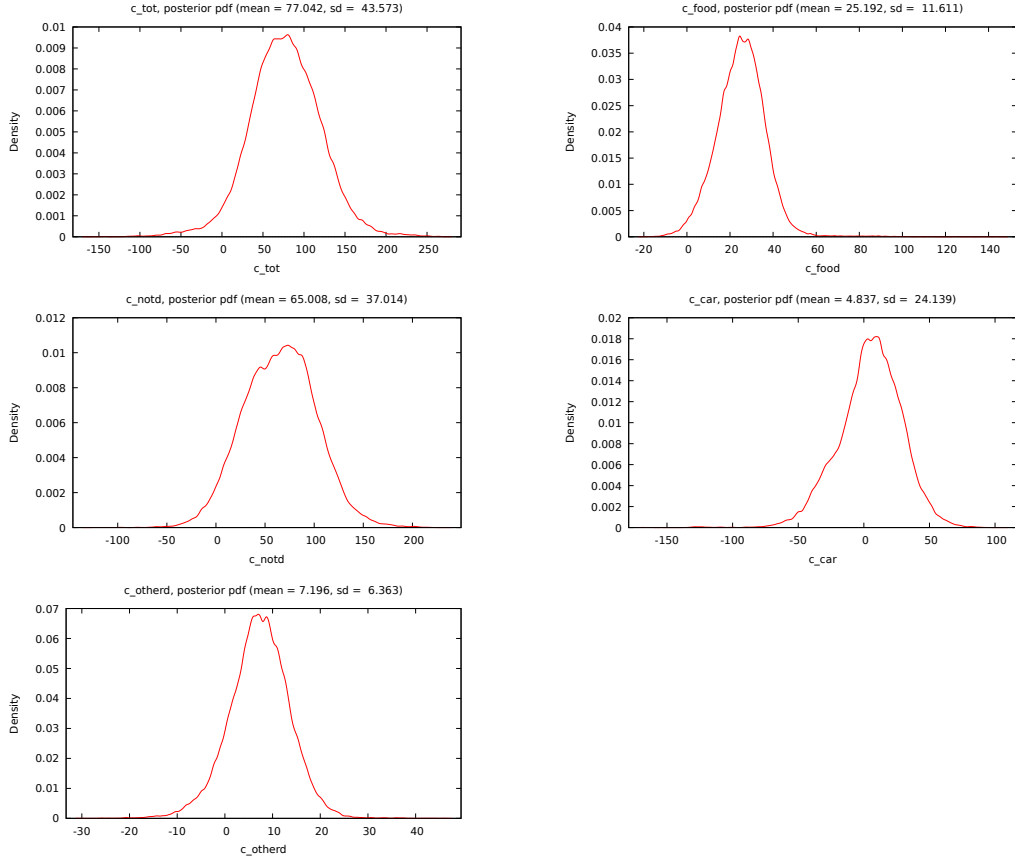
Figure 1 shows the distributions of the ATET estimates across the model space for the five outcomes we consider. Each panel contains a kernel estimate of ATET density across the whole set of sampled models. All plots span a large range in ATET values. Null or negative impacts of the tax rebate can be found on the left tail of the densities for total, nondurables, food and other durables consumption; the probability of such effects is even larger for car-related expenditure.

The final BMA estimates of the ATET are reported at the top of Table 6, where the averages across samples are computed as in Equation (5) and the associated standard error is evaluated as the square root of the variance in Equation (6). Note that none of the ATET estimates are statistically significant. For comparison, the rest of the Table reports the ATET estimates based on the selected specification choices for the PS

Table 5: Balancing conditions in BMA-RJMCMC: Paired and Unpaired t -tests p-values

	Unpaired p-value	Paired p-value
NEQU	0.62100	0.42821
CLY2 (Ref: Category 1)		
Category 2	0.53506	0.32832
Category 3	0.57794	0.38833
Category 4	0.53781	0.3228
Category 5	0.62168	0.41719
AREA3 (Ref: North)		
Centre	0.37074	0.24227
South/Islands	0.51887	0.32847
ETA5 (Ref: ≤ 30)		
31 – 40	0.48597	0.32464
41 – 50	0.62956	0.43011
51 – 65	0.61098	0.36042
> 65	0.69072	0.35273
STUDIO (Ref: None)		
Elementary school	0.66068	0.48884
Middle school	0.40851	0.21517
High school	0.45373	0.26462
Bachelor's degree	0.51339	0.37247
Post-graduate	0.43066	0.41001
QUALP10 (Ref: Blue collar)		
Office worker	0.61926	0.42765
Junior manager	0.60105	0.48635
Manager	0.49527	0.42713
Member arts/professions	0.60515	0.48871
Sole proprietor	0.39222	0.38406
Freelance	0.59269	0.45853
Family business	0.58023	0.49427
Active shareholder	0.40669	0.37210
Not employed	0.75256	0.50859
CONDGEN (Ref: Great difficulty)		
Difficulty	0.22100	0.13715
Some difficulty	0.57952	0.41325
Fairly easy	0.62906	0.45421
Easy	0.55725	0.43848
Very easy	0.61658	0.50111

Figure 1: ATET distributions across the model space



model, that are Full and M1 to M10. Test results for the related balancing conditions are available upon request.

Note that the results produced by models M1-M10 are quite heterogeneous. To be more specific, models M1, M2 and M6 produce a significant effect for total, non-durables and food consumption, model M4 and M10 show significant ATET for food consumption while models M8 and M9 exhibit significant effects for nondurables/food and total/nondurables consumption, respectively.

The implication of Figure 1 and Table 6 is clear: blindly committing to model selection methods may lead to ignoring the fact that very similar specification, nearly equivalent in terms of fit and Information Criteria, may lead to dramatically different estimates of the ATET.

Table 6: ATET estimates: BMA, Full and 10 top-ranked models

		BIC	ATET				
			c_tot	c_food	c_notd	c_car	c_otherd
BMA			77.042 (77.394)	25.192 (20.630)	65.008 (66.969)	4.837 (32.857)	7.196 (11.159)
Full	3633.0		59.398 (72.915)	23.519 (20.004)	51.373 (64.384)	13.297 (22.978)	-5.272 (10.049)
M1	3563.4		102.826 * (61.661)	33.083 ** (15.803)	90.360 * (54.292)	2.329 (20.734)	10.137 (8.594)
M2	3566.2		120.210 * (64.365)	36.187 ** (16.083)	114.348 ** (56.676)	1.013 (20.642)	4.850 (8.745)
M3	3564.8		70.292 (62.055)	25.430 (15.949)	64.482 (53.891)	5.898 (21.300)	-0.089 (9.961)
M4	3564.5		80.933 (61.708)	29.035 * (16.024)	87.173 (53.681)	-13.498 (22.854)	7.258 (8.835)
M5	3567.7		91.222 (63.273)	27.845 (17.215)	60.156 (54.031)	29.010 (23.628)	2.056 (8.955)
M6	3567.1		116.929 * (62.398)	32.164 * (16.429)	98.521 * (53.931)	6.220 (21.578)	12.188 (9.461)
M7	3567.3		69.808 (64.815)	17.924 (17.383)	43.890 (57.261)	14.364 (21.579)	11.554 (9.147)
M8	3567.0		88.561 (63.614)	38.410 ** (16.510)	98.652 * (55.961)	-11.562 (21.027)	1.472 (8.502)
M9	3568.6		118.153 ** (57.891)	9.884 (17.447)	98.390 * (51.281)	19.230 (20.694)	0.534 (8.724)
M10	3566.8		44.476 (61.937)	34.488 ** (15.785)	72.549 (53.624)	-30.044 (23.337)	1.970 (8.976)

*: p-value < 0.10, **: p-value < 0.05, ***: p-value < 0.01. Standard errors in parentheses.

In order to further investigate the potential role of liquidity constraints, we estimate the effect of the tax rebate on a subset of households that are subject to financial strain and/or economic deprivation. Following Neri et al. (2017), two subsamples are selected. The first one is based on financial wealth, where a household is classified as liquidity constrained if its net financial wealth (difference between financial activities and liabilities) is smaller than half of its labour income (Broda and Parker, 2014). The second one embeds those families who perceive great difficulties to make ends meet, i.e. those belonging to the first category of the CONDGEN variable.

Table 7 reports the ATET estimates for the BMA-based PSM: the effect of the tax rebate on food consumption is statistically significant if only households facing economic deprivation are considered, which validates the liquidity constraint hypothesis. It is worth emphasising that a model selection procedure would have led us to conclude that the rebate was effective (see the ATET for food and nondurables consumption using M1 in Table 6) for the whole sample of households, which would have been somewhat at odds with extant evidence in the related literature. Instead, model averaging coherently points toward a role for liquidity constraints, as the tax rebate has a significant effect on food expenditure only for those households struggling to make ends meet. This example effectively shows that model selection matters and how BMA makes it possible to handle model uncertainty, leading to a more complete portrait of the phenomenon and a more genuine evaluation of the ATET.

5.3. Robustness analysis

In order to assess the robustness of our results with respect to the modelling choices we made in terms of priors, we ran two variations on the procedure used in the previous subsection. First, we adopt different choices for the model prior $P(M_j)$; next, we consider an alternative possibility for the priors on the coefficients $P(\beta|M_j)$.

As for the former check, we consider different values for the parameter ϕ in Equation (7), that is $\phi = \{0.25, 0.40, 0.60, 0.75\}$. It should be noted that the cases when $\phi < 0.5$ tend to favour parsimonious specifications, while $\phi > 0.5$ produce models with a larger

Table 7: BMA ATET estimates: liquidity constraints

	c_tot	c_food	c_notd	c_car	c_otherd
Financial wealth	9.422 (76.538)	23.397 (23.129)	3.217 (62.675)	6.193 (34.451)	0.012 (12.627)
Economic deprivation	9.649 (134.659)	79.327** (37.961)	10.137 (128.707)	-5.340 (24.708)	4.912 (11.536)

*: p-value < 0.10, **: p-value < 0.05, ***: p-value < 0.01. Standard errors in parentheses. Financial wealth: 2,781 households, 628 treated (23%) - RJMCMC with 100000 iterations after 10000 of burn-in. Economic deprivation: 841 households, 152 treated (18%). For the second subsample, the QUALP10 variable is expressed in three categories (employee, self-employed, not employed) for identification purposes as well as the post-graduate educational level class is embodied in the Bachelor's degree category. RJMCMC with 500000 iterations after 50000 of burn-in.

number of covariates. As a consequence, the extreme values $\phi = 0.25$ and $\phi = 0.75$ should be considered as rather unrealistic, since they produce PS models that are extremely tight in the former case and very large in the latter one. Given the need for balancing bias and variance in this type of application (see Section 1), the results for the extreme choices of ϕ should just be taken as a robustness check, and not as a serious modelling possibility.

Table 8: ATET with different choices for the model prior

	Total consumption	Food	Nondurables	Car	Other durables
$\phi = 0.25$	11.88 (98.19)	56.07 (39.13)	47.58 (73.74)	-49.60 (59.25)	13.89 (15.47)
$\phi = 0.40$	69.36 (81.17)	28.30 (22.86)	64.44 (67.50)	-3.20 (37.72)	8.12 (11.45)
$\phi = 0.50$	77.04 (77.39)	25.19 (20.64)	65.01 (66.97)	4.84 (32.86)	7.20 (11.16)
$\phi = 0.60$	80.20 (76.66)	23.10 (20.64)	63.23 (67.29)	10.37 (30.65)	6.60 (11.10)
$\phi = 0.75$	80.03 (76.24)	20.05 (21.16)	56.99 (67.26)	17.32 (29.01)	5.72 (11.04)

The results for the ATET are shown in Table 8, where the middle line $\phi = 0.5$ reproduces our earlier estimates for ease of comparison. As can be seen, the results for the ATET are qualitatively homogeneous in all cases; the largest discrepancies arise in the $\phi = 0.25$ case and the category which appears to be most affected by different choices of ϕ is “car”. In all cases, the balancing conditions were met.

The robustness check we carried out on the coefficient priors is based on replacing the Zellner-type prior in (8) with an identity matrix (*i.e.*, $V_{0,j} = I_j$). This choice gives rise to a ridge prior, which is the most commonly employed alternative in the BMA literature. The difference between these two possibilities lies in the different assumptions on the correlations between coefficients, which are data-driven for the Zellner prior and fixed at 0 *a priori* for the ridge prior. With the ridge prior, the prior probability for each coefficient β_i is independent of all other coefficients in the model, thereby implementing a position of complete agnosticism. Table 9 contains the results. Like in the previous case, results appear to be remarkably robust, with the only possible exception of the “car” category.

Table 9: ATET with different choices for the parameter prior

	Total consumption	Food	Nondurables	Car	Other durables
Zellner prior	77.04 (77.39)	25.19 (20.64)	65.01 (66.97)	4.84 (32.86)	7.20 (11.16)
Ridge prior	84.13 (76.70)	20.62 (21.34)	60.64 (67.47)	17.98 (29.00)	5.51 (11.03)

6. Discussion and conclusions

Model uncertainty is an impelling problem in any model building procedure, but little attention has been devoted in the PSM literature to this topic. In this article, we propose a BMA solution for the PSM estimation of treatment effects. Making use of RJMCMC allows us to avoid BIC approximations of model posteriors and two-step procedures, that reduce estimation accuracy, while framing our approach within the standard sequential framework for PSM eliminates the risk of feedback effects affecting ATET estimates.

Using the SHIW dataset, we estimate the consumption effects of the Italian 2014 tax rebate. Results show that the impact of the tax rebate on consumption is quite different across the expenditure categories considered. In fact, the additional consumption is mainly driven by an increase in the expenditure for non durable goods. This result is in line with that by Neri et al. (2017), who report similar findings for the self-reported spending of the tax rebate in the SHIW survey. However, such increase in expenditure is statistically insignificant for all of the consumption categories considered here.

Perhaps more interestingly, the problem of model uncertainty is evident in our application. By using a model selection procedure, we show that slightly different choices for the PS specification can lead to markedly different estimates of the ATET. Consider, for instance, the estimated ATET resulting for the effects on total, nondurables and food consumption. If one considered only the PS models with the smallest BIC (see Section 5.1), the estimated effect would be deemed to be statistically significant. Such results may enable the conclusion that tax credit is indeed an overall effective policy instrument, in contrast to most theoretical literature which predicts negligible effect due to consumption smoothing, except in the case of liquidity constraints (Jappelli and Pistaferri, 2010; Kaplan and Violante, 2014). However, after taking model uncertainty

into account, this positive effect is called into question, as BMA-based PSM estimates paint a much blurrier picture, where treatment effects for all consumption categories are hardly significant at all, except perhaps for liquidity constrained households, in line with previous findings (Souleles, 2002; Johnson et al., 2006; Agarwal et al., 2007).

In our opinion, the evidence we present should make it imperative to at least consider the issue of model uncertainty in all applied research using PS matching to evaluate policy intervention, and a larger number of empirical applications would be most welcome to generalise the findings reported here.

References

- Agarwal, S., Liu, C., Souleles, N.S., 2007. The reaction of consumer spending and debt to tax rebates. Evidence from consumer credit data. *Journal of Political Economy* 115, 986–1019.
- An, W., 2010. Bayesian propensity score estimators: incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology* 40, 151–189.
- Andini, M., Ciani, E., de Blasio, G., D’Ignazio, A., Salvestrini, V., 2018. Targeting with machine learning: an application to a tax rebate program in Italy. *Journal of Economic Behavior & Organization* 156, 86–102.
- Belloni, A., Chernozhukov, V., Fernández-Val, I., Hansen, C., 2017. Program evaluation and causal inference with high-dimensional data. *Econometrica* 85, 233–298.
- Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* 81, 608–650.
- Brock, W.A., Durlauf, S.N., 2001. What have we learned from a decade of empirical research on growth? Growth empirics and reality. *The World Bank Economic Review* 15, 229–272.

- Brock, W.A., Durlauf, S.N., 2015. On sturdy policy evaluation. *The Journal of Legal Studies* 44, S447–S473.
- Brock, W.A., Durlauf, S.N., West, K.D., 2003. Policy evaluation in uncertain economic environments. Technical Report. National Bureau of Economic Research.
- Broda, C., Parker, J.A., 2014. The economic stimulus payments of 2008 and the aggregate demand for consumption. *Journal of Monetary Economics* 68, S20–S36.
- Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., Stürmer, T., 2006. Variable selection for propensity score models. *American Journal of Epidemiology* 163, 1149–1156.
- Bryson, A., Dorsett, R., Purdon, S., 2002. The use of propensity score matching in the evaluation of active labour market policies. LSE Department for Work and Pensions Working Paper .
- Caliendo, M., Kopeinig, S., 2008. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* 22, 31–72.
- Cefalu, M., Dominici, F., Arvold, N., Parmigiani, G., 2017. Model averaged double robust estimation. *Biometrics* 73, 410–421.
- Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A* 158, 419–466.
- Chen, M.H., Huang, L., Ibrahim, J.G., Kim, S., 2008. Bayesian variable selection and computation for generalized linear models with conjugate priors. *Bayesian Analysis* 3, 585–614. doi:100.1214/08-BA323.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, C1–C68.

- Dehejia, R.H., Wahba, S., 1999. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94, 1053–1062.
- Eris, M.N., Ulasan, B., 2013. Trade openness and economic growth: Bayesian model averaging estimate of cross-country growth regressions. *Economic Modelling* 33, 867–883. URL: <https://www.sciencedirect.com/science/article/pii/S026499931300206X>, doi:<https://doi.org/10.1016/j.econmod.2013.05.014>.
- Fernandez, C., Ley, E., Steel, M.F., 2001a. Benchmark priors for Bayesian model averaging. *Journal of Econometrics* 100, 381–427.
- Fernandez, C., Ley, E., Steel, M.F., 2001b. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* 16, 563–576.
- George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88, 881–889.
- Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 711–732.
- Green, P.J., 2003. Trans-dimensional Markov chain Monte Carlo. *Highly Structured Stochastic Systems* Ch.6, 179–198.
- Hastings, W.K., 1970. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* 57, 97–109.
- Heckman, J.J., Ichimura, H., Todd, P.E., 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies* 64, 605–654.
- Heim, B.T., 2007. The effect of tax rebates on consumption expenditures: evidence from state tax rebates. *National Tax Journal* , 685–710.

- Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. *Statistical Science* , 382–401.
- Holmes, C.C., Held, L., 2006. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* 1, 145–168.
- Horvath, R., 2011. Research & development and growth: A Bayesian model averaging analysis. *Economic Modelling* 28, 2669–2673. URL: <https://www.sciencedirect.com/science/article/pii/S0264999311001994>, doi:<https://doi.org/10.1016/j.econmod.2011.08.007>.
- Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J.R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M., Stopnitzky, Y., 2021. The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry* 59, 944–960. doi:<https://doi.org/10.1111/ecin.12992>.
- Jappelli, T., Pistaferri, L., 2010. The consumption response to income changes. *Annual Review of Economics* 2, 479–506.
- Jappelli, T., Pistaferri, L., 2017. *The economics of consumption: theory and evidence*. Oxford University Press.
- Johnson, D.S., Parker, J.A., Souleles, N.S., 2006. Household expenditure and the income tax rebates of 2001. *American Economic Review* 96, 1589–1610.
- Kaplan, D., Chen, J., 2012. A two-step Bayesian approach for propensity score analysis: simulations and case study. *Psychometrika* 77, 581–609.
- Kaplan, D., Chen, J., 2014. Bayesian model averaging for propensity score analysis. *Multivariate Behavioral Research* 49, 505–517.
- Kaplan, G., Violante, G.L., 2014. A model of the consumption response to fiscal stimulus payments. *Econometrica* 82, 1199–1239.

- Kitagawa, T., Muris, C., 2016. Model averaging in semiparametric estimation of treatment effects. *Journal of Econometrics* 193, 271–289.
- Lamnisos, D., Griffin, J.E., Steel, M.F., 2009. Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations. *Journal of Computational and Graphical Statistics* 18, 592–612.
- Lamnisos, D., Griffin, J.E., Steel, M.F., 2013. Adaptive Monte Carlo for Bayesian variable selection in regression models. *Journal of Computational and Graphical Statistics* 22, 729–748.
- Lucchetti, R., Pedini, L., 2021. Parma: parallelised Bayesian model averaging for generalised linear models. *Journal of Statistical Software* forthcoming.
- Madigan, D., York, J., Allard, D., 1995. Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique* , 215–232.
- Man, G., 2015. Competition and the growth of nations: International evidence from bayesian model averaging. *Economic Modelling* 51, 491–501. URL: <https://www.sciencedirect.com/science/article/pii/S0264999315002242>, doi:<https://doi.org/10.1016/j.econmod.2015.08.012>.
- Millimet, D.L., Tchernis, R., 2009. On the specification of propensity scores, with applications to the analysis of trade policies. *Journal of Business & Economic Statistics* 27, 397–415.
- Neri, A., Rondinelli, C., Scoccianti, F., 2017. Household spending out of a tax rebate: Italian “€80 tax bonus”. Technical Report 379. Bank of Italy (Occasional Papers).
- Parker, J.A., 1999. The reaction of household consumption to predictable changes in social security taxes. *American Economic Review* 89, 959–973.
- Parker, J.A., Souleles, N.S., Johnson, D.S., McClelland, R., 2013. Consumer spending and the economic stimulus payments of 2008. *American Economic Review* 103, 2530–53.

- Raftery, A.E., 1995. Bayesian model selection in social research. *Sociological Methodology* , 111–163.
- Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.
- Rosenbaum, P.R., Rubin, D.B., 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79, 516–524.
- Rubin, D.B., Thomas, N., 1996. Matching using estimated propensity scores: relating theory to practice. *Biometrics* , 249–264.
- Sala-i-Martin, X., Doppelhofer, G., Miller, R.I., 2004. Determinants of long-term growth: a bayesian averaging of classical estimates (bace) approach. *American Economic Review* 94, 813–835. URL: <https://www.aeaweb.org/articles?id=10.1257/0002828042002570>, doi:10.1257/0002828042002570.
- Shapiro, M.D., Slemrod, J., 2003a. Consumer response to tax rebates. *American Economic Review* 93, 381–396.
- Shapiro, M.D., Slemrod, J., 2003b. Did the 2001 tax rebate stimulate spending? Evidence from taxpayer surveys. *Tax Policy and the Economy* 17, 83–109.
- Shapiro, M.D., Slemrod, J., 2009. Did the 2008 tax rebates stimulate spending? *American Economic Review* 99, 374–79.
- Souleles, N.S., 1999. The response of household consumption to income tax refunds. *American Economic Review* 89, 947–958.
- Souleles, N.S., 2002. Consumer response to the Reagan tax cuts. *Journal of Public Economics* 85, 99–120.
- Steel, M.F., 2020. Model averaging and its use in economics. *Journal of Economic Literature* 58, 644–719.

- Wilcox, D.W., 1989. Social security benefits, consumption expenditure, and the life cycle hypothesis. *Journal of Political Economy* 97, 288–304.
- Xie, Y., Zhu, Y., Cotton, C.A., Wu, P., 2019. A model averaging approach for estimating propensity scores by optimizing balance. *Statistical Methods in Medical Research* 28, 84–101.
- Zellner, A., 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti* 6, 233–243.
- Zigler, C.M., Dominici, F., 2014. Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association* 109, 95–107.