Exploiting Zero-Inflated Consumption Data using Propensity Score Matching and the Infrequency of Purchase Model, with Application to Climate Change Policy

Bardsley, Nicholas and Buechs, Milena

2013

Exploiting Zero-Inflated Consumption Data using Propensity Score Matching and the Infrequency of Purchase Model, with Application to Climate Change Policy

Nicholas Bardsley*
Milena Büchs†

*University of Reading, School of Agriculture Policy and Development, Department of Food Economics and Marketing, Reading RG6 6AR. n.o.bardsley@reading.ac.uk

†University of Southampton, Social Sciences, Southampton SO17 1BJ, United Kingdom

DRAFT – Please do not cite without permission

This version July 2013

Exploiting Zero-Inflated Consumption Data using Propensity Score Matching and the Infrequency of Purchase Model, with Application to Climate Change Policy

Nicholas Bardsley and Milena Büchs

We apply propensity score matching (PSM) to the estimation of household motor fuel purchase quantities, to tackle problems caused by infrequency of purchase. The results are compared to an alternative, regression-based, imputation strategy using the infrequency of purchase model (IPM). Using data from the UK's National Travel Survey (NTS) we observe that estimated mean obtained from the PSM imputation is closer to the estimated mean from the consumption diary, than that obtained from fitted values from the IPM. The NTS also contains an interview question on household mileage which can be used to assess the results of imputation. We find that the order statistics of the imputed distribution are more plausible for the PSM estimates than those obtained using the IPM, judging by the sample distribution of household mileage. We argue that there are some applications for which the PSM method is likely to be superior, including estimates of distributional effects of policies. On the other hand, the IPM is more suitable for analysing conditional effects and associations of consumption with covariates. We illustrate our arguments using a simple microsimulation exercise on CO_2 emissions reduction policies, an area where methods for coping with zero-inflated data seem currently to be under-used.

1. Introduction

Data on household consumption of goods and services, including those underlying national level statistics, often come from purchase diaries. The resulting data pose analytical problems because the diary is typically of a relatively short duration, say 1-2 weeks, with the result that a sampled household will often not be observed to make a purchase despite consuming the good. For example, it is known from everyday life that practically everyone consumes clothing, but there are many households that will not buy any clothes in a given week. A weekly diary instrument will record a substantial proportion of households as purchasing no clothing and the others making purchases of varying amounts. Assuming these are accurate records of purchases at the level of one week, the situation is unsatisfactory because in most applications the variable of interest is a rate of consumption, which can be expressed as weekly, monthly or yearly. Interpreting the diary data as yearly rates for each sampled household would yield the absurd conclusion that many or most households consume no clothing at all and others consume very large amounts. Also, applying standard OLS regression techniques results in biased coefficients and spurious standard errors.

Economists have developed models to deal with this and related problems, based on multivariate regression techniques and economic theory (Deaton and Irish (1984), Blundell and Meghir (1987), Kimhi (1999)). We are concerned in this article only with the case in which all households (or individuals, depending on the survey) consume the good or service in question, and so zero-valued observations only arise from infrequency of purchase. In this case the appropriate model is the "Infrequency of Purchase Model" (IPM) as set out in Blundell and Meghir (1987, p183). The IPM estimates simultaneously a logistic regression model of the purchase decision and a linear (or log-linear) regression model of the quantity purchased. Unbiased regression coefficients and valid standard errors can then be obtained, conditional on other modelling assumptions stipulating the error term and functional forms.

In this paper we are interested in exploiting Propensity Score Matching (PSM) to impute the distribution of rates of consumption.[1] We show below that the properties of propensity scores imply their usefulness, under certain conditions, for this imputation. This is, to our knowledge, a novel application of PSM, which is more usually applied to the problem of estimating treatment effects in

[1] We thank Dr. James Brown for suggesting to us the possible use of propensity scores in this context (personal correspondence), and Dr Sylke Schnepf for helpful comments, at the University of Southampton. This research was funded by the Economic and Social Research Council (project RES-000-22-4083).

observational studies. Little (1986) applies PSM to missing item-data problems in sample surveys, but does not consider purchase infrequency. The IPM might also be used for such imputation. *A priori* it is not clear which method, PSM or IPM, should be preferred for this task, since on the one hand the PSM method requires only one (logistic regression) equation to be estimated and so does not rely on the assumptions IPM makes about the error term or functional form of the second (consumption regression) equation. On the other hand the IPM uses more information in its estimate of probabilities, since these are jointly estimated with the quantity equation.[2] We therefore compare the approaches empirically using a survey that is especially suited to this task, namely the UK's National Travel Survey (NTS).

The NTS is distinctive in its use both of a consumption diary and an interview question, to assess household consumption of motorised transport. The diary records litres of liquid transport fuel *and* monetary expenditure on this item in one week. The interview question asks household representatives to state the mileage of any vehicle in their possession in the previous year. There should be a close relationship between a household's annual mileage and their actual weekly rate of fuel consumption, albeit confounded by the fuel efficiency of the vehicle and question-specific error. The inclusion of both questions holds constant any survey design effect across the two measures.

One check on the adequacy of imputation is consistency of the estimated mean with the sample mean from the recorded diary data. In addition to this we exploit the mileage information in the NTS by comparing features of the sample distribution of household mileage with those of the imputed distribution of household rates of consumption. The next section sets out the theoretical basis for our application of PSM to consumption data. We then elucidate the NTS data and set out the details of the imputation procedure and results. Next, we use the IPM to attempt the same imputation exercise, using the same specification of the logistic regression model, and compare the results of the two imputations. We then discuss the relative strengths and limitations of PSM and IPM for dealing with different types of research questions associated with consumption data.

2. Theory

Let Z denote a binary event with outcome r_1 if Z=1 and r_0 if Z=0. A propensity score, $e_i(\mathbf{X})$ is the conditional probability that Z occurs, given a vector of observed characteristics **X** of a unit of observation *i*. Rosenbaum and Rubin (1983) show that the propensity score is a 'balancing item', meaning that the distribution of **X** will tend to be the same for random samples of observations with similar values of $e(\mathbf{X})$, whether Z=1 or Z=0. This is a large sample property of propensity scores. True propensity scores are always unknown and can only be estimated, for example using a logistic regression on observed covariates. On condition that there are no "unobserved confounders", that is, no unmeasured covariates that affect both the probability of exposure and the potential Rosenbaum and Rubin (1983) also show that propensity scores can be used to correct for certain kinds of selection biases. In their framework, the key assumptions are

$$Z \perp (r_1, r_0) \mid \mathbf{X} \qquad (1)$$

and

$$0 < p(Z=1 \mid \mathbf{X}) < 1 \text{ for all } \mathbf{X}.$$

Together these imply $Z \perp (r_1, r_0) \mid e(\mathbf{X})$ and $0 < p(Z=1 \mid e(\mathbf{X})) < 1$ for all $e(\mathbf{X})$.

Each household has a value of both r_1 and r_0, referred to as its potential outcomes, only one of which is recorded in the dataset; r_{1i} is recorded iff $Z_i=1$ and r_{0i} is recorded iff $Z_i=0$. In the context of purchase infrequency, Z represents the event that a household is observed (Z=1) or not observed

[2] Gibson and Kim (2011) test the IPM in datasets where recorded purchases are highly infrequent, finding considerable bias compared to results on measured stocks. However, this evidence concerns a more complicated variant of IPM in which an additional source of zeros is allowed, namely non-consumption of the good. In this paper we study the IPM variant where zeros arise only from infrequent purchase.

(Z=0) making a purchase, and r_1 is the purchase under Z=1. Here, r_1 contains the only unknown quantities since $r_0 \equiv 0$ for each household. We therefore require only that $Z \perp r_1 | \mathbf{X}$ and $0<p(Z=1|\mathbf{X})<1$ for all **X**.

Estimated propensity scores, $\hat{e}(\mathbf{X})$, providing they are of sufficient quality, can be used to balance samples on their observed characteristics. If there are no unobserved confounders, property (1) implies that matching each household for whom Z=0 with one for whom Z=1, with approximately the same value of $\hat{e}(\mathbf{X})$, yields an estimate of the missing values of r_1 at that value of $\hat{e}(\mathbf{X})$. It follows that the set of households matched to Z=0 households provides an estimate of the entire set of missing values of r_1. The quality of these estimates will depend on both sample size and the quality of the estimated propensity scores.

Each value of r_1 is then multiplied by the corresponding estimated propensity score to yield an estimated rate of consumption per diary window time period. That is, values given by

$$\hat{c}_i = \begin{array}{ll} \hat{e}_i(\mathbf{X}).r_{1i} & if\ \mathrm{Z}_i = 1 \\ \hat{e}_i(\mathbf{X}).r_{1j:\hat{e}_j(\mathbf{X}) \cong \hat{e}_i(\mathbf{X}),\, \mathrm{Z}_j=1} & if\ \mathrm{Z}_i = 0 \end{array} \qquad (2)$$

constitute the estimated distribution of consumption. Although $\hat{c}$ is subscripted it is important to realise that a given imputed value is not a prediction *for that household*, since each value of e(**X**) is associated with a distribution of values of **X**, not a specific value of **X**. We discuss this point further in section 6.

The argument just given supporting inference from PSM to the distribution of r_1 is distinct from that given for causal inferences in observational studies. There, inferences from PSM are only supported about the mean of the variable of interest. Mean treatment effects can be estimated, but minimum, median and maximal effects, for example, cannot. In that context, *both* potential outcomes, r_1 and r_0, are of interest, and one of these is unknown for *each* observation. In the present context, the situation is different, in that only one of a household's potential outcomes is of interest.

In common with other applications of PSM, choices the analyst has to make include the method used to estimate the propensity scores, how to assess the quality of balance achieved, and the details of the matching algorithm.

3. Zero-inflated consumption data in the UK National Travel Survey

We consider data from the National Travel Survey, pooling data for years 2002-2008 to achieve a large sample size.[3] For these years there is a total of 57,069 fully cooperating households. Of these, 42,712 have vehicles, either cars, vans or motorbikes, but 17,485 (41% of the motoring households) did not buy fuel during the diary week. But the mileage question in the interview data, for the same households, reveals that only 70 vehicle-owning households report zero mileage. So apparently only around 0.2% of motoring households in the sample could have no fuel consumption. Therefore, almost all of the recorded zeros are attributable to infrequency of purchase. A histogram of the diary data is shown in Figure 1 below, showing a spike at zero and an extended tail to the right of the mean (26 litres, 1 s.f.). A histogram of the mileage data is shown in Figure 2 below.

We assume that the diary data, shown in Figure 1, are an accurate representation of what sampled households purchased in the diary week. It follows from the mileage data, however, that the real distribution of households' weekly rate of fuel purchase calculated over longer periods is very different. In theory, the mean of the diary-sampled fuel purchase variable nonetheless provides an unbiased estimate of the latter, given that the survey is a probability sample. Concerning the mileage data, we assume that each household's actual mileage is functionally related to its fuel

[3] The data are available on request through the Department for Transport.

purchases, but that this function can be heterogeneous between households. We therefore anticipate a strong relationship between the true distribution of mileage and the true distribution of fuel purchases. The mileage data are not wholly unproblematic, however, as Figure 2 displays several modes at salient numbers. In particular, each exact multiple of 5,000 miles is a local mode. It is possible therefore that some bias is introduced by a tendency for salient numbers to be reported.

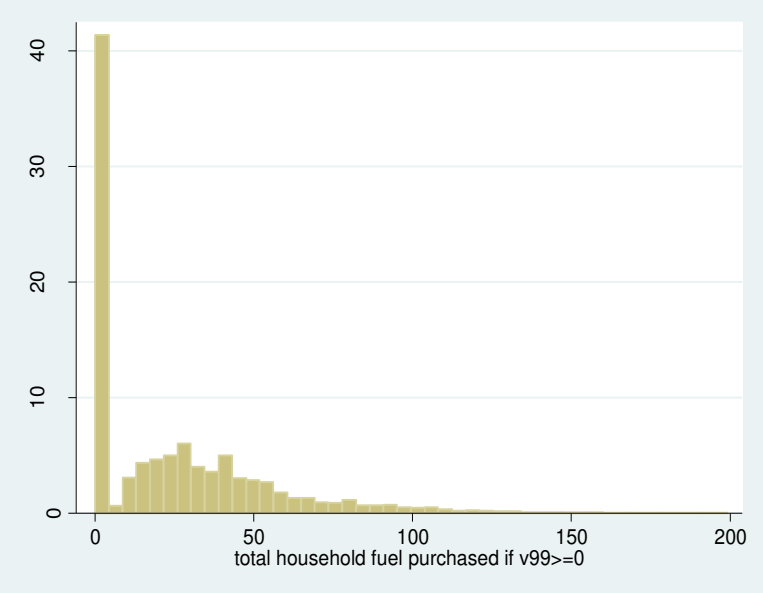


Figure 1 Histogram of Motoring Households' Fuel Purchases from the NTS One Week Expenditure Diary. Source: NTS 2002-2008. Censored at 200litres (>99th percentile; 40 observations excluded).

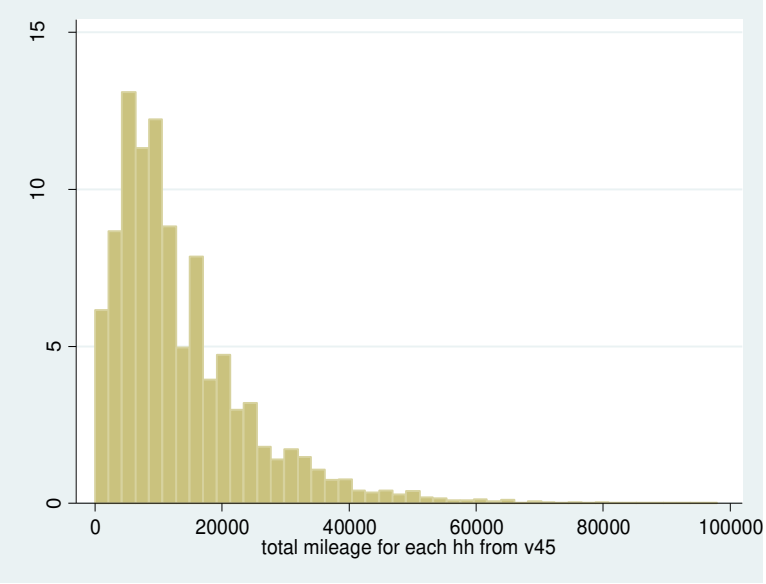


Figure 2 Histogram of Motoring Household's Recorded Mileage from the NTS Interview. Source: NTS 2002-2008. Censored at 100,000 miles (>99th percentile; 15 observations excluded).

4. Using PSM to estimate the distribution of rates of fuel purchase

It follows from the above discussion that the zero-inflation of the diary data, interpreted as estimates of each household's mean weekly purchase rate, is expected to be balanced in large samples by an over-representation of relatively high values. A desirable property of imputed purchases, therefore, is that they yield the same overall mean. We define the propensity score as

Probit regression				N	42600	
				LR chi2(35)	2862.7	
				Prob > chi2	0	
Log likelihood	-28684.4			Pseudo R2	0.0475	
	Coef.	Std. Err.	z	P>z	[95% Confidence	Interval]
spring	-0.08	0.02	-4.40	0.00	-0.11	-0.04
autumn	-0.14	0.02	-6.51	0.00	-0.18	-0.10
summer	-0.12	0.02	-6.28	0.00	-0.15	-0.08
h150	0.00	0.00	5.90	0.00	0.00	0.00
h14	0.05	0.01	6.59	0.00	0.04	0.07
h15	-0.04	0.01	-7.05	0.00	-0.05	-0.03
h20	0.02	0.01	1.50	0.13	-0.01	0.04
h24	0.02	0.01	3.77	0.00	0.01	0.03
h26	-0.03	0.01	-2.33	0.02	-0.06	0.00
h29	-0.01	0.01	-2.22	0.03	-0.03	0.00
h63	-0.08	0.02	-5.35	0.00	-0.11	-0.05
rural	-0.06	0.02	-2.83	0.01	-0.10	-0.02
adult2	-0.08	0.02	-4.10	0.00	-0.11	-0.04
adult3	-0.06	0.02	-3.11	0.00	-0.10	-0.02
child1	-0.04	0.02	-1.99	0.05	-0.09	0.00
child2	0.01	0.03	0.42	0.68	-0.04	0.06
child3	-0.05	0.04	-1.27	0.20	-0.12	0.03
bike1	-0.07	0.02	-4.40	0.00	-0.10	-0.04
bike2	-0.01	0.02	-0.65	0.51	-0.05	0.02
motorbikes	0.13	0.03	4.22	0.00	0.07	0.19
vehicles	-0.22	0.01	-16.23	0.00	-0.24	-0.19
large car	-0.04	0.02	-2.65	0.01	-0.07	-0.01
pensioners	0.12	0.02	5.13	0.00	0.08	0.17
working	0.09	0.07	1.35	0.18	-0.04	0.22
renters	-0.07	0.02	-3.71	0.00	-0.11	-0.03
professional	0.01	0.02	0.55	0.58	-0.03	0.06
clerical	0.02	0.02	0.76	0.45	-0.02	0.05
othermanan~s	-0.01	0.03	-0.21	0.84	-0.06	0.05
retired	0.32	0.07	4.59	0.00	0.18	0.46
econinactive	0.30	0.08	3.98	0.00	0.15	0.45
detatched	0.02	0.02	1.03	0.30	-0.02	0.06
semi	-0.02	0.02	-1.37	0.17	-0.06	0.01
flat	0.03	0.03	1.31	0.19	-0.02	0.09
convertedf~t	0.17	0.04	3.82	0.00	0.08	0.26
Estimated income (£)	-8.89E-08	4E-07	-0.24	0.813	-8.3E-07	7E-07
_cons	0.25	0.10	2.66	0.01	0.07	0.44

Table 1 Probit Regression stage of PSM imputation

the probability to purchase liquid transport fuel conditional on a household's covariate vector, and estimate it as a function of observed covariates. We perform this estimation using a probit regression, a common approach in PSM studies. The results of the probit estimation are shown in Table 1. The dependent variable is the decision *not* to buy fuel, for reasons of computing

convenience. The purpose of the regression is not primarily explanatory. It provides classification for matching purposes, and prediction for imputation purposes in the calculation shown in (2).

We use the results to obtain a set of paired households that bought and did not buy fuel, matched on the predicted propensity score obtained from the probit. The aim is that the two sets have very similar distributions of **X**. A commonly-used check on the quality of covariate balance between these sets is to calculate the standardised bias for each variable before and after matching (Rosenbaum and Rubin, 1985). The largest standardised bias here for any coefficient after matching is 1.9%. Since standardised bias of less than 10% seems generally to be regarded as negligible in PSM applications (Austin, 2011), we take this to indicate a high quality of matching on observed covariates on this measure. Also, despite the large sample size, there is no statistically significant difference between covariate means even at the 10% level.

For each vehicle-owning household we then have either a recorded purchase or we use the recorded purchase of its matched partner, as r_1. This is then multiplied by $\hat{e}(\mathbf{X})$, calculated as 1-$\hat{p}(\mathbf{X})$, where 1-$\hat{p}(\mathbf{X})$ is the estimated probability *not* to by fuel, calculated from the coefficients in Table 1. We then multiply each value of r_1, whether imputed or observed, by the household's value of $\hat{e}(\mathbf{X})$, the probability of purchase, to obtain a set of estimates of consumption.

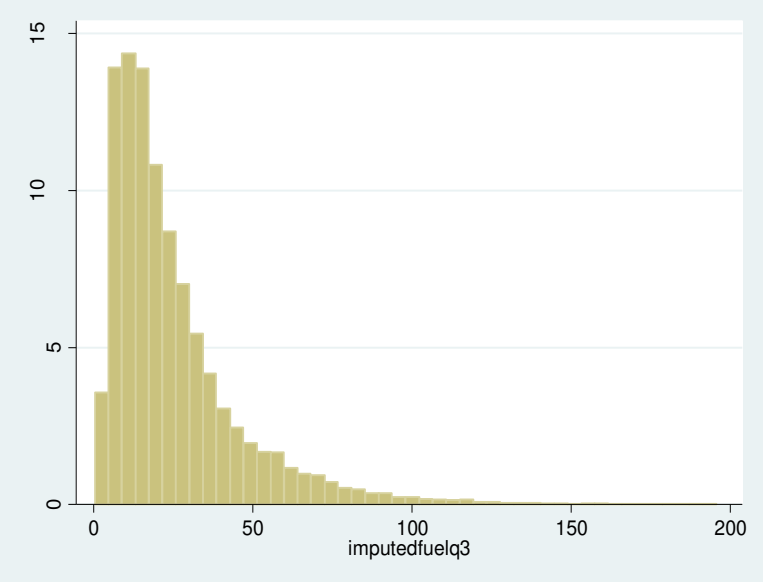


Figure 3 Histogram of imputed fuel purchases of motoring households, generated using PSM of Non-Purchasing to Purchasing Households. Source: NTS 2002-2008 and authors' calculations.

The results are shown in the histogram of Figure 3 above. This bears a strong qualitative resemblance to Figure 2, excepting the local modes of the latter at points of numerical salience. We now derive estimates of purchased fuel using fitted values from the IPM. A comparative evaluation of the estimates, analysing the extent of their isomorphism with the mileage data, is then conducted in section 6.

5. Using IPM to estimate the distribution of rates of fuel purchases

For full exposition of the IPM, see Blundell and Meghir (1987). It can be stated succinctly using its log likelihood function,

$$\text{Log L} = \sum_0 \log\left(1 - \Phi(z_i \cdot \theta)\right) + \sum_+ [-\log\sigma + \log\varphi\left(\frac{\Phi(z_i \cdot \theta) y_i - x_i \cdot \beta}{\sigma}\right) + 2\log\Phi(z_i \cdot \theta)] \qquad (3)$$

where y_i is the recorded purchase, z_i and x_i are covariate vectors for unit *i*, and θ and β are vectors of parameters to be estimated in the purchase and consumption equations respectively. This particular

specification is based on assumptions that zeros only arise from infrequent purchase, that recorded consumption entries (interpreted as rates over the diary period) are inflated in inverse proportion to the probability of purchase, that error terms in both the purchase decision and quantity decision are independent and gaussian with mean zero, and linear functional forms for both the purchase and quantity decisions.

We estimate the model using maximum likelihood in STATA.[4] We use the same specification of regressors for the purchase decision equation as for the PSM exercise to aid comparison. We modify the set of regressors for the purchase quantity equation by removing the dummy variables for the season in which the survey week fell, since this can be expected to affect the purchase decision in the diary week but not the seasonally-adjusted rate of fuel consumption. Similar adjustments between the two vectors of regressors are made by Blundell and Meghir (1987). Full results of the IPM and fitted consumption values are shown respectively in Table 2, and Figure 4 below.

6. Discussion

Comparing the histograms of distributions in Figure 3 and Figure 4, the PSM estimates reproduce the strong positive skew in the mileage data, but the IPM estimates do not. Basic statistics on the distributions of estimates, and of the mileage data, are provided in Table 3 below. These show that the PSM estimates are also closer to the observed mean for the NTS fuel purchase variable. The estimated coefficient of variation for the PSM estimates is also close to that obtained for the mileage data, whereas that for NTS fuel purchases exhibits over-dispersion.

A 2-sample t-test nonetheless rejects the null hypothesis of no difference in means between the PSM estimates and NTS fuel purchases (2-tailed $p = 0.04$; Satterthwaite's test). But the magnitude of the difference in estimates is 0.4 litres per week, (95% confidence interval $0.02 < x < 0.79$). In our view this does not represent a large substantive difference. It represents only ~2% of the estimated mean from the diary data, and would amount to 21 litres, so perhaps one or two acts of purchase over the course of a year.[5]

The imputation exercise using IPM produces a distribution of estimated fuel purchases closer to a normal distribution than that from the PSM exercise. It has also produced an estimated mean of 23 litres (95% c.i. $22.9 < x < 23.0$ litres), which is further away from the mean of recorded purchases than the mean of the PSM estimates. Arguably this is still, substantively, fairly close to the latter however. An improved mean prediction could perhaps be obtained via experimentation with the regressors. Of more concern is the basic shape of the distribution. The fitted values suppress the error term estimated by the model, but incorporating this would add noise symmetrically to the estimates, and so would not alter the skewness of the predicted distribution substantially.

[4] We adapt the program code given in the supplementary material of Gibson and Kim (2012).

[5] It should also be noted that the results are potentially sensitive to the analyst's decision about how to deal with topcoding of the NTS income data, since household income appears to play a key role in the estimated propensities and is also strongly associated with reported mileage. We experimented with various values to represent the midpoint of this income band, and excluding topcoded observations from the exercise, but found that these manipulations did not substantially affect the estimated mean. However, results derