DEPARTMENT OF ECONOMICS

WORKING PAPER SERIES

2016-12

McMASTER UNIVERSITY

Department of Economics
Kenneth Taylor Hall 426
1280 Main Street West
Hamilton, Ontario, Canada
L8S 4M4

http://www.mcmaster.ca/economics/

Playing by the rules? Agreement between predicted and observed binary choices

Stephanie Thomas, Ph.D.

McMaster University
Department of Economics

Abstract

Empirical economics frequently involves testing whether the predictions of a theoretical model are realized under controlled conditions. This paper proposes a new method for assessing whether binary ('Yes'/'No') observations ranging over a continuous covariate exhibit a discrete change which is consistent with an underlying theoretical model. An application using observations from a controlled laboratory environment illustrates the method, however, the methodology can be used for testing for a discrete change in any binary outcome variable which occurs over a continuous covariate such as medical practice guidelines, firm entry and exit decisions, labour market decisions and many others. The observations are optimally smoothed using a nonparametric approach which is demonstrated to be superior, judged by four common criteria for such settings. Next, using the smoothed observations, two novel methods for assessment of a step pattern are proposed. Finally, nonparametric bootstrapped confidence intervals are used to evaluate the match of the pattern of the observed responses to that predicted by the theoretical model. The key methodological contributions are the two innovative methods proposed for assessing the step pattern. The promise of this approach is illustrated in an application to a controlled experimental lab data set, while the methods are easily extendable to many other settings. Further, the results generated can be easily communicated to diverse audiences.

JEL Classification: C18, C14, C4, C9
Keywords: Evaluation of theoretical predictions, binary outcome data, applied nonparametric analysis, data from experiments

1 Introduction

The objective of this work is to determine whether data collected from economic activities support the predictions derived from a theoretical model of economic behaviour. The specific prediction described here is a step-wise relationship between a binary 'Yes'/'No' outcome variable, and two explanatory variables. For example, a consumer may decide to purchase a product over a range of low prices, so the outcome is 'Yes', and decide not to purchase at prices above a certain price threshold, so the outcome becomes 'No' from there on. In the raw form, the observations would include some noise due perhaps to impulse purchases or lack of attention. Using a standard parametric framework for smoothing such noise the location of a discrete change is masked, prompting the investigation of alternative approaches to evaluation. Using flexible nonparametric regression opens the possibility of locating a discrete change or 'switch-point' within the smoothed observations. Three methods for locating candidate switch-points are suggested. The final proposed framework for evaluation uses the nonparametric smoothing approach combined with a maximum absolute gradient switch-point candidate identification strategy. The candidate switch-points are compared to the predictions of the theoretical model via constructing nonparametric bootstrapped confidence intervals which acknowledge the interdependence of the observations.

The theoretical model and observations used to illustrate the techniques discussed here are taken from a single treatment of the experiment reported in Buckley et al. (2015). The observations consist of a set of observed decisions to participate or not participate in an activity and are conditioned on two explanatory variables. The methods developed here maintain as strict an independence as possible between the observations and the predictions of the theoretical model. This means that rather than attempting to explain the matches of the theory with the observations, the theory is defined first, and, if necessary, the observations are secondly smoothed using a regression framework. The first and second parts are then compared. This means that if the researcher wishes to compare a different theory to the observations it is straightforward to do so and requires no alteration to the description of the smoothed observations. It also means that if one suspects that the smoothed observations are improperly described that the regression framework can be altered independently of the theoretical predictions. An advantage of this approach is that the results can be intuitively illustrated, which offers substantial appeal to researchers wishing to communicate with diverse audiences. Applications of the methodological framework extend naturally to archival data and data from field experiments as well as to controlled laboratory experiments. Examples of these applications include decisions to look for a job or not (labor force participation) or adherence to professional practice guidelines in accounting, law or medicine (See Chapter 2 of this thesis).

2 Methodology

Empirical economics frequently involves testing whether the predictions of a theoretical model are realized under controlled conditions. This paper proposes a new method for assessing whether binary ('Yes'/'No') observations ranging over a continuous covariate exhibit a discrete change which is consistent with an underlying theoretical model. An application using observations from a controlled laboratory environment illustrates the method, however, the methodology can be used for testing for a discrete change in any binary outcome variable which occurs over a continuous covariate such as medical practice guidelines, firm entry and exit decisions, labour market decisions and many others. The observations are optimally smoothed using a Nonparametric approach which is demonstrated to be superior, judged by four common criteria for such settings. Next, using the smoothed observations, two novel methods for assessment of a step pattern are proposed. Finally, nonparametric bootstrapped confidence intervals are used to evaluate the match of the pattern of the observed responses to that predicted by the theoretical model. The key methodological contributions are three innovative methods proposed for assessing the step pattern. The promise of this approach is illustrated in an application to a controlled experimental lab data set, while the methods are easily extendable to many other settings.

Once the basic overview of the match with theory is established using a classification matrix approach the discussion moves on to smoothing, comparing three techniques to achieving this objective while incorporating the effects of two covariates. The first technique, an 'Empirical approach,' simply calculates basic proportions. The next is the standard parametric technique in which a probit estimation strategy is employed.[1] The last is a Nonparametric approach in which the conditional density of the positive participation decisions is estimated. An Appendix compares the approaches in detail. The results of each smoothed model are then compared to the theoretical predictions and statistical significance established using a bootstrapped confidence interval approach.

A key weakness of the parametric approach is illustrated here. The parametric technique suggests very tight confidence intervals but is clearly misspecified. Using the Nonparametric approach the observations demonstrate reasonable support for the theoretical model. Evidence for a match of the candidate switch-points with those suggested by the theoretical model is found in most instances using the final proposed framework. Section 3 describes

[1] All results were also carried out using a logit technique with virtually indistinguishable results.

the theory and observations, Section 4 deals with smoothing the observations. Section 5 proposes two new methods for identifying switch-points, and Section 6 constructs nonparametric bootstrapped confidence intervals to evaluate the match of the candidate switch-points with the predictions of the theoretical model. Section 7 concludes, discussing alternatives and extensions of the framework.

The techniques explored here were developed in response to a particular situation arising in the experimental lab in which a theoretical model suggested that the outcome of interest would exhibit a clearly defined cutoff. There were few well defined options for accepting or rejecting the suggestions of the theoretical model, and the more standard options produced uninformative results. The insights gathered in the process of investigation and presented here offer a new direction for the confrontation of theory with evidence.

In addition to the analysis presented here variants of structural breaks and regression discontinuity were considered, treating the continuous covariate as the variable over which breaks occur, as year does in the familiar macroeconomic sense of structural breaks. In the case of testing for unknown structural breaks the single model of the relationship between the outcome and the two covariates proposes 5 breaks in the continuous covariate, each dependent upon the ordered covariate level. Testing for every possible combination of breaks implies 118,755 tests.[2] Taking the approach of a known break one could also simply suppose the true breaks to be those of the theoretical model and test for a match. In the macroeconomic context, this approach suggests that the model takes on a different form before and after the break. In the case of a binary outcome variable this implies a model for the positive outcomes of *participate* and another for the negative outcomes. Conceptually this would imply a belief that the covariates have different influences upon the participation decision

[2]The number of ways to choose 5 break-points from 29 possibilities

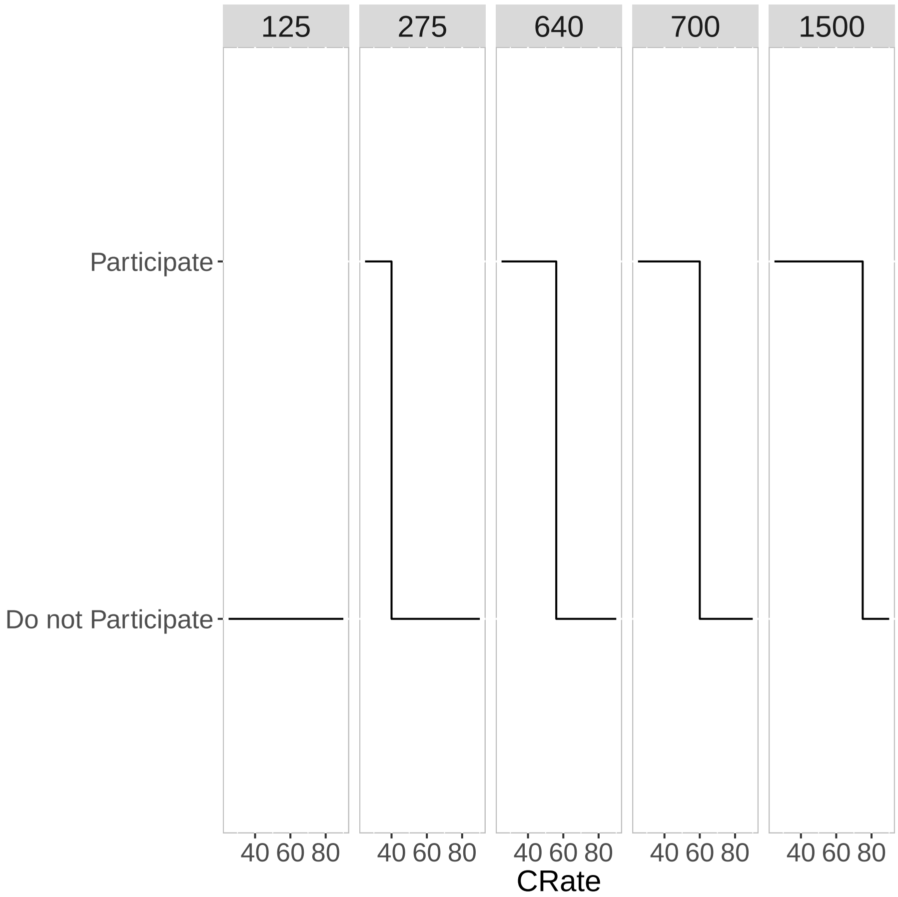
based on whether the particular participation decision is taken before or after a certain level in the continuous covariate. Similarly, regression discontinuity suggests that the application of a treatment effect at a known break has the potential to result in two different models before and after the break. In this experiment participants chose the outcome variable based on the two covariates where the continuous covariate was determined in the experiment and not applied as a treatment variable. As an adaptation, the match of the observations against all possible cutoffs is provided in the paper, using the goodness of fit metrics of Correct Classification Ratio, Adjusted Correct Classification Ratio, The Area Under the Receiver Operator Characteristics Curve and Cohen's κ. Despite being able to select a maximum value in order to identify a candidate switch point, the true behavior of the observations is masked by this technique because a break in the sense of a clear shift in the observations may or may not actually exist.

3 Theoretical predictions versus observations

3.1 A theoretical model of participation

The theory and observations forming the basis of the example illustrated in this paper were drawn from an experiment presented in Buckley et al. (2015). The experiment sought to explore the role of mixed finance arrangements whereby a private good is funded publicly but also available to purchase privately. Participants in the experiment were asked to submit a preferred contribution rate to a fund financing public provision of the good, and then contributed to the public fund at a rate determined by the median of the preferred contribution rates. This rate is the tax rate that is applied to the incomes of all participants in the session and provides resources to publicly provide the private good. Any income remaining could be used to purchase additional amounts of the private good in a manner defined by the particular treatment. The public fund was invested and each participant then received an equal share of the total fund, i.e. a private good. For the purpose of this investigation only the theory and observations of the 'top-up' treatment in Buckley et al. (2015) are used. The focus here is on the development of a method for determining whether or not behavior within a treatment supports a theoretical proposition. In this context, participation, referred to as '*participate*' in this paper, is defined as 'topping-up' in Buckley et al. (2015); that is, purchasing an amount of the good privately in addition to the level provided through the public system. Non-participation is defined as consuming only the amount provided publicly. The decision to participate is dependent upon *'income'* and the tax rate (labelled here as the contribution rate *'CRate'*).

The theoretical model of top up behaviour outlined in Buckley et al. (2015) describes a pattern of participation decisions which are dependent upon income and the rate at which all the participants in a group contributed to the public fund. For each level of income in the model there is a contribution rate at which participation switches from being an optimal to a non-optimal decision. The theoretical predictions thus follow, for each level of income, a pattern of participation in privately topping up the consumption of the publicly provided private good at low CRates and then stopping this topping up once the CRate reaches a sufficiently high value. This creates a distinct 'step' or 'switch' in the relation between topping up and not topping up as the CRate rises. The step patterns in Figure 1 illustrate the relationship. As income level increases the contribution rate at which participation becomes non-optimal increases.



Source: Buckley et al. (2015).

Figure 1: Predictions of the theoretical model of participation by income and contribution rate.

3.2 Observations

The set of observations consists of 500 participation decisions along with the associated income levels and contribution rates. In each period subjects first learned their income level (*'income'*) and then submitted their preferred rate of contribution to the public fund. The median of the submitted rates was selected as the *'CRate'* for the group and the *CRate* proportion of income deducted from each participant's funds. Each member of the group was free to purchase additional investments independently of the group fund with the remaining proportion of their income as part of the 'top-up' treatment. If a participant made any additional purchases the variable *'participate'* was classed 'Participate', and as 'Do not Participate' otherwise.[3]

The data set includes decisions made by 50 individuals for each of 10 decision periods. The experiment was designed as a repeated one-shot game and no statistically significant period effects were reported in Buckley et al. (2015). Voting to determine the contribution rate to the public fund took place in 5-person groups. No statistically significant group effects were identified by Buckley et al. (2015). An independent observation of a determination of the CRate is defined as the result of a 5 person group in each period of the experiment, so the experiment contains 100 independent observations.

The outcome variable *participate* is an unordered factor variable taking on a value of 'Participate' if the individual purchases a nonzero amount of the good privately, and 'Do not Participate' otherwise. *Income* is a discrete ordered variable taking on values {125, 275, 640, 700, 1500} which were randomly assigned to individuals and were distributed so as to ensure that no two group members experienced the same level of income within a period. Each participant experienced each level of income twice during the 10 decision periods but was not informed of which level of income would occur prior to the start of any period and so experienced the assignment of income in a random manner. *CRate* is a continuous variable which could take on any value in the range $[0, 100]$. In the data set we observe 29 unique values in the range $[25, 90]$. Table 1 provides a summary of the data. The total number of observations is 500. 229 participant decisions involved purchasing a positive amount of private investment, 271 did not participate in private purchasing. There are 100 observations at each level of income. Contribution rates to the public fund ('CRate') range from 25 to 90, with mean 55.3 and median 54.5. See Appendix Section A Table A.1 for

[3]The sum of the contributions to the public fund was invested and the fund increased in a pre-defined manner with $\frac{1}{5}$ of the total returned to each participant. Private purchases were also augmented in the same way but were not shared among group members.

the frequency of observations of *CRate*.

participate	n	Income	n	CRate	value
		125	100	Minimum	25
Participate	229	275	100	Maximum	90
Do not Particicpate	271	640	100	Mean	55.3
		700	100	Median	54.5
		1500	100		

Table 1: Summary of observations.

3.3 Matching theoretical predictions and observations

		Observation		
		Do not Participate	Participate	Total
Theoretical Prediction	Do not Participate	223	54	277
	Participate	48	175	223
	Total	271	229	500

Table 2: Classification matrix of observations and theoretical model predictions.

A basic way to summarize the overall match of the collected observations with the predictions of the theoretical model can be done simply by classifying each observation based on whether or not the observation is in agreement with the appropriate theoretical prediction. There are four possible classes.

Class 1: the theoretical model predicts participation and participation was observed,

Class 2: the theoretical model predicts no participation and participation was not observed,

Class 3: the theoretical model predicts participation and participation was not observed ('under participation'), and

Class 4: the theoretical model predicts no participation and participation was observed ('over participation').

Table 2 presents the number of observations in each of the described classes in tabular form, commonly referred to as a 'classification matrix' or 'confusion matrix'. In this case 48 of a total 500 observations qualify as under-participation (10%) while 54 observations qualify as over-participation (11%).

To evaluate the fit of the observations with the predictions of the theoretical model, the Correct Classification Ratio (CCR) is frequently used as a measure of accuracy. This measure is simply the proportion of observations which match the predictions of the theoretical model (i.e. Class 1 and Class 2). In this case, this refers to the accuracy of the theoretical model at producing predictions which describe the observations collected in the experiment. Here 80% of observations are in agreement with the predictions of the theoretical model. The CCR can be biased, however, because even if chosen by chance, there is a higher probability of simply choosing the most frequent outcome and being correct. Adjusting for the probability of choosing the most frequent outcome by chance this value falls to 55% using the adjusted correct classification ratio (adj-CCR, defined in Section 4.4). Another measure used to describe agreement between the observations and predictions is the the area under the receiver operator characteristics curve (AUC), which is described in detail in Section 4.4.2. AUC values range from 0.5 to 1, with higher values indicating greater levels of agreement between observations and predictions. In this case the value of the AUC is 0.79, which suggests substantial agreement between the observations and theoretical predictions, and is nearly identical to the CCR result. Finally, Cohen's κ describes the amount of agreement between the theoretical predictions and the observations beyond that occurring by chance, where a value of 1 indicates perfect agreement and 0 no agreement (Cohen, 1960). The value of Cohen's κ in this case is 59%, which is similar to the adj-CCR result and serves to confirm substantial agreement beyond random chance between the theoretical predictions and the observations.

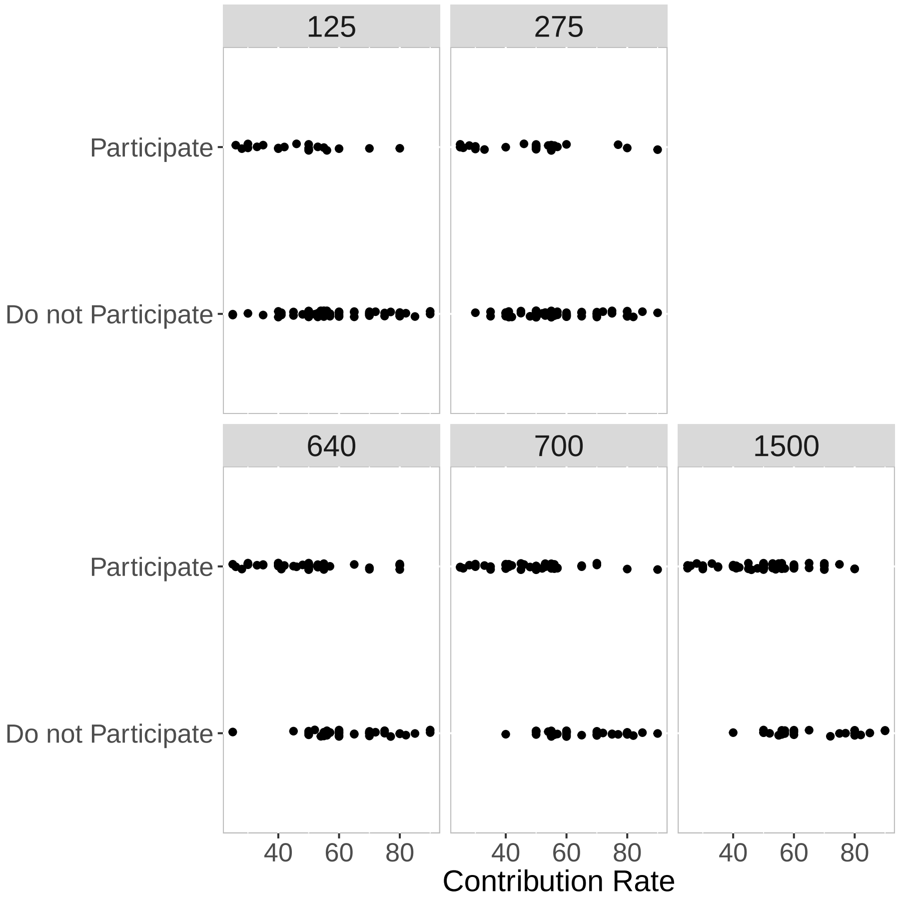
Of the subset of predictions which are in disagreement, 53% are cases which qualify as 'over participation' and 47% qualify as 'under participation'. A one sample proportions test with continuity correction fails to reject a null of equal proportions (p-value 0.6205), indicating that there is no reason to suspect any systematic tendency towards over- or under- participation among decisions which are not in agreement with the theoretical model.

The classification matrix technique is a simple means to assess the overall fit of the observations with the predictions of the theoretical model, but it does not address the relative influence of the explanatory variables upon the outcome or whether the observations do in fact exhibit a discrete 'step'. The remainder of this chapter will explore the relationship between *participate*, *income* and *CRate* in greater detail.

4 Observations: Describing the relationship between *participate*, *income* and *CRate*

This section presents three approaches to investigating the relationship between *participate*, *income* and *CRate*. The first method, the 'Empirical' approach, simply examines the frequency of participation at each combination of *income* and *CRate*. The second method, the 'Standard' approach, attempts to estimate a line of best fit using Probit regression. The third approach, the 'Nonparametric', estimates the relationship by nonparametric conditional density estimation which calculates an optimal bandwidth and uses kernel regression. Goodness of fit is assessed using four commonly used criteria: the adjusted correct classification ratio (adj-CCR), Cohen's κ (Cohen, 1960), area under the receiver operator characteristics curve (AUC) and Youden's J (Youden, 1950). Each criteria is described in detail in Section 4.4. Of the smoothing strategies considered only the Nonparametric approach is capable of unmasking switch-points in the data. The Nonparametric approach also dominates the Standard approach in terms of fitting the observations on all four criteria.

The purpose of investigating the relationship of *participate* with *income* and *CRate* in the observations is ultimately to assess whether the pattern of observations is similar to the step pattern suggested by the predictions of the theoretical model. A conclusive step pattern would present as two distinct groups of observations which do not overlap across *CRate* for each level of *income* and all that would be required would be to determine the *CRate* at which the observations of *participate* 'switch' from 'Participate' to 'Do not Participate'. Figure 2 presents the observations of the experiment and offers motivation for smoothing. In order to facilitate a visual assessment the observations have been jittered vertically in order to show multiple observations which occur at the same *CRate*. Each level of *income* is presented in a different pane. Despite these two visual adjustments it is clear that it is not possible to identify a clear 'switch' from 'Participate' to 'Do not Participate' for each level of *income*. The main objective of this section is to condense the observations into a form conducive to identifying a candidate switch-point, should it exist.



Source Buckley et al. (2015). Slight vertical jittering (displacement) of points to show multiple observations.

Figure 2: Participation observations by contribution rate and income.

4.1 The Empirical Approach

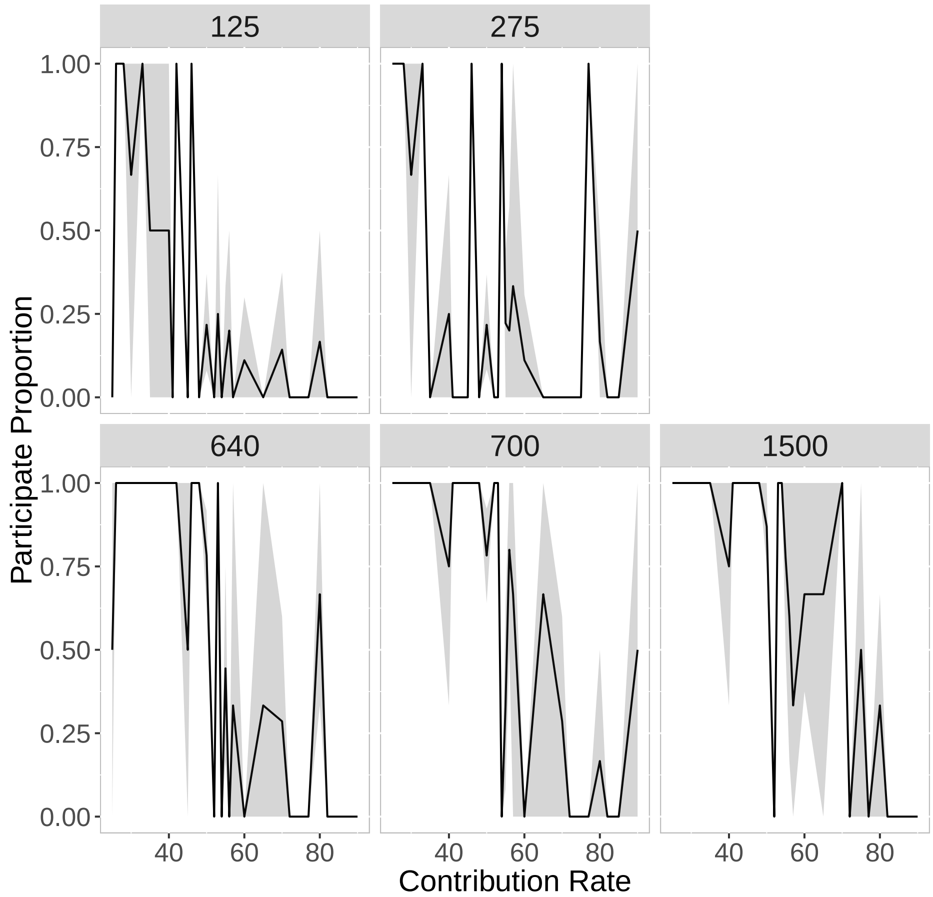
A Pearson Chi-squared test is used as a preliminary assessment of the existence a relationship between *participate*, *income* and *CRate*. This test proposes a null hypothesis of independence among all the three variables; failure to reject this null indicates a lack of a relationship. Here the null is rejected (p-value of 0.0000) suggesting that a relationship in fact exists.

The Pearson Chi-squared test is a frequency-based test which relies on comparing the observed frequencies with the expected frequencies if the null were true and no relationship were to exist. In this case *income* has 5 levels and

CRate is treated as a discrete variable of 29 values, so there are 145 frequencies to compare. In order to avoid a biased result the number of observations in each *income* - *CRate* cell should be greater than 5 as per the specifications of the Chi-square test. Table A.1 in the Appendix Section A shows that this is not the case for the observations at hand; many of the cells contain only one observation. As a rudimentary remedy to achieve the necessary minimum of 5 observations in each cell the Chi-squared test was conducted over a grouped *CRate* which was arbitrarily categorized into classes with values of less than or equal to cutoffs of $\{30, 40, 50, 60, 70, 80, 90\}$.[4] The result is a rejection of the null of independence among the variables (p-value of 0.0000, as reported above). The test using the uncategorized *CRate* provided the same result (p-value of 0.0000). Both results are suggestive of the existence of a relationship.

The first method for condensing the raw observations is done by simply plotting the proportion of *'participate'* decisions which were to positively 'Participate' at each level of *income* and *CRate*. The noisiness of the results is indicative of the sensitivity of this approach to the number of observations at each *income-CRate* cell; nonetheless this is useful for gaining an initial idea of the relationships and suggests that the observations may in fact exhibit 'switch' type patterns. The lines in Figure 3 trace the proportion of positive 'Participate' decisions and suggest that, apart from noise, *participate* may exhibit distinct changes over *CRate* for at least some *income* levels. For instance, in the 125 pane the probability of *participate* falls sharply at a *CRate* of approximately 50. Similar changes are also visually identifiable in the 640 and 700 panes. The next section will attempt to address this noise using a Standard regression approach.

[4]The optimal bandwidth for 'CRate' is 3.1305, however this bandwidth results in less than the necessary 5 observations in each cell required to conduct the Chi-squared test. The arbitrary cutoff used here therefore represents an 'over-smoothed' comparison case.



Solid line is the proportion of 'Participate' outcomes. Grey shaded area is the bootstrapped 90 percent confidence interval.

Figure 3: Proportion of participation by contribution rate and income using the Empirical approach.

4.2 The Standard approach

The main weakness of the Empirical approach is that the results are so noisy that multiple 'switches' in *participate* could potentially be identified for each *income* level. To smooth out this noise, the Standard approach described here employs Probit regression. Probit regression is among the most frequently used regression frameworks for estimating binary outcomes and so serves as a reference point for a broad audience of analysts. In this instance the approach treats *CRate* as a continuous variable, overcoming the pitfall of having of too few observations in each *income*-*CRate* cell, and estimates a line of best fit which is much smoother than that of the Empirical approach. The reduced noise should assist in identifying a unique candidate switch-point. This regression strategy is appropriate to the task of estimating proportions and can be alternatively interpreted as estimating the predicted probability of the decision *participate* being 'Participate' [5]. The drawback of this approach, as will be shown, is that the results are so smooth that switch-points are masked completely.

The conditional probability $\widehat{participate}$ is defined here by:

$$Pr(Y = 1|X) = \Phi(X'\hat{\beta}), \tag{1}$$

where in this case X is composed of the two explanatory variable vectors $X = (X_1 = income, X_2 = CRate)$ and Y is the binary outcome variable *participate* which is conditional on X. $X'\beta$ is referred to as the index function and the results are estimated by maximum likelihood estimation. Φ is the standard normal cumulative density function and is used to ensure that the predicted probabilities lie within the range $[0, 1]$. The complete details are provided in Appendix Section B.

	Coefficient	Std. Error	z value	Pr(>|z|)
(Intercept)	1.44	0.30	4.82	0.00
CRate	-0.04	0.01	-8.41	0.00
Income275	0.11	0.21	0.52	0.61
Income640	1.03	0.20	5.05	0.00
Income700	1.18	0.20	5.76	0.00
Income1500	1.69	0.21	7.87	0.00

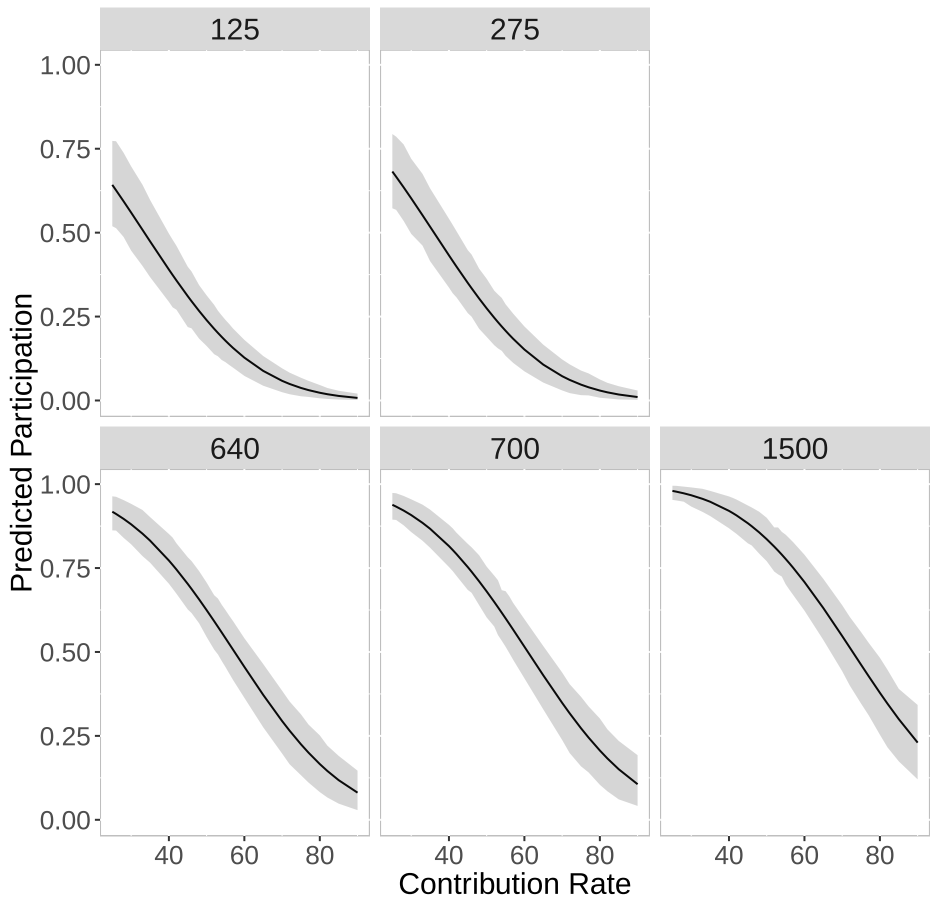
Table 3: Probit regression estimates

Table 3 provides the regression estimates [6]. The outcome exhibits significant differences from the reference income level of 125 when participants face the

[5]Logistic regression for the odds of 'Participate' returned virtually identical results.

[6]The estimation is done using R's glm function in the stats package R Core Team (2015), or the mfx package Fernihough (2014) which also provides marginal effects.

income levels of 640, 700 and 1500 but not the 275 income level. Both both the sign and ordering of the magnitudes are as expected. An income of 1500 has a larger effect on the estimated probability of participation than the 700 income level. The 700 income level, in turn, has a larger effect on the estimated probability than the 640 income level, and so on. The variable *CRate* also has the expected sign and is significant. As *CRate* increases the estimated probability of participation decreases.



The solid line is the probability of the 'Participate' outcome. The grey shaded area is the bootstrapped 90 percent confidence interval.

Figure 4: Predicted probability of participation by contribution rate and income using the Standard approach.

The estimated probability of participation, $\widehat{participate}$, is illustrated for each *income* level in Figure 4, along with the bootstrapped confidence intervals (discussed in more detail in Section 6). This figure demonstrates the exceptional smoothing achieved by the Standard approach. The multiple potential switches identified by the Empirical approach are now completely masked; no distinct switches are observable. Taking the results from Table 3 and Figure 4 together it is reasonable to infer that the probability of *participate* is declining as *CRate* increases and increasing as *income* increases.

For the probit model the predicted probabilities follow a smooth pattern by design, as dictated by the parametric structure. This is of no consequence if the population from which the data are drawn in fact follow this exact distribution. If our sample of observations is, however, not drawn from a population specified precisely by the parametric form estimated in Equation 1 then any inference derived from this model is misleading. Based on this fact one might be concerned with whether the probit model is a reasonable approximation to the population from which the data are drawn. A detailed examination of (pseudo) coefficients of determination which attempt to measure the amount of variation in the outcome (*participate*) attributable to variation in the explanatory variables *income, CRate*, is provided in Appendix Section D. As well, Wald tests for the joint significance of all variables, income variables alone, and contribution rate alone are all rejected confirming that the variables are jointly significant. According to these commonly used metrics this model appears to deliver a fairly good fit to the observations, however, it does not suggest the discrete changes predicted by the theoretical model. Nor do any of these common tests target misspecification directly. Running a variant of Ramsey's RESET Test (Ramsey, 1969) suggested by Ramalho and Ramalho (2012) on the function described by:

$$Pr(Y = 1|X, \hat{y}^2) = \Phi(X'\beta + \theta\hat{y}^2) \tag{2}$$

reveals that the square of the fitted values of the original regression is significant (i.e. θ has a value of 5.43 and a p-value 2.4×10^{-4}) indicating that the null of correct specification should be rejected in favour of the alternative: that the model is misspecified. This invalidates any inference based upon this model because it is likely biased and inconsistent. In addition, confidence intervals are not proper confidence intervals since they are centered on a biased estimate.

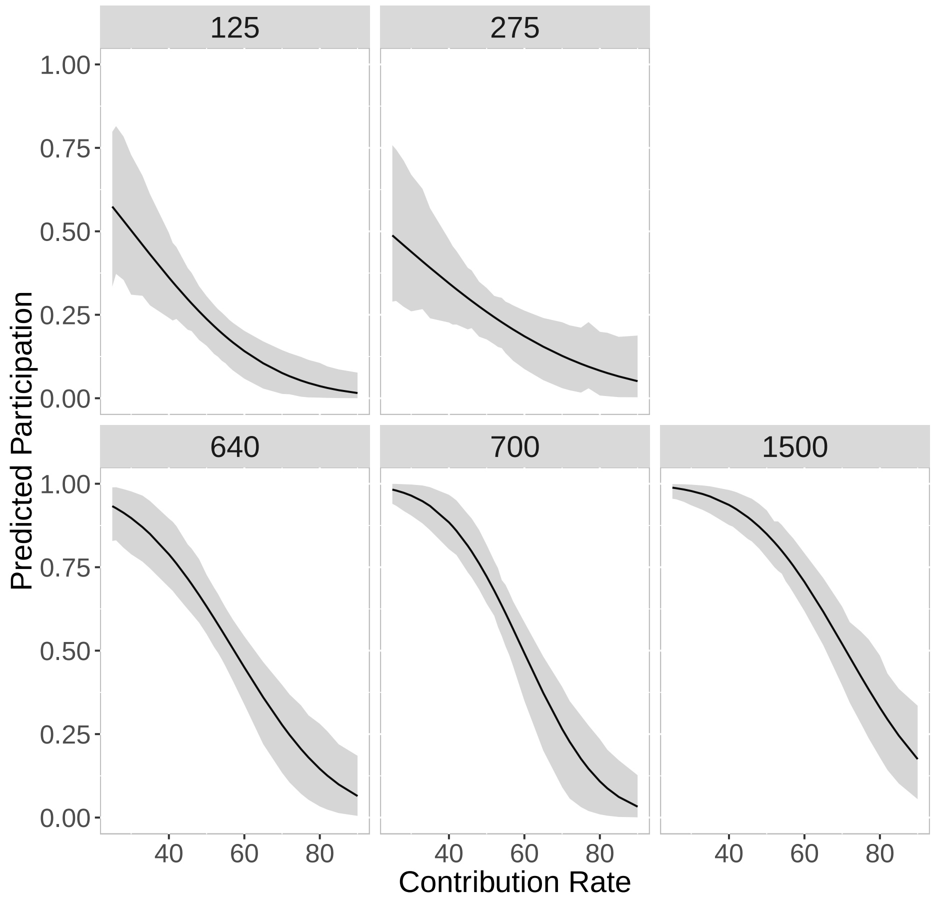
4.2.1 Including interactions

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.0883	0.6282	1.73	0.0832
CRate	-0.0361	0.0120	-3.01	0.0026
Income275	-0.5020	0.8540	-0.59	0.5566
Income640	1.5701	0.8858	1.77	0.0763
Income700	2.5412	0.9612	2.64	0.0082
Income1500	2.4063	0.9482	2.54	0.0112
CRate:Income275	0.0114	0.0160	0.71	0.4781
CRate:Income640	-0.0104	0.0163	-0.64	0.5247
CRate:Income700	-0.0248	0.0175	-1.42	0.1562
CRate:Income1500	-0.0132	0.0168	-0.78	0.4325

Table 4: Probit regression estimates with interaction terms

The RESET test run in the previous section suggests that the probit model is misspecified. While there is no way of knowing the particular form of the misspecification suggested by the RESET result, one possibility is that the model failed to account for potential interactive effects of *income* and *CRate*. Including this interaction changes the index function but does not improve the situation. Table 4 shows the results with the addition of an interaction between *income* and *CRate*. Again, in order to interpret the coefficients in Table 4 as probabilities the index function is distributed according to the cumulative normal distribution function. The results are similar to those without interactions, none of the interactive terms are significant. Reading the coefficients, including interactions, does not assist in describing the relationship.

Figure 5 illustrates the predictions of the Standard model with interactions. Here the results still fail to clearly delineate a switch-point, a reflection the particular choice of model. The larger confidence bounds suggest that the inclusion of interactions leads to a loss of precision, however, both forms of the parametric approach fail to reject misspecification, which suggests that the results are inconsistent in both cases.



The solid line is the probability of the 'Participate' outcome. The grey shaded area is the bootstrapped 90 percent confidence interval.

Figure 5: Predicted probability of participation by contribution rate and income using the Standard approach with interactions.

Comparing the results of the approaches without and with interactions, McFadden's Adjusted R^2 value (McFadden, 1973) is 0.24 for the model without interactions and 0.23 with interactions, indicating that the inclusion of interactions does not assist in explaining variation in the decision to participate. The R^2 values of Cragg and Uhler (1970) are also commonly used to compare approaches and tell a similar story. The values are 0.4 for the model without interactions and 0.41 with interactions included, indicating that the approach with interactions does not offer much improvement in explanatory power over the model without interactions.

McFadden's and Cragg and Uhler's R^2 values are applicable only to the parametric approaches, whereas the Adjusted Count R^2_{AC} value has the advantage that it does not depend on the approach used and so can be directly compared regardless of the approach to estimation. This metric is essentially the same as the adj-CCR described in Section 4.4, except that in this case the values of $\widehat{participate}$ are compared to the observed values of *participate* by classifying the $\widehat{participate}$ with values greater or equal to 0.5 as positive 'Participate' decisions and all others as negative 'Do Not Participate' decisions. The statistic therefore summarizes the degree of match between the classified predictions and the observations. The value of this statistic is 0.5 without the interactions and 0.52 with interactions included, indicating a better fit in the case of the approach with interactions. Overall the difference between the approaches with and without interactions is small. When taken in conjunction with the lack of significance of the interaction terms, the overall results are supportive of concluding that the interaction terms are uninformative.

Another means of comparing results across models is via Akaike's Information Criterion (AIC) (Akaike, 1974) which describes the amount of information lost by using a model to describe a set of data. For the model with interactions the value is 526.19 which is larger than that of the probit estimation without the interaction term (524.36), indicating that including the interaction, while offering a worse fit according to McFadden's Adjusted R^2, a better fit according to Cragg and Uhler's R^2, and a better fit according to the Adjusted Count R^2_{AC} results in a greater loss of information. The more parsimonious form, without interactions is preferable, indicating that the interaction terms possibly introduce a degree of multicollinearity. In support of the notion that the model is misspecified, the Ramsey RESET test variant again rejects the null of correct specification.[7].

A Wald test for the joint insignificance of the interaction coefficients confirms that these terms are not relevant in determining the participation decision, while the remaining coefficients remain jointly significant. Including the interaction term suggests about as good a fit to our data as the specification without interactions. The higher R^2 and AIC values combined with the rejection of the Ramsey RESET test variant suggest that the effect of *CRate* and/or *income* upon the *participate* decision is potentially more nonlinear than the specified probit model. Adding higher order terms in addition to, or instead of, the interaction term could improve the fit of this model. However, if we begin to adjust our model in order to achieve better results we run the risk of forcing the data to tell us the story we want to hear.

The Standard approaches illustrated here produce smooth declines in the predicted values of $\widehat{participate}$. These smooth declines suggest rejection of a hypothesis of switching patterns in the observations. Yet, the smoothness is largely the result of choosing to employ the probit technique, which, as was shown, is not a correct specification of the relationship between the variables at hand. In the next section an alternative smoothing technique will be explored, and in Section 4.4 a comparison of all the approaches is presented.

[7] The square of the predicted values in the regression (as in equation 2) is significant (p-value 0.01365)

4.3 The Nonparametric approach

The Standard approach investigated in the previous section smoothed the observations but did so in such a way as to completely mask the switch patterns observed under the unsmoothed Empirical approach. This section considers a Nonparametric alternative which smooths noise effectively while revealing switch patterns in the observations. Nonparametric regression presents a robust alternative to the Standard parametric approaches; along with being insensitive to a small proportion of outliers in the observations, these methods circumvent issues of model misspecification and have excellent in-sample fit. Because a specific form is not specified at the outset of the investigation the estimates of $\widehat{participate}$ may take on any shape, including those of the Standard approach. For experimentalists seeking to investigate relationships within relatively small, but carefully collected data sets these features are particularly attractive and simple to implement.

The Nonparametric approach relies upon the data at hand to form predictions, smoothing weighted observations within small sections of data called bandwidths. First an optimal bandwidth for each variable is calculated by minimizing a cross validation function. Then, the observations within each bandwidth are weighted according to a specified weight function and combined to produce a product kernel. This approach automatically takes into account any interactions and has the ability to exclude variables which are not relevant. By circumventing the need to choose the form of an estimating model this approach avoids issues of misspecification while retaining the ability to reproduce the results of any Standard approach. The nonparametric alternative of conditional density estimation as described first by Stone (1977) and more recently by Hall, Racine, and Li (2004) is implemented here in R using the np package developed by Hayfield and Racine (2008).

The problem at hand is to estimate the conditional density function $g(y|x) = \frac{f(x,y)}{\mu(x)}$ where $f(x,y)$ is the joint probability distribution of the outcome y and explanatory variables x and $\mu(x)$ is the mean of explanatory variables x . This is done via estimating the function:

$$\hat{g}(y|x) = \frac{\hat{f}(x,y)}{\hat{\mu}(x)} \tag{3}$$

Under the approach described by Li and Racine (2007) the numerator and denominator of the conditional probability function are described by:

$$\hat{f}(x,y) = \frac{1}{n}\sum_{i=1}^{n} K_{\gamma}(x, X_i) k_{\lambda_0}(y, Y_i) \tag{4}$$

$$\hat{\mu}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{\gamma}(x, X_i), \tag{5}$$

with $K_{\gamma}(x, X_i)$ and $k_{\lambda_0}(y, Y_i)$ representing kernel density functions.

In this study the unordered dependent variable *participate* is estimated using the kernel suggested by Aitchison and Aitken (1976) and defined by:

$$\begin{aligned} k_{\lambda_0}(y, Y_i) &= l(Y_{is}, y_s, \lambda_s) \\ &= \begin{cases} 1 - \lambda_s & \text{if } Y_{is} = y_s \\ \frac{\lambda_s}{c_s - 1} & \text{if } Y_{is} \neq y_s \end{cases}, \end{aligned} \tag{6}$$

where y_s can take on c_s ordered values $0, 1, c_s - 1$. If $\lambda_s = 0$ then $l(Y_{is}, y_s, \lambda_s) = 1$ is an indicator function, and if $\lambda_s = \frac{c_s - 1}{c_s}$, then $l(Y_{is}, y_s, \frac{c_s - 1}{c_s}) = \frac{1}{c_s}$, a constant. Thus the range for the smoothing parameter associated with *participate* is $[0, \frac{2-1}{2} = 0.5]$.

The dependent variables both enter into the product kernel $K_{\gamma}(x, X_i)$ which takes the general form:

$$K_\gamma(x, X_i) = W_h(x^c, X_i^c)L(x^d, X_i^d, \lambda), \tag{7}$$

where $\gamma = (h, \lambda)$ is a vector of continuous and discrete bandwidths, in this case for *CRate* and *income*. The superscript c denotes the continuous variable *CRate* and d the discrete variable *income*. The ordered levels of *income* are estimated using the kernel proposed by Racine and Li (2004) while the Epanechnikov (1969) kernel is used for the continuous variable *CRate*.

The Racine and Li (2004) kernel is described by:

$$L(x_i^d, x^d, \lambda) \;=\; \begin{cases} 1 \text{ if } |x_i^d - x| & = 0, \\ \lambda^{x_i^d - x} \text{ if } |x_i^d - x| & \geq 1 \end{cases}, \tag{8}$$

where λ must lie between 0 and 1.

The Epanechnikov (1969) kernel is defined by:

$$W(u) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}u^2) & \text{if } u^2 < 5 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$\text{where } u = \frac{x^c - X_i^c}{h}$$

$$\text{and } h > 0,$$

4.3.1 Bandwidth selection

The choice of the particular kernel weight functions has little influence on the results of the nonparametric method while the bandwidth selection method has great impact (Li and Racine, 2007). Two bandwidth selection routines are considered here without altering the chosen kernels in order to investigate the impact of bandwidth selection upon the resulting estimates. Least squares cross validation is the preferred method because it has the ability to remove irrelevant regressors [8] but it can be computationally intensive, and even prohibitive for large data sets [9]. In this case, the routine takes less than one minute. An alternative to least squares cross validation is maximum likelihood cross validation, which can be less computationally intensive but has the drawback that it can oversmooth if the tails of the distribution are fat. This oversmoothing can potentially lead to an inconsistent estimate but it can also be beneficial if it effectively removes irrelevant regressors. The details of each of the methods are included in Appendix Section E.

Variable	Least Squares	Maximum Likelihood	Upper Bound
Participate	0.0000	0.0512	0.5
Income	0.9926	0.9923	1
CRate	3.1305	3.1305	inf

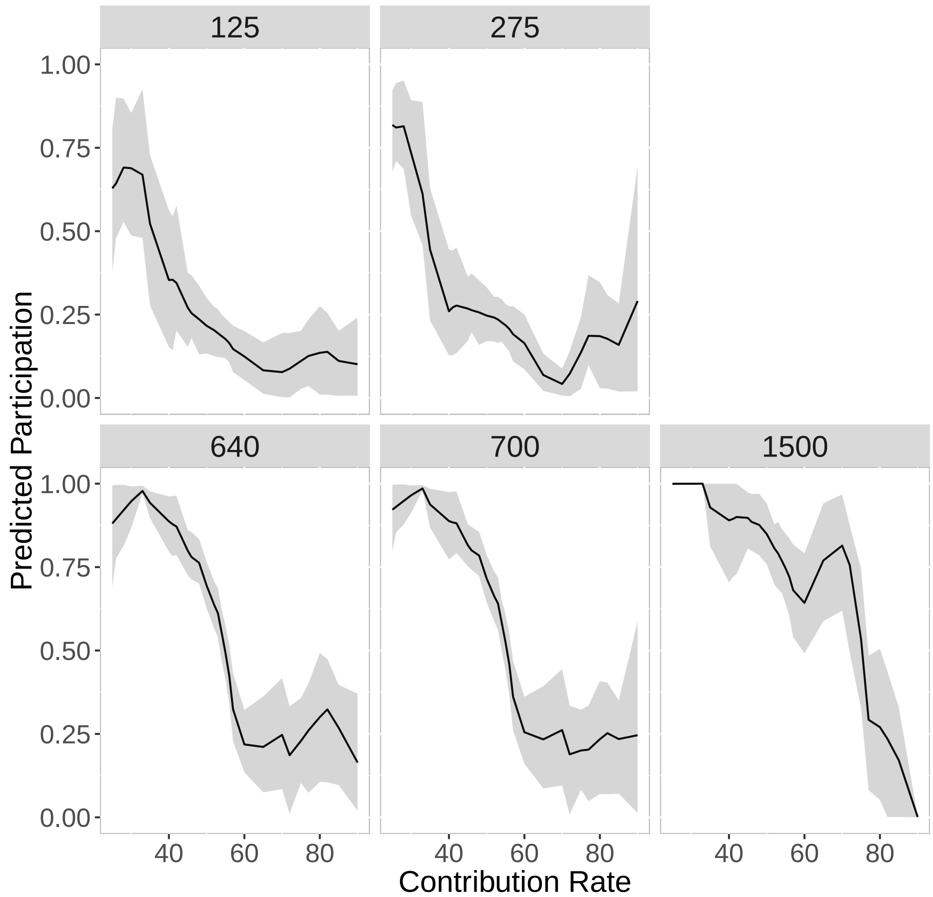
Table 5: Bandwidths generated using least squares cross validation and maximum likelihood cross validation

The bandwidths resulting from each selection method as well as the maximum value these bandwidths can take on given the chosen kernels are presented in Table 5. Bandwidths which are closer to their maximum values are indicative of variables which are irrelevant. The table suggests that *participate* and *CRate* are relevant, while *income* is very close to being smoothed out of the regression. In this case maximum likelihood cross validation is not suspected of oversmoothing, since the results of the least squares and maximum likelihood cross validation routines are very similar. Appendix Section F investigates the relevance of *income* in detail, finding *income* to be highly relevant. In what follows, the least squares cross validated bandwidths of the approach with both *income* and *CRate* will be used.

[8] An irrelevant regressor is a variable whose variations do not contribute to variation in the outcome.

[9] Prohibitive computational intensity means that the routine may take months to determine a result using current computational technology.

Figure 6 illustrates the results of the Nonparametric approach for the predicted probability that *participate* = 'Participate': $\hat{g}(y|x) = \widehat{participate}$. The predictions retain some noise, but unlike both the completely unsmoothed Empirical approach or the oversmoothed Standard approach, these results support the existence of unique switch-points at each *income* level. While the confidence bounds are wider in some regions due to a lack of observations, the changes in $\widehat{participate}$ are much steeper than under the Standard approach. Section 4.4 will demonstrate the superiority of the Nonparametric approach to smoothing in terms of fitting with the observations.



Proportion of 'Participate' outcomes is the solid line. Grey shaded area is the bootstrapped 90 percent confidence interval.

Figure 6: Predicted probability of participation by contribution rate and income using the Nonparametric approach.

4.4 Comparison of approaches

The Nonparametric approach dominates in terms of producing predicted probabilities of participation which match with the collected observations on participation (in-sample fit). This better in-sample fit of the Nonparametric approach is a particularly attractive feature for analysts of experimental data because often the point of analysis is not to approximate an unknown population from which the results are a sample but to explore patterns within a carefully collected population of laboratory observations. As long as the number of explanatory variables are few, the relatively small data sets encountered by experimentalists are suitable for the least squares cross validation procedure which removes irrelevant variables with little to no risk of oversmoothing. In what follows, the use of the Nonparametric approach will serve as the reference case to evaluate the prediction that individuals will participate or will not participate in 'topping-up' consumption given the explanatory variables *income* and *CRate*. Comparisons will be made with the Standard approach as a point of illustration.

In order to compare approaches four criteria are presented here. The Adjusted Correct Classification Ratio (adj-CCR), or adjusted accuracy rate, forms the basic criterion for comparison (Fawcett, 2006). Two further measures include Receiver Operator Characteristics (ROC) curves (See Swets (2014) and Green and Swets (1966)) and Youden's J. As well, Cohen's κ, which adjusts for the probability of selecting the matching outcomes by chance, is provided. These approaches are considered because the traditional pseudo-R^2 values such as the Adjusted McFadden's R^2 reported for the results of the Standard approach are not comparable across models. In what follows, the Nonparametric approach shows the strongest performance regardless of the assessment criterion.

4.4.1 Confusion matrices and correct classification ratio

		Observed Participation	
		Do not Participate	Participate
Predicted Participation	Do not Participate	true negative	false negative
	Participate	false positive	true positive

Table 6: Outline of a confusion matrix

For each smoothing approach, a confusion matrix (or correct classification matrix) summarizes how well the resulting smoothed values match the originally observed outcomes. In Section 3 a classification matrix was introduced to compare the observations with the predictions of the theoretical model. Now, the same technique will be used to compare the observations with the predicted probabilities of participation for each of the Standard and Nonparametric approaches. For these comparisons, a 'threshold' must additionally be specified for sorting the predictions in to the classes 'Participate' and 'Do not Participate'. For example, if the Standard approach estimates that the conditional probability of participation is 70% for a participant with an *income* of 125 and a *CRate* of 55, then given a threshold such as 50% (the typical default threshold) this prediction would be classed as 'Participate' and compared to the actual observation. Whenever this positive decision matches the decision recorded in the experiment the total in the cell 'True Positive' increases. Table 6 provides the naming convention. If the prediction is positive but the actual observation was negative, a false positive is recorded whereas if the prediction was negative but the actual observation was positive a false negative is recorded.

The CCR is a simple way to summarize the entries in the confusion matrix. This measure consists of the fraction of outcomes which match the actual outcomes. As with the confusion matrix this measure is also dependent upon the particular threshold employed when converting the predicted probabilities into binary outcomes. The default threshold used here for demonstration is 0.5. Unfortunately the CCR does not control for the probability of correctly choosing the more frequent outcome. To account for this, the adj-CCR is used. This measure is defined as:

$$\text{adj-CCR} \quad = \quad \frac{\text{True Positive} + \text{True Negative} - \text{M}}{n - \text{M}} \tag{10}$$

where M is the count of the most frequent outcome.

Approach	Threshold	Adjusted CCR	Lower Bound	Upper Bound
Empirical	0.5	0.66	0.55	0.75
Standard	0.5	0.50	0.41	0.62
Standard (Interaction)	0.5	0.52	0.41	0.63
Nonparametric	0.5	0.56	0.45	0.66

Values rounded to the nearest hundredth.
Bounds are bootstrapped 95 percent confidence intervals.

Table 7: Adjusted correct classification ratios for each approach.

The adj-CCRs for the approaches explored thus far are reported in Table 7. The Empirical approach offers the best fit and the Standard the worst. This is no surprise since the Empirical approach is completely responsive to the data at hand, however, this approach does not smooth noise very well, masking switch-type patterns in noise. The Standard approaches with and without interaction terms are very similar, adding interactions improves the fit by only 4%. As discussed earlier, while the Standard approach smooths out noise, it does so at the cost of masking potential switch-type patterns. The Nonparametric approach using least squares cross-validation offers a 12% improvement over the Standard approach, (8% over the Standard approach with interaction included). The Nonparametric approach is 15% worse than the Empirical approach, however, it clearly indicates a step pattern while outperforming the

Standard approaches.[10]

4.4.2 Receiver operator characteristics curves

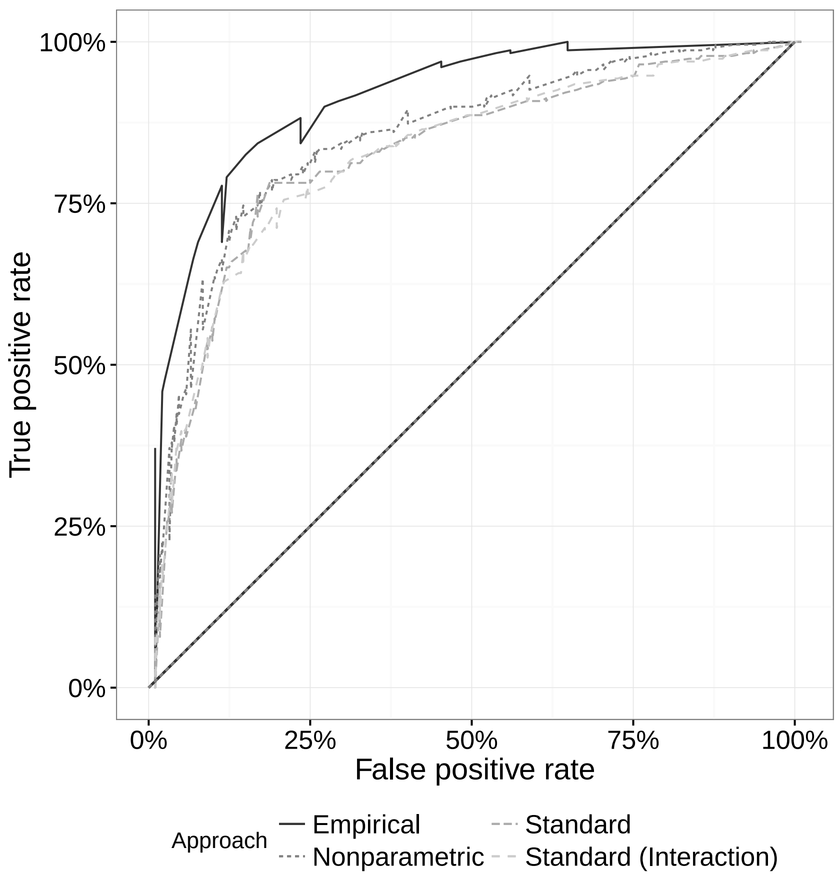


Figure 7: Receiver operator characteristics curves for each approach.

All the results of Table 7 are dependent upon the arbitrary choice of a threshold of 0.5. Receiver Operator Characteristic (ROC) curves, however, explore the effect of varying the threshold value, plotting the True Positive and False Positive Rates as the threshold is varied. These curves, presented in Figure 7, compare the predictive performance of the different estimation ap-

[10]For reference, the Nonparametric approach excluding *income* as a regressor (not shown in Table 7) worsens the fit by 16% over the Nonparametric approach including income, offering further support for including *income* in spite of the large smoothing parameter.

proaches as the threshold changes. A detailed explanation of the Receiver Operator Characteristics Curve is provided in Appendix Section G. More bowed out ROC curves indicate better predictive ability, so the Empirical approach is the best performing estimation approach and the two Standard approaches the poorest.

Area Under the ROC (AUC) is the preferred statistic used to quantify the ability of the results of a smoothing approach to fit an observed data set and for comparing ROC Curves. Perfect predictive ability produces an AUC of 100% while zero predictive ability produces an AUC of 50%. Table 8 presents these results and their bootstrapped 95% confidence intervals.[11] The Nonparametric dominates the smoothed approaches. The results indicate that the Empirical approach explains the data most effectively, while the Nonparametric approach is more efficient than the Standard approach.

Approach	AUC	Lower Bound	Upper Bound
Empirical	91.76	89.41	93.94
Standard	83.41	79.71	87.00
Standard (Interaction)	83.12	79.65	86.84
Nonparametric	86.15	82.68	89.47

Bounds are bootstrapped 95 percent confidence intervals.

Table 8: Area under the receiver operator characteristics curve for each approach.

The AUC corresponding to the Standard approach in Figure 7 is 83.41% and the AUC corresponding to the Standard technique with an interaction term is 83.12%, slightly lower than the original Standard approach. A bootstrap test of the difference between the two areas fails to reject the null of no difference.[12] This suggests that adding the interaction term to the Standard approach did not significantly improve the predictive ability of this approach.

Additionally, testing the difference between the AUCs suggests that the

[11]Using the bootstrapping embedded within the Robin et al. (2011) package.

[12]Using the bootstrapping test provided within the Robin et al. (2011) package.

Approach	Standard (Interaction)	Nonparametric	Empirical
Standard	0.91057	0.27482	0.00
Standard (Interaction)		0.2215	0.00
Nonparametric			0.01

Table 9: P-values of bootstrap tests of differences in areas under receiver operator characteristics curves for each approach.

advantage of the Nonparametric approach over the Standard approach is not significant. The p-value results using the bootstrapping test of the unpaired difference in AUCs provided within the pROC package of Robin et al. (2011) are presented in Table 9. Values less than 0.05 suggest rejection of the null hypothesis of no difference at the 5 percent level of significance. Only the Empirical approach is suggestive of a significant difference. While the improvement in predictive ability of the Nonparametric approach is small, the reduction in misspecification error combined with the substantial difference in capacity to visually suggest a switch-point afforded by this approach are important. This lends support to a descending ranking of preferability of the estimation strategies in terms of AUCs from least to most smoothed.

4.4.3 Youden's J

Another means for exploring the impact of varying the threshold upon the predictive power of the smoothing strategies is to compare Youden's J values (Youden, 1950) at each threshold. This index is described by:

$$J = \text{True Positive Rate} + \text{True Negative Rate} - 1, \tag{11}$$

where the True Positive Rate (TPR) and True Negative Rates (TNR) are defined by

$$TPR = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{12}$$

$$TNR = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \tag{13}$$

J is a measure of relative effectiveness and takes on a value of 0 if the true positive and true negative results are correctly classified at the same rate, and 1 if there are no false positives and no false negatives. Higher values of J are indicative of a more effective approach. The results from a threshold of 0.5 for each smoothing strategy are presented in Table 10. Larger values indicate better match of sorted predictions with the observations. Although the overlap in the confidence bounds of the Youden's J values indicates a lack of difference in the statistical sense, the relative ranking of the values is consistent with the adj-CCR and AUC results reported in Tables 7 and 8, respectively.

Approach	Threshold	J	Lower Bound	Upper Bound
Empirical	0.5	0.69	0.59	0.77
Standard	0.5	0.56	0.45	0.65
Standard (Interaction)	0.5	0.56	0.45	0.66
Nonparametric	0.5	0.59	0.49	0.68

Threshold for sorting predictions is 0.5.
Bounds are bootstrapped 95 percent confidence intervals.

Table 10: Youden's J values for each approach.

This measure again depends on the particular threshold employed. In Appendix Section I a method which searches for a threshold by maximizing Youden's J is investigated.

4.4.4 Cohen's κ

Cohen's κ is a measure of the amount of agreement between the predictions and the observations beyond that occurring by chance. The measure is defined as:

$$
\begin{aligned}
\kappa &= \frac{p_0 - p_e}{1 - p_e} \qquad (14)\\
\text{where } p_0 &= \frac{\text{true positive} + \text{true negative}}{\text{Total}}\\
\text{and } p_e &= \frac{(\text{true negative} + \text{false negative})}{\text{Total}} * \frac{(\text{true negative} + \text{false positive})}{\text{Total}} +\\
&\frac{(\text{false positive} + \text{true positive})}{\text{Total}} * \frac{(\text{false negative} + \text{true positive})}{\text{Total}}
\end{aligned}
$$

Approach	Threshold	kappa	Lower Bound	Upper Bound
Empirical	0.5	0.69	0.62	0.74
Standard	0.5	0.54	0.47	0.61
Standard (Interaction)	0.5	0.56	0.48	0.63
Nonparametric	0.5	0.59	0.52	0.66

Threshold for sorting predictions is 0.5.
Bounds are bootstrapped 95 percent confidence intervals.

Table 11: Cohen's kappa values for each approach.

A kappa value of 1 indicates perfect agreement and 0 no agreement. Table 11 presents the results of sorting the predictions of each approach according to the arbitrary threshold value of 0.5 and calculating Cohen's κ. The results indicate that again, the Empirical approach offers the most agreement between predictions and observations, but among the smoothing options the Nonparametric approach outperforms the parametric approach. The ranking is supported by the results of bootstrapped 95% confidence intervals. The overlap of the confidence intervals of the Empirical and Nonparametric approaches suggests a lack of a statistically significant difference. However, the Empirical approach is significantly different from the Standard approach suggesting that the Nonparametric approach offers a relevant compromise between the over-smoothed Standard and under-smoothed Empirical approaches. In Appendix Section I a method which searches for a threshold by maximizing Cohen's κ is investigated.

5 Identifying switch-points

5.1 Identifying candidates using observations: The cumulative summation method

Within a step-type pattern the 'switch-point' is the x axis coordinate marking the location of the step. With the observations at hand, this could be thought of simply as the point at which the 'Do not Participate' outcome becomes relatively more frequent than the 'Participate' outcome. A way to locate the *CRate* where this occurs is to simply plot the cumulative summation (CS) of the 'Do not Participate' decisions against the inverse of the cumulative summation (ICS) of the 'Participate' decisions on the same graph and locate the *CRate* where the two lines intersect; as presented in Figure 8. This intersection is called the Cumulative Summation Intersection (CSI) and the method used to identify the CSI is the Cumulative Summation Method (CSM). The details are contained in Appendix Section H.

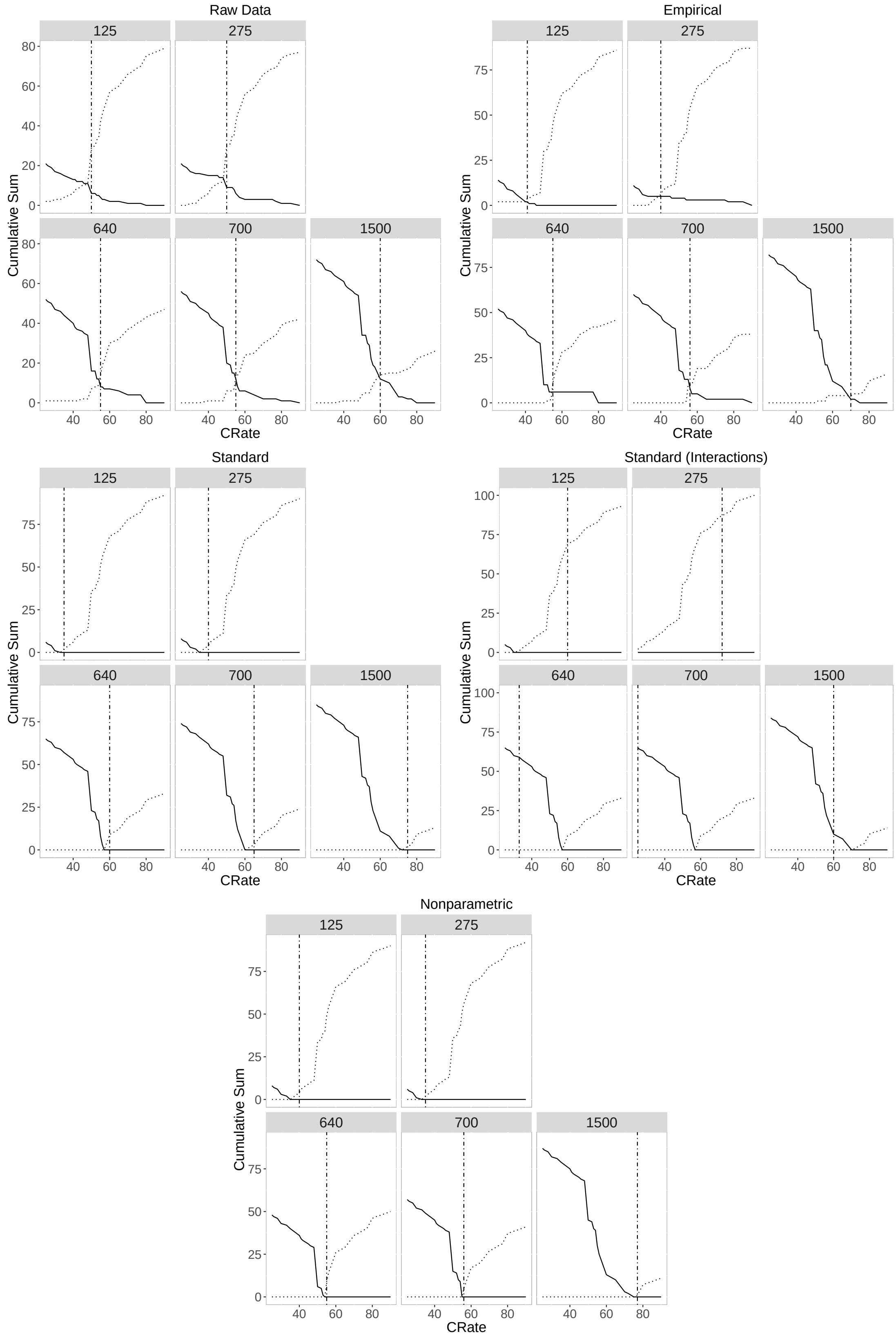
Figure 8 presents the CS and ICS data by CRate using solid and dotted lines respectively. The contribution rate where the distributions cross is taken as the estimated switch-point and is represented by the dashed lines in the figure. While one attractive feature of this method is that it does not require any smoothing, it applies equally well to smoothed predictions as long as these are classified according a threshold (as was done when calculating the CCR). Taking a 0.5 threshold for classification of the predictions, Table 12 presents the candidate switch-points for the Empirical, Standard and Nonparametric approaches, as well as directly to the raw data. The first *CRate* at which the CS exceeds the ICS is the CSI and taken as the switch-point.

Switch-points identified directly with the raw data are convenient because no estimation is required, however this technique cannot differentiate perfect switch-points from equally distributed data with no switch-point. To illustrate this point, the CSI diagrams of two example data sets are presented. The first data set features data which are distributed in such a way as to represent a perfect switch. The second example incorporates data which are distributed equally and therefore have no switch at all. Both illustrative data sets are composed of 0's and 1's associated with a range from 1 to 100 and both with a mean of 0.5. Figures 9 and 10 show the weakness of applying the CSM to the raw data. In Figure 9 the data are presented in the first pane and have a clear switch-point. In the second pane, the CS and ICS are plotted and suggest a switch-point at 0.5, an exact match to the obvious location of the switch. Figure 10 however presents a data set with no switch-point at all in the first pane. The CSI obtained by plotting the CS and ICS suggests a switch point at 0.5. This method thus loses validity as the data become less representative of a switch. One might consider a method for incorporating the strength of the switch by recognizing that the perfect switch occurs at a cumulative sum of 50 in the case with a switch and at 25 in the non-switch case, but this is not explored here.

Income	Data	Empirical	Standard	Standard (Interaction)	Nonparametric
125	50	41	35	33	40
275	50	40	40	25	35
640	55	55	60	60	55
700	55	56	65	60	56
1500	60	70	75	72	77

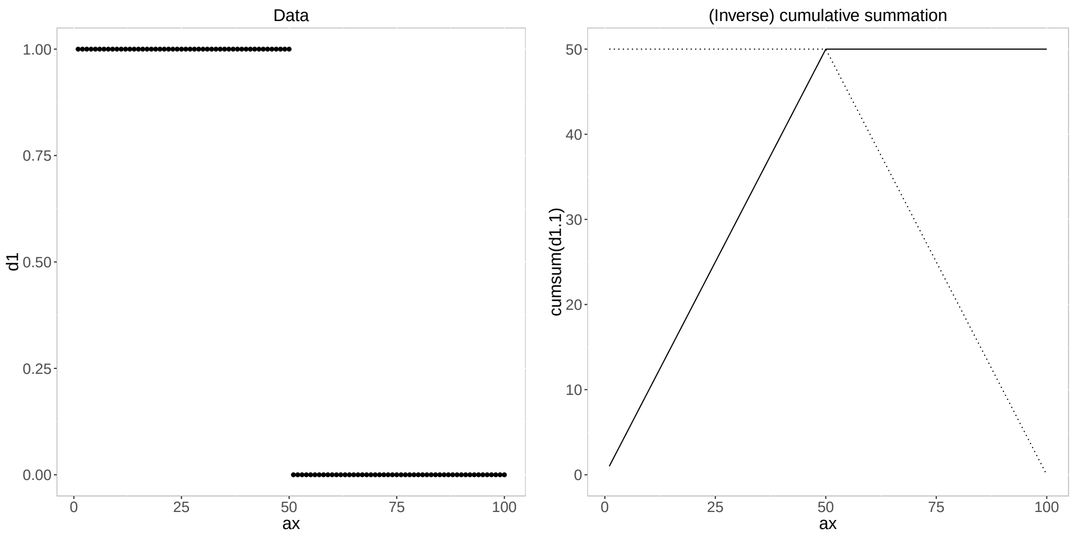
Predictions classified using a threshold of 0.5.

Table 12: Candidate switch-points identified by intersection of cumulative summation and inverse cumulative summation of participate outcomes.



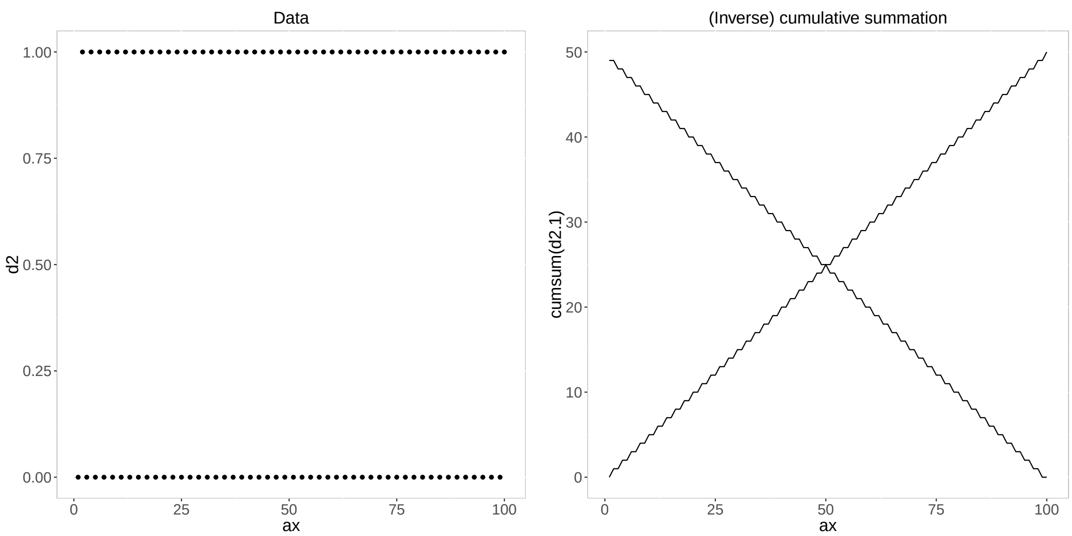
Predictions classified using a threshold of 0.5.

Figure 8: Switch-point identification using the intersection of cumulative summation of ‘Do not Participate’ and inverse cumulative summation of ‘Participate’ outcomes for each approach.



Cumulative summation of 1's and inverse cumulative summation of 0's using example data.

Figure 9: Example of a perfect switch-point.

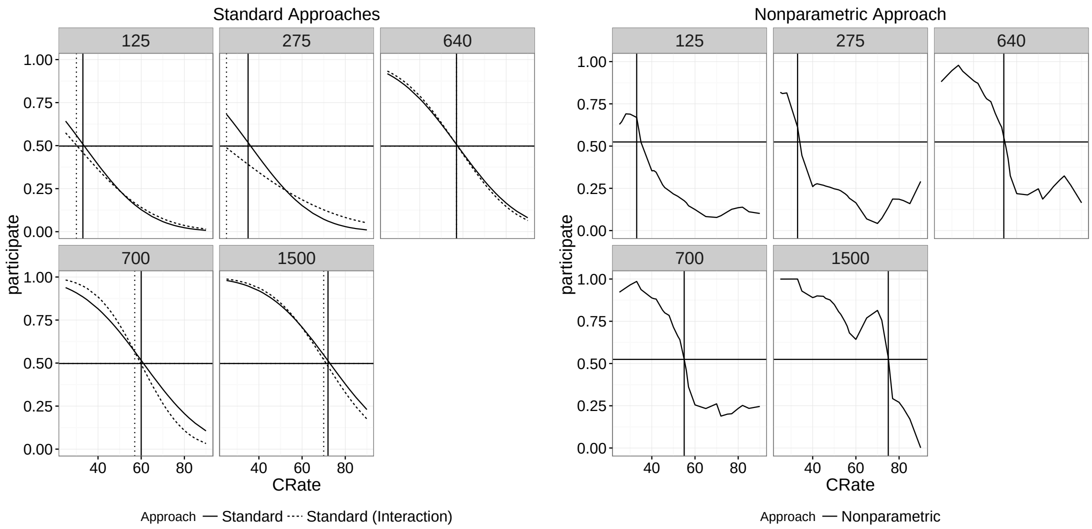


Cumulative summation of 1's and inverse cumulative Summation of 0's using example data.

Figure 10: Example of absence of a switch-point.

5.2 Identifying candidates using predicted values: Youden's optimal J

For each of the approaches to smoothing taken in Section 4.2, candidate switch-points can be identified by mapping the single Youden's Optimal J (YOJ) of the approach onto the *CRate* for each level of *income*. Thus five candidate switch points are identified for each smoothing approach. The results of the Empirical approach are omitted here since these cross the YOJ value in multiple locations, leading to multiple values of *CRate*s as candidate switch-points. Multiple intersections do not occur with the parametric approaches by design, and rarely occur under the Nonparametric approach. In cases where this does occur the median of the candidates identified is used. Figure 11 illustrates the method for



Horizontal lines indicate optimal Youden's J value, vertical lines indicate intersection with the profile of predictions.

Figure 11: Identification of switch-points by mapping optimal Youden's J values to contribution rates.

the Standard approach both with and without the interaction term in the first pane and for the Nonparametric approach in the second pane. The horizontal lines represent the optimal Youden's J values, and the vertical lines indicate the *CRate* where the YOJ intersects the predictions of a particular approach. The intersection point represents a candidate switch-point since all predictions to the left can be sorted as 'Participate' and all to the right as 'Do not Participate'. Following this sorting, the resultant candidate switch-points can be compared to those predicted by the theoretical model. Table 13 summarizes the candidate switch-points identified using this method. If the predictions do not intersect the YOJ value (as in the 275 *income* level for the Standard approach with interaction) the candidate switch-point is recorded as the lowest *CRate* which occurred in the observations.

Income	Standard	Standard (Interaction)	Nonparametric
125	33	30	33
275	35	25	33
640	57	57	54
700	60	57	55
1500	72	70	75

Table 13: Switch-points identified using the optimal Youden's J value.

The results for OYJ are very similar across smoothing strategies. The two Standard approaches return identical results for the 125, 700 and 1500 income levels, and very similar results for the remaining 275 and 640 income levels. The results of the Nonparametric approach identify candidate switch-points which are much more clearly reflected in the plotted predictions than in the Standard approaches. Since this method provides a result regardless of the steepness of the change in the predictions, no information is provided about the merit of the candidate switch-points identified.

5.3 Identifying candidates using predicted values: maximum absolute gradients

Because switch-points are substantively large changes in predicted participation decisions ($\widehat{participate}$) over a very small range of *CRate* at each level of *income* we can use the gradients of each of these approaches to compare the relative merits of each of the candidate switch-points.

A gradient is simply the rate of change in predicted probability at a particular *CRate* and *income*. Gradients can provide information about the relative intensity of the candidate switch-points both across and within approaches. Comparing across approaches, the sizes of the gradients can offer support, or lack thereof, with larger gradients indicative of a stronger candidate. Within an approach the gradients for each *income* level are a measure of the relative strength of the candidate switch-points, which will be discussed in this section. The details of the calculations are provided in Appendix Section J. The results for the Standard approaches are the smooth lines plotted in Figure 12 while the Nonparametric gradients are the jagged lines. For the Standard approach the gradients become less negative as *CRate* increases because the predictions decline at a decreasing rate in a smooth manner. The gradients of the Nonparametric approach reflect the less smooth nature of the predictions with sharp downward points representing steep changes in the predictions. As previously mentioned, steep changes in the predictions are indicative of switch-points. Using the Nonparametric approach rather than the Standard approach, switch points are unmasked and the gradients point directly to the candidates. Table 14 reports the *CRate*s indicated by the largest gradients in absolute value. For the Standard approaches these are simply the first *CRate*s encountered due to the nature of the predictions, while for the Nonparametric approach clear switch-points within the predictions are indicated.

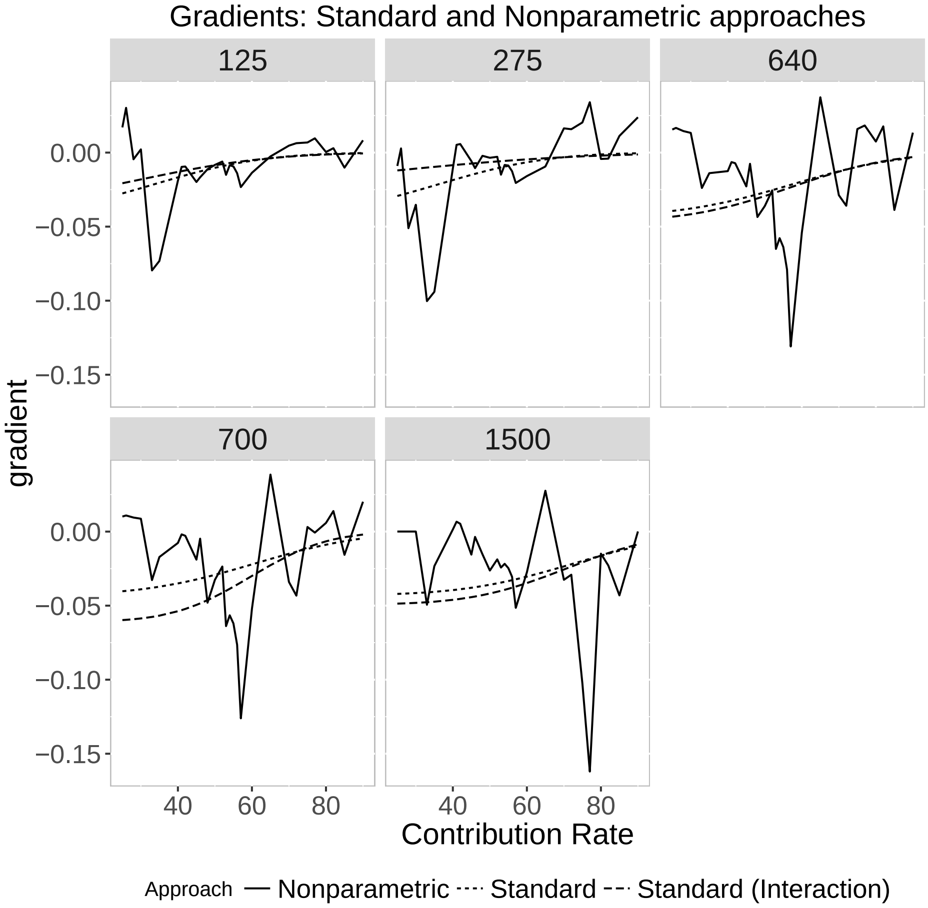


Figure 12: Gradients by contribution rate and income for each approach.

Income	Nonparametric	Standard	Standard (Interaction)
125	33	25	25
275	33	25	25
640	57	25	25
700	57	25	25
1500	77	25	25

Table 14: Switch-points identified using the maximum absolute gradient.

5.4 Comparison of candidate switch-points

Thus far five methods for identifying candidate switch-points have been proposed: Three for the observations alone and two for the predicted probabilities $\widehat{participate}$. The cumulative and inverse cumulative sum intersection (CSM) method is applied to the observations directly (See Appendix Section I for the details of these methods). The Youden's optimal J (YOJ) mapping, and the maximum absolute gradient (MAG) were applied to the predictions of each approach. While the CSM method can be used on the observations without transformation, as well as any of the predictive approaches, it does not differentiate candidates from observations with an obvious discrete change from non-candidates and so is less attractive than the other options.

The YOJ method can be applied only to predictions which are smoothed. This disqualifies the application of the YOJ method to the Empirical approach since the YOJ value intersects the predicted probabilities multiple times leading to multiple *CRate* candidates. For the Standard and Nonparametric approaches the YOJ identifies at least one threshold for optimally sorting predictions into 'Participate' and 'Do not Participate' categories regardless of the degree of smoothing, and so cannot discriminate approaches which exhibit steep changes in the predictions from gentle slopes.[13] The MAG method can be applied also only to adequately smoothed predictions [14] and identifies the point at which the largest change in the predictions occurs. This method is the more direct for locating candidates and evaluating their relative merits because both the location of the change in terms of *CRate* and the value of the the gradient at this point (a measure of the degree of the change in predictions) are

[13] If multiple values are encountered during the bootstrap process the median of the candidate *CRates* is used. This is an issue only for the Nonparametric approach since the Standard approaches are uniformly downward sloping.

[14] Inadequately smoothed predictions may result in non-unique maximum absolute gradients.

provided.[15]

6 Results

6.1 Bootstrapping procedure

The main purpose of the identification and assessment of candidate switch-points is to compare the observations gathered in the experiment with the predictions of the theoretical model of Section 3. Confidence intervals are constructed using the following simple nonparametric bootstrapping approach:

1. An identifier is applied to demarcate independent observations for each group of 5 participants in each period of the experiment. This leads to 100 unique identifier values.
2. A sample, with replacement, of 100 observations of the identifiers is taken.
3. For each identifier, the set of 5 triads of observations of *Income*, *CRate*, *participate* are pulled into the bootstrap sample (i.e. a new set of 500 observations). From this re-sample of 500:
   - Smooth the new set of observations using the approach of choice (Standard or Nonparametric) and record the results.
   - Identify candidates using the CSM, YOJ or MAG methods and record.
4. Repeat steps 2 and 3 1000 times.
5. Record the 5th and 95th percentiles of the observations and candidates as the upper and lower bounds of confidence intervals.

This is the same approach taken for constructing all confidence intervals unless otherwise reported.

[15] A method for constructing a measure of strength of the YOJ candidates using both YOJ and MAG information is provided in Appendix Section K.

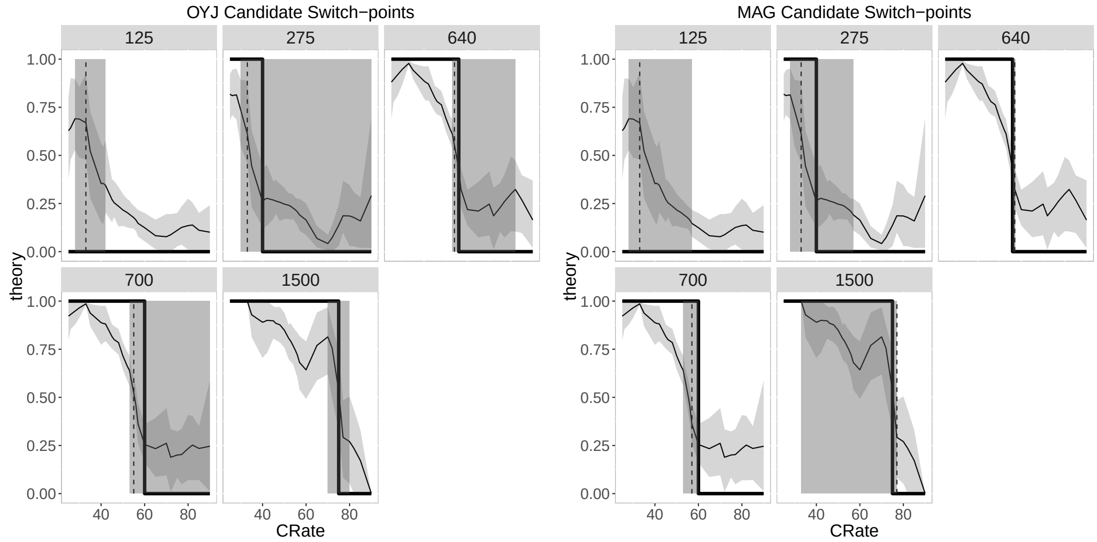
6.2 Observations vs theoretical predictions

Approach	Income	Theoretical	MAG	Lower	Upper	OYJ	Lower	Upper
Standard	125	20	25	25	26	33	25	35
Standard	275	35	25	25	26	35	25	25
Standard	640	56	25	25	26	57	50	57
Standard	700	58	25	25	26	60	53	65
Standard	1500	75	25	25	26	72	60	75
Standard (Interaction)	125	20	25	25	26	30	25	40
Standard (Interaction)	275	35	25	25	26	25	25	35
Standard (Interaction)	640	56	25	25	26	57	50	60
Standard (Interaction)	700	58	25	25	26	57	53	65
Standard (Interaction)	1500	75	25	25	26	70	60	75
Nonparametric	125	20	33	28	57	33	28	42
Nonparametric	275	35	33	28	57	33	30	90
Nonparametric	640	56	57	56	57	54	53	82
Nonparametric	700	58	57	53	60	55	53	90
Nonparametric	1500	75	77	33	77	75	70	80

Maximum absolute gradient (MAG) and optimal Youden's J (OYJ) switch-point identification.
Upper and lower bounds of the bootstrapped 90 percent confidence interval.

Table 15: MAG and OYJ method switch-point candidates and confidence intervals for each approach.

The confidence intervals for the candidate switch-points identified using the MAG and YOJ methods are presented in Table 15. For the Nonparametric approach, all candidates identified using the MAG method, except the 125 *income* level, have associated confidence intervals which include the theoretical prediction. This indicates that the observed switch-point is in agreement with the prediction of the theoretical model. Figure 13 illustrates. In a number of cases agreement with the theoretical predictions is driven by the finding of wider confidence intervals. The 640 *income* level strongly supports the theoretical prediction. Similar results hold for the candidates identified using the YOJ method, though the confidence intervals are generally wider, indicating a lower degree of precision associated with this method of identifying switch-point candidates. This combination of results suggests that the theoretical model cannot be rejected as an explanation of behaviour for all but the lowest income group. For the Standard approaches illustrated in Figures 14 and 15, the results of the MAG and YOJ method differ substantially. The

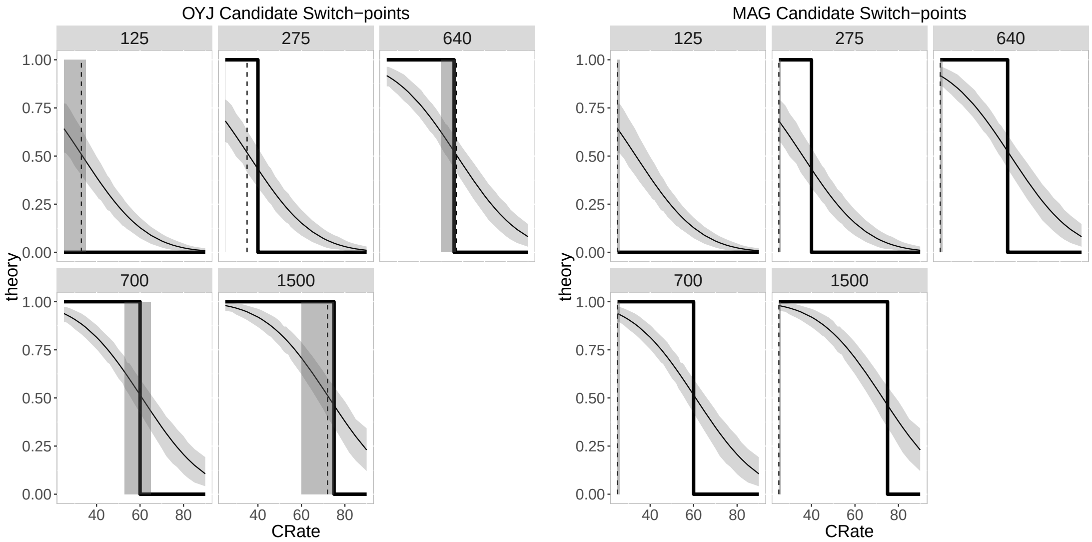


Theoretical cutoffs bold lines, predictions solid lines, and candidate switch-points dotted lines. Bootstrapped 90 percent confidence intervals shaded in grey. Maximum absolute gradient (MAG) and optimal Youden's J (YOJ) switch-point identification methods.

Figure 13: Candidate switch-points of the MAG and OYJ methods for the Nonparametric approach.

MAG confidence intervals suggest a complete rejection of the coincidence of the candidates with the predictions of the theoretical model. All candidates are identified as the first *CRate* encountered and the associated confidence intervals are very small. While the small confidence intervals might be suggestive of a high degree of precision, it is precision associated with an inconsistent candidate, a result of the particular specification of the Standard technique. This is further evidence against the use of the Standard method of smoothing. The YOJ candidates are closer to the theoretical predictions for all except the lowest *income* level, suggesting that the theoretical model partially explains the results. The confidence intervals are wider than those using the MAG method and contain the estimates of the theoretical model in all but the two lowest *income* instances.

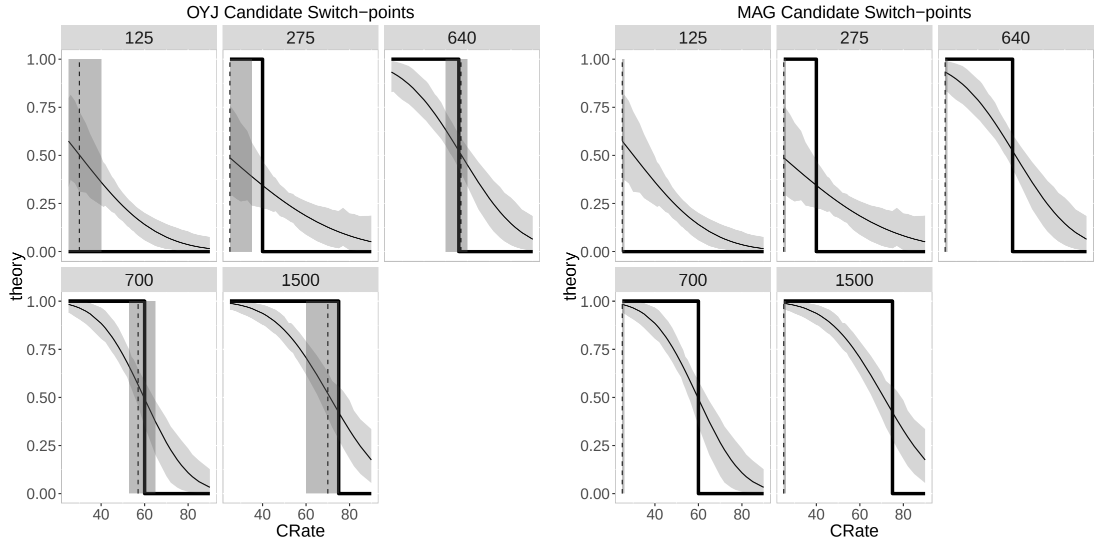
For the Standard approach with interactions the results are similar to the Standard without interactions and raise an important consideration for the treatment of predictions which do not intersect the optimal Youden's J value. In this case the 275 level of *income* predictions do not intersect the Youden's optimal J value and therefore have no associated *CRate*. Rather than leaving this case undefined the *CRate* can be substituted as the minimum *CRate* which occurs in the data set, or as a 0. Here the minimum value in the data set is used.



Theoretical cutoffs bold lines, predictions solid lines, and candidate switch-points dotted lines. Bootstrapped 90 percent confidence intervals shaded in grey. Maximum absolute gradient (MAG) and optimal Youden's J (YOJ) switch-point identification methods.

Figure 14: Candidate switch-points of the MAG and OYJ methods for the Standard approach.

Overall, for the Standard approaches the more accurate MAG method for identifying candidate switch-points rejects the notion that the identified candidates are consistent with the theoretical model, while the less precise YOJ method identifies candidates which are closer to the theoretical predictions. Similar to the CSM approach, under the YOJ approach a switch-point can be suggested even in the case of predictions which exhibit a constant decline rather than a rapid change resembling a switch point.



Theoretical cutoffs bold lines, predictions solid lines, and candidate switch-points dotted lines. Bootstrapped 90 percent confidence intervals shaded in grey. Maximum absolute gradient (MAG) and optimal Youden's J (YOJ) switch-point identification methods.

Figure 15: Candidate switch-points of the MAG and OYJ methods for the Standard approach with interactions.

7 Conclusion and discussion

In this paper observations of an economic experiment were compared to predictions from the theoretical model on which the experiment was based. The theoretical model predicts a discrete change in the binary outcome, forming a 'step' pattern. Three approaches to smoothing observations were investigated, the Nonparametric approach was shown to be the preferred approach by a number of of criteria and, within the predicted $\widehat{participate}$, the MAG method the most effective for identifying candidate switch-points. While the candidate switch-points of each method are similar across approaches, a visual inspection of the gradients demonstrates that the Nonparametric candidates are better at capturing substantive changes in predictions than the Standard candidates.[16] Finally, Nonparametric bootstrapped confidence intervals were

[16]The statistic E was proposed in Appendix Section K as a means of condensing the YOJ results into a measure of the intensity of each candidate switch-point for comparison across and within models and potentially across data sets.

constructed around the candidate switch-points and the match between the identified switch points and cutoffs suggested by the theoretical model evaluated.

While the Standard approach with interactions suggested rejection of the coincidence of the candidate switch-points with the theoretical cutoffs for the two lowest levels of *income* these results are likely invalid. Based on comparison to the confidence intervals of the Nonparametric approach, the Standard approach appears to offer a high level of precision around incorrect estimates. Combined with results which suggest that the parametric models are misspecified and weakly indicative of switch-points, inference should not be made about the behaviours captured in the experimental data based on the Standard approaches.

The Nonparametric approach and MAG method extend naturally to other fields in which analysts wish to confront observed data with an externally determined cutoff for the purpose of evaluating the effectiveness of such a cutoff. The potential policy applications of this framework are diverse. For example, medical practice guideline evaluation could be improved by the application of the more robust Nonparametric approach combined with the MAG method for identifying switch points in observed practitioner behavior. Firm entry and exit decisions over a range of prices can be analysed using this method, as can the uptake of public programming over a range of incomes of individuals. The flexibility of the smoothing also allows for evaluation of more complex guidelines. In addition, by smoothing the observations themselves, rather than the observations which are in agreement with the theoretical predictions, the smoothing approach, switch-point identification method and the guideline can be easily compared against alternative specifications visually. This intuitive visual appeal of the results provides the added benefit of fostering the communication of the results of evaluations undertaken using this method with diverse audiences.

References

Aitchison, J. and Aitken, C. G. G. (1976). "Multivariate binary discrimination by the kernel method". In: *Biometrika* 63.3, pp. 413–420.

Akaike, H. (1974). "A new look at the statistical model identification". In: *Automatic Control, IEEE Transactions on* 19.6, pp. 716–723.

Buckley, N., Cuff, K., Hurley, J., Mestelman, S., Thomas, S., and Cameron, D. (2015). "Support for public provision of a private good with top-up and opt-out: A controlled laboratory experiment". In: *Journal of Economic Behavior & Organization* 111.0, pp. 177 –196.

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46.

Cragg, J. G. and Uhler, R. S. (1970). "The demand for automobiles". In: *The Canadian Journal of Economics/Revue canadienne d'Economique* 3.3, pp. 386–406.

Epanechnikov, V. A. (1969). "Non-Parametric Estimation of a Multivariate Probability Density". In: *Theory of Probability & Its Applications* 14.1, pp. 153–158. DOI: `10.1137/1114019`. eprint: `http://dx.doi.org/10.1137/1114019`. URL: `http://dx.doi.org/10.1137/1114019`.

Fawcett, T. (2006). "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8, pp. 861–874.

Fernihough, A. (2014). *mfx: Marginal Effects, Odds Ratios and Incidence Rate Ratios for GLMs*. R package version 1.1. URL: `http://CRAN.R-project.org/package=mfx`.

Green, D. and Swets, J. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hall, P., Racine, J. S., and Li, Q. (2004). "Cross-validation and the estimation of conditional probability densities". In: *Journal of the American Statistical Association* 99, pp. 1015–1026.

Hayfield, T. and Racine, J. S. (2008). "Nonparametric Econometrics: The np Package". In: *Journal of Statistical Software* 27.5.

Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton University Press.

McFadden, D. (1973). *Conditional logit analysis of qualitative choice behavior*. Institute of Urban and Regional Development, University of California.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Racine, J. S. and Li, Q. (2004). "Nonparametric estimation of regression functions with both categorical and continuous data". In: *Journal of Econometrics* 119.1, pp. 99–130.

Ramalho, E. A. and Ramalho, J. J. S. (2012). "Alternative Versions of the RESET Test for Binary Response Index Models: A Comparative Study*". In: *Oxford Bulletin of Economics and Statistics* 74.1, pp. 107–130. ISSN: 1468-0084.

Ramsey, J. B. (1969). "Tests for specification errors in classical linear least-squares regression analysis". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 350–371.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves". In: *BMC Bioinformatics* 12, p. 77.

Stone, C. J. (1977). "Consistent Nonparametric Regression". In: *The Annals of Statistics* 5.4, pp. 595–620.

Swets, J. A. (2014). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. New York: Psychology Press.

Youden, W. J. (1950). "Index for rating diagnostic tests". In: *Cancer* 3.1, pp. 32–35. ISSN: 1097-0142.

Appendices

A Frequency of observations

	25	26	28	30	33	35	40	41	42	45	46	48	50	52	53	54	55	56	57	60	65	70	72	75	77	80	82	85	90
125	2	1	1	3	1	2	4	2	1	2	1	1	23	1	4	1	9	5	3	9	3	7	1	2	1	6	1	1	2
275	2	1	1	3	1	2	4	2	1	2	1	1	23	1	4	1	9	5	3	9	3	7	1	2	1	6	1	1	2
640	2	1	1	3	1	2	4	2	1	2	1	1	23	1	4	1	9	5	3	9	3	7	1	2	1	6	1	1	2
700	2	1	1	3	1	2	4	2	1	2	1	1	23	1	4	1	9	5	3	9	3	7	1	2	1	6	1	1	2
1500	2	1	1	3	1	2	4	2	1	2	1	1	23	1	4	1	9	5	3	9	3	7	1	2	1	6	1	1	2

Table A.1: Number of observations by income and contribution rate

B Standard approach

The index function (the $X'\beta$) is described by:

$$X'\beta \;=\; \alpha + \beta_1 \, CRate \; + \beta_2 \mathbb{I}_{275} + \beta_3 \mathbb{I}_{640} + \beta_4 \mathbb{I}_{700} + \beta_5 \mathbb{I}_{1500}, \qquad \text{(B.1)}$$

where *CRate* and *income* are explanatory variables, $\beta_2, \beta_3, \beta_4$, and β_5 are associated with indicators, $\mathbb{I} = 1$, for the sub-scripted income level and $\mathbb{I} = 0$ otherwise.

In order to interpret the coefficients in Table 3 as probabilities the index function must be distributed according to a cumulative normal distribution for the probit model. For example, the probability of a subject with an income of

125 choosing to participate can be described by:

$$\begin{aligned} \Pr(participate = 1|income = 125, CRate) &= \Phi(\alpha + \beta_1 * CRate) \quad \text{(B.2)} \\ &= \Phi(1.438 - 0.043 * CRate). \end{aligned}$$

The model is estimated by maximum likelihood using iteratively weighted least squares.

C Standard approach with interaction

The probit model is updated to include interaction terms:

$$\begin{aligned} X'\beta &= \alpha + \beta_1 CRate + \beta_2 \mathbb{I}_{275} + \beta_3 \mathbb{I}_{640} + \beta_4 \mathbb{I}_{700} + \beta_5 \mathbb{I}_{1500} + \quad \text{(C.1)} \\ & \beta_6 \mathbb{I}_{275} * CRate + \beta_7 \mathbb{I}_{640} * CRate + \beta_8 \mathbb{I}_{700} * CRate + \beta_9 \mathbb{I}_{1500} * CRate, \end{aligned}$$

which is again, distributed according to $\Phi(\cdot)$ in order to interpret the results as probabilities. The $\mathbb{I}$ represent indicator variables equal to 1 at the sub-scripted levels of *income* and 0 otherwise, as before.

So now, for example, the probability of a subject with an *income* level of 275 choosing to participate can be described by:

$$\begin{aligned} \Pr(Y = 1|income = 275, CRate) &= \Phi(\alpha + \beta_1 * CRate + \beta_2 \mathbb{I}_{275} + \beta_6 \mathbb{I}_{275} * CRate \\ &= \Phi(1.088 - 0.036 * CRate - 0.502 + \\ & 0.011 * CRate). \quad \text{(C.2)} \end{aligned}$$

D Coefficients of determination

Often a coefficient of determination is used as a measure of goodness of fit, with various pseudo-R^2 calculations for predictions not calculated by ordinary least squares, as is the case here.

For example, the pseudo-R^2 described by McFadden (1973) is:

$$R^2 = 1 - \frac{ln\hat{L}(M_{FULL}) - K}{ln\hat{L}(M_{INTERCEPT})}, \tag{D.1}$$

where $\hat{L}$ is the estimated likelihood, K the number of parameters, M_{FULL} is the full model and $M_{INTERCEPT}$ is the model with only an intercept included. The value of this statistic is 0.24, which is indicative of a good model fit according to Louviere (2000).[17] By this metric alone it is plausible that this model sufficiently describes the data. Another option is the R^2 of Cragg and Uhler (1970), which is an adjusted version of the R^2 value of Cox and Snell (1989), and is described by :

$$R^2 = \frac{1 - \{\frac{L(M_{INTERCEPT})}{L(M_{FULL})}\}^{\frac{2}{N}}}{1 - L(M_{INTERCEPT})^{\frac{2}{N}}}, \tag{D.2}$$

where L is the log likelihood, M_{FULL} and $M_{INTERCEPT}$ are the same as previously defined and N is the number of observations in the data set. This statistic takes on values between 0 and 1, with 0 indicating a poor fit and 1 a very good fit. The value here is 0.4. Another R^2 which is independent of the particular approach used is the adjusted count R^2 described by:

$$R^2 = \frac{Correct - n}{Total - n}, \tag{D.3}$$

where *Correct* are the number of predicted outcomes ≥ 0.5, *Total* is the total

[17] Louviere (2000) suggests that a value between 0.2-0.4 indicates a very good model fit.

number of observations, and n is the count of the most frequent outcome. Here it has a value of 0.5. Each of these measures has a slightly different interpretation so they are not directly comparable, however it is not uncommon to see values in these ranges cited as indicative of good model fits.

E Bandwidth selection

The benefits of least squares cross validation are described in detail by Hall et al. (2004). This method minimizes the weighted integrated squared error described by:

$$ISE = \int \hat{g}(y|x) - g(y|x)^2 \mu(x) M(x^c) dxdy, \tag{E.1}$$

where $M(x^c)$ is a weight function which can then be minimized by least squares cross validation, described by:[18]

$$CV_{g_0} = \int \frac{[\hat{(}f)(x,y) - \hat{(}x)g(y|x)]^2 M(x^c)}{\mu(x)^2} dxdy$$

and

$$CV_{g_0}(h_0, h, \lambda) = n^{\frac{-q}{(q+4)}} \chi_g(a_0, a, b) \tag{E.2}$$

This approach has the distinctive benefit that irrelevant elements of X can be smoothed completely out of the regression.

An alternative to least squares cross validation is to use maximum likelihood cross validation, which can tend to oversmooth for fat-tailed distributions:

$$\hat{g}_{-i}(Y_i|X_i) = \frac{\hat{f}_{-i}(X_i, Y_i)}{\hat{m}_{-i}(X_i)}. \tag{E.3}$$

[18]See Li and Racine (2007) chapter 5, pp 157-160 for more details about this function

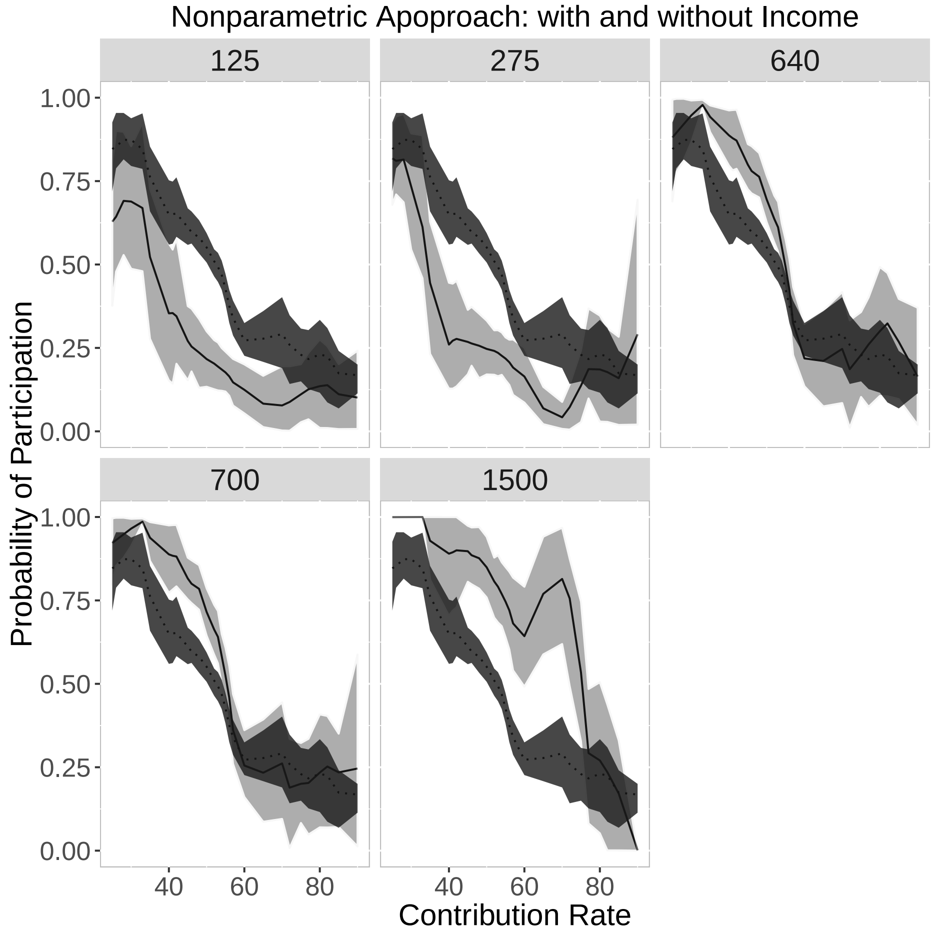
F Is *income* relevant?

Variable	Least Squares	Maximum Likelihood	Least Squares Without Income	Upper Bound
Participate	0.0000	0.0512	3e-04	0.5
Income	0.9926	0.9923	-	1
CRate	3.1305	3.1305	2.8407	inf

Table F.1: Bandwidths generated using least squares cross validation (with and without income) and maximum likelihood cross validation.

It is worth noting that even when using the maximum likelihood strategy *income* is not completely smoothed out, as would be indicated if the bandwidth were equal to 1.00. As noted in Racine and Li (2004) there may be times in which a smoothing parameter may be close to it's upper bound but not be irrelevant. As a check for the relevance of *income* the least squares cross validation bandwidth selection routine is run for the conditional probability of *participate* on *CRate* alone. The column 'Least Squares Without Income' in Table F.1 shows the resulting smoothing parameters when *income* is omitted from the estimation. The bandwidth for *participate* is nearly identical to the least squares bandwidth when *income* is included, while the bandwidth of *CRate* is lower, indicating that more of the variation in *CRate* is now explaining variation in the conditional probability of participation.

Comparing the Adjusted Count R^2_{AC} of each estimation described by Equation D.3 assists in clarifying the relevance of *income* in estimating the conditional probability of *participate* because the value is 0.559 in the case where *income* is included and 0.284 when it is not. These results support the continued inclusion of *income* despite having a smoothing parameter near its upper bound. If *income* were truly irrelevant these R^2 values should be nearly identical. In what follows, the bandwidths resulting from the least squares cross validation procedure with probability of participation conditional on both *income* and *CRate* will be used.



Solid lines and lighter gray confidence bands include income as an explanatory variable. Dotted lines and darker gray confidence bands exclude income as an explanatory variable. Confidence intervals are bootstrapped 90 percent bounds.

Figure F.1: Predicted probability of participation by contribution rate (with and without income) using the Nonparametric approach.

Figure F.1 plots the results for the Nonparametric approach both with and without *income* as an explanatory variable, where without *income* the predictions remain constant across the five panes of the figure. It is clear that the predictions of the model without *income* differ substantially from the model with *income*, supporting the inclusion of *income* since if income were truly irrelevant the results should be identical.

G Receiver operator characteristics curve

The True Positive Rate (TPR) is the percentage of predicted positive decisions which match the observed positive decisions. This is expressed as:

$$TPR = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}, \tag{G.1}$$

and is a measure of the ability of the theoretical model to correctly predict a positive response. Taking an example from the medical literature where such methods are frequently used, this measures the ability of a diagnostic test to correctly identify a disease in a person who indeed has the disease.

Increasing the ability of a model to correctly predict positive outcomes is typically done at the expense of increasing the False Positive rate: of incorrectly predicting illness in a healthy patient. This measure is the False Positive Rate (FPR) and is expressed:

$$FPR = \frac{\text{false positives}}{\text{false positives} + \text{true negatives}} \tag{G.2}$$

Combining the TPR and FPR, ROC curves capture the responsiveness of a particular estimation strategy by varying the threshold from 0 to 1, recalculating the confusion matrix and plotting the TPR against the FPR. Figure 7 in Section 4.4.2 above plots the results for the models discussed in Section 4.2. Since the measures constructed from a confusion matrix do not depend on underlying assumptions about the model from which the predictions were generated, the results are comparable regardless of the source of the predictions. When the threshold is 0 all predictions are classified as 0's and thus no positives (1's) are identified at all. When the FPR is 0 it must also be the case that that the TPR is zero since there are no positives identified at all. This point occurs at the origin of Figure 7. When the threshold is 1 all predictions are classified as 1's,

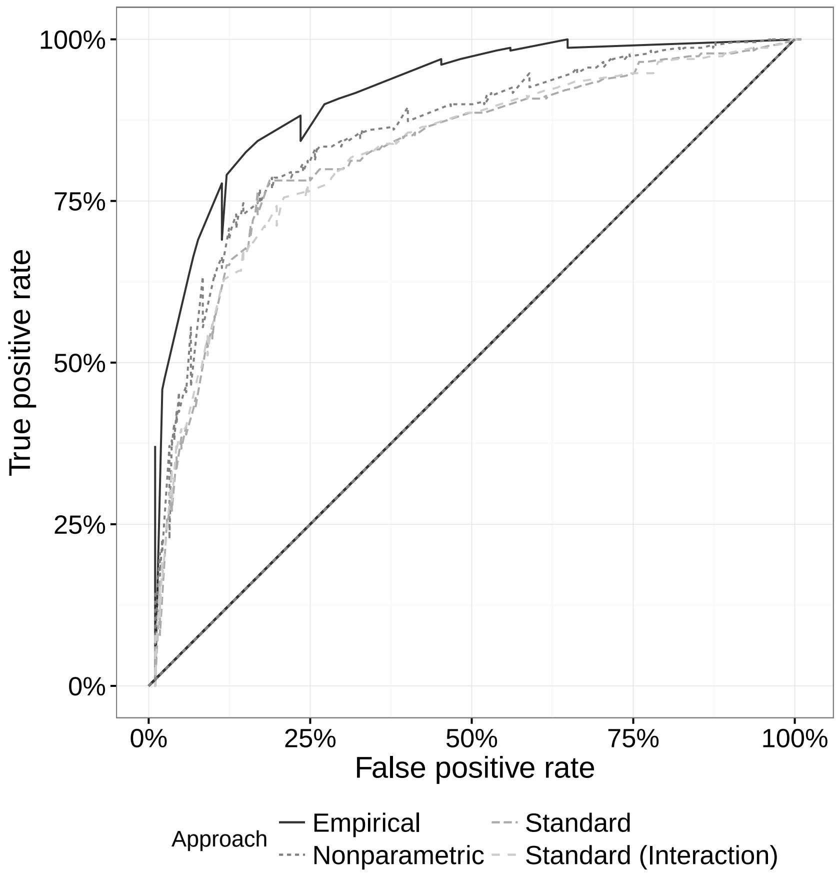


Figure 7: Receiver operator characteristics curves for each approach. (repeated from page 32)

no negatives are identified, and the TPR and FPR are thus both 100%. This point occurs in the North East corner of Figure 7.

More concave ROC curves are indicative of predictions which are closer to the observed data. A curve going from (0,0) to (0,100), to (100,100) represents a situation in which the predictions perfectly match the outcomes regardless of the threshold used. The 45° line on the other hand is representative of a situation in which varying the threshold causes an exactly proportional increase in the FPR and TPR indicating that the results of the predictive strategy do not describe the data at hand.

H Cumulative summation intersection

Cumulative summation is the sum of the frequencies of an outcome up to a particular *CRate* and *income*. Here it is defined as the sum of the 0 observations up to a particular *CRate*(j) for each *income*(k):

$$CS_{jk} = \sum_{1}^{j} \sum_{i}^{N} I\{(y_i|CRate = j, income = k) = 0\} \tag{H.1}$$

As more observations are encountered the cumulative summation will either increase or stay the same, depending on the outcome of interest. The inverse of the cumulative summation is the cumulative frequency of an outcome up to a particular *CRate* and *income* subtracted from the maximum value of the cumulative summation.

$$\begin{aligned} ICS_{jk} = &\sum_{1}^{J} \sum_{i}^{N} I\{(y_i|CRate = j, income = k) = 1\} - \\ &\sum_{1}^{j} \sum_{i}^{N} I\{(y_i|CRate = j, income = k) = 1\} \end{aligned} \tag{H.2}$$

I Identifying candidates using observations: search method

Another way the raw observations can be used to identify candidate switch-points is via searching for a candidate which maximizes a criteria. This method is convenient because it does not require smoothing, but it may remain sensitive to noise.

To carry out a search for an optimal candidate,for each level of income:

1. Choose a switch-point candidate,
2. Create a ‘pseudo-step’: define all *CRate*s equal or less than the candidate

as 'Participate' and all *CRate*s greater as 'Do not Participate'.

3. Construct a confusion matrix of the observations and pseudo-step.

4. Calculate the J value.

5. Repeat the procedure for each possible candidate (every possible *CRate*).

The *CRate* associated with the maximum J value is then the 'optimal J' (oj) candidate. The results are presented in Table I.1. The largest value of J is obtained at the 640 and 700 levels of *income* for which the thresholds are 53.

Income	Candidate	J
125	46	0.35
275	33	0.29
640	53	0.60
700	53	0.60
1500	55	0.43

Table I.1: Candidate switch-points by search approach and optimal Youden's J method.

Similarly, maximizing the correct proportion instead of the value of J is another interpolation option. This method leads to multiple maxima in some instances. Table I.2. In cases with multiple maxima the median value is reported as the candidate switch-point. The 640 level of *income* exhibits the highest proportion of correctly classified observations at the optimal threshold of 53.

Income	Candidate	Min	Max	Correct Proportion
125	35	33	40	0.10
275	33	33	33	0.26
640	53	53	53	0.57
700	55	53	57	0.50
1500	70	70	70	0.31

Table I.2: Candidate switch-points by search approach and maximum correct proportion method.

Finally, the same procedure is replicated in Table I.3, maximizing the value of Cohen's κ. Here the 640 level of *income* candidate is the strongest, exhibiting 60% agreement beyond chance at the optimal threshold of 53.

Income	Candidate	Cohen's kappa
125	46	0.36
275	33	0.38
640	53	0.60
700	53	0.58
1500	70	0.45

Table I.3: Candidate switch-points by search approach and maximum Cohen's kappa method

Overall there is agreement between methods for locating an optimal threshold on income levels of 275 and 640. The search methods are convenient because they require no smoothing but do not lend well to intuitive graphical presentation of the results.

J Gradients

For the Standard model the gradient is calculated by taking the first derivative of the function $\Phi(X'\beta)$ and is defined as:

$$\frac{\partial Pr(Y=1|X)}{\partial CRate} = \phi(X'\beta) * \frac{\partial(X'\beta)}{\partial CRate}, \quad \text{(J.1)}$$

where $\Phi' = \phi$ is the derivative of the cumulative normal distribution. This leads to an equation into which the coefficients from 3 can be substituted and the exact predictions calculated for each value of *CRate* and *income* level encountered. For the 275 *income* level in the probit specification without interaction terms this amounts to:

$$\frac{\partial Pr(Y=1|X)}{\partial CRate} = \phi(\alpha + \beta_1 * CRate + \beta_2 * D1_{income=275}) * \beta_2 \quad \text{(J.2)}$$

$$= \phi(1.438 + -0.043 * CRate + 0.109) * 0.109 \quad \text{(J.3)}$$

For the Nonparametric approach the gradient is simply computed as the derivative at every point in the predictions:

$$\frac{\partial \hat{g}(y|x)}{\partial x} \tag{J.4}$$

Approach	Income	MAG	Lower Bound	Upper Bound
Standard	125	-0.0218	-0.0430	-0.0186
Standard	275	-0.0237	-0.0429	-0.0203
Standard	640	-0.0218	-0.0539	-0.0301
Standard	700	-0.0243	-0.0542	-0.0310
Standard	1500	-0.0184	-0.0558	-0.0331
Standard (Interaction)	125	-0.0166	-0.0565	-0.0055
Standard (Interaction)	275	-0.0101	-0.0409	-0.0017
Standard (Interaction)	640	-0.0234	-0.0736	-0.0253
Standard (Interaction)	700	-0.0344	-0.1030	-0.0399
Standard (Interaction)	1500	-0.0189	-0.0721	-0.0316
Nonparametric	125	-0.0796	-0.1921	-0.0381
Nonparametric	275	-0.1003	-0.2180	-0.0410
Nonparametric	640	-0.1309	-0.1994	-0.0858
Nonparametric	700	-0.1261	-0.1924	-0.0834
Nonparametric	1500	-0.1620	-0.3307	-0.0675

Bounds are bootstrapped 90 percent confidence intervals.

Table J.1: Gradients at the MAG candidate switch-points.

Approach	Income	OYJ	Lower Bound	Upper Bound
Standard	125	-0.0218	-0.0323	-0.0181
Standard	275	-0.0237	-0.0429	-0.0203
Standard	640	-0.0240	-0.0319	-0.0189
Standard	700	-0.0258	-0.0321	-0.0192
Standard	1500	-0.0198	-0.0331	-0.0190
Standard (Interaction)	125	-0.0166	-0.0404	-0.0055
Standard (Interaction)	275	-0.0101	-0.0327	-0.0017
Standard (Interaction)	640	-0.0260	-0.0422	-0.0155
Standard (Interaction)	700	-0.0372	-0.0568	-0.0224
Standard (Interaction)	1500	-0.0208	-0.0426	-0.0178
Nonparametric	125	-0.0796	-0.1568	0.0213
Nonparametric	275	-0.1003	-0.1788	0.0457
Nonparametric	640	-0.0578	-0.0977	0.0226
Nonparametric	700	-0.0620	-0.1103	0.0295
Nonparametric	1500	-0.1027	-0.1538	0.0165

Bounds are bootstrapped 90 percent confidence intervals.

Table J.2: Gradients at the OYJ candidate switch-points.

K A strength measure for OYJ candidates

To construct a measure of the strength of a candidate switch-point which captures both the proportion of the largest change and the relative distance of the candidate to the maximum change in terms of *CRate* the following are identified:

$$Gradient_M : \text{The true value of the largest gradient (when ranked by absolute value)} \tag{K.1}$$

$$Gradient_Y : \text{The true value of the gradient at the candidate switch-point} \tag{K.2}$$

$$CRate_M : \text{The } CRate \text{ associated with the largest gradient} \tag{K.3}$$

$$CRate_Y : \text{The } CRate \text{ of the candidate switch-point} \tag{K.4}$$

The largest gradient is identified by ranking the gradients in terms of absolute value and selecting the maximum. The actual value of this gradient is recorded because in the Nonparametric approach a difference in sign from the candidate switch-point can be indicative of noise. This is in contrast to the parametric model which imposes unidirectionality upon the predictions. If the candidate switch-point and the maximum gradient are the same or very close the data exhibit a strong switch-point since the optimal cutoff for sorting (the optimal Youden's J mapped into a *CRate*) accurately captures the location where the most substantive change occurs in the predictions. If the optimal Youden's J *CRate* and the *CRate* where the maximum gradient occurs are far apart then most of the changes in the predictions are occurring apart from the candidate, which should weaken the candidate's attractiveness.

One way to concisely compare the maximum gradients and candidate switch-points is to take the percentage of the maximum gradient ($Gradient_M$) captured by the candidate ($Gradient_Y$) in terms of the relative distance of the associated *CRates*. If the $Gradient_M$ occurs close to the candidate then an area within the predictions exhibiting substantive change has been identified, if it occurs far from the candidate then either noise or an inappropriate model is indicated. The percentage deviation from the *CRate* at the maximum gradient from the optimal Youden's J *CRate* provides a unit-free measure of the spread between these two points. *CRate*s located far apart may indicate noise or a model which fails to provide a switch-point. A simple measure of the relationship is suggested by the following:

$$\text{GradientP} = \frac{\text{Gradient}_Y}{\text{Gradient}_M} \tag{K.5}$$

$$\text{CRateP} = \left|\frac{CRate_Y - CRate_M}{CRate_Y}\right| \tag{K.6}$$

$$E = \frac{\text{Gradient}_P}{CRate_P}, \tag{K.7}$$

where the absolute value of $CRateP$ is used because the direction of the the deviation is irrelevant. $\text{Gradient}_P \leq 1$ and may be negative if the maximum (in absolute terms) gradient differs in direction from the gradient identified at the candidate switch-point for a particular *income* level. The resulting measure E is akin to simply calculating the slope between the maximum gradient point and the one identified by the maximum Youden's J value, but avoids issues relating to the scale of *CRate* and thus enables comparison across models and, potentially across data sets. Larger values of E indicate steeper slopes and thus stronger candidate switch-points. A perfect match of the candidate switch-point and maximum gradient provides the largest value E can take on ($\frac{100}{0} = \infty$).

Income	Standard	Standard (Interaction)	Nonparametric
125	4.12	6.00	Inf
275	3.50	Inf	Inf
640	1.96	1.98	7.94
700	1.82	1.93	13.53
1500	1.65	1.71	23.76

Table K.1: Strength measure of candidate switch-points identified using the optimal Youden's J method.

Table K.1 presents the results from each approach. The final values of E range from a low of 1.65 to a high of ∞. In absolute terms, a value less than $|1|$ indicates particularly poor evidence of a switch-point since the relative distance between the YOJ and MAG candidate is greater than the percentage of the MAG gradient captured by the OYJ gradient. Values above $|1|$ are indicative of steeper relationships, as seen for the 125 and 275 *income* levels in Table K.1, but these results offer little information given that the location of the maximum gradients are uniformly the first *CRate* encountered, as discussed in the previous section.

The Nonparametric OYJ candidates are more reflective of steep changes than Standard OYJ candidates, as evidenced by the larger values of E. Nonparametric E values range from 7.94 to ∞, with the minimum value exceeding the maximum value of the Standard approaches. For the 125 and 275 *income* levels the candidate identified by the Nonparametric YOJ is identical to the

Nonparametric MAG candidate and the E of ∞ is ideal.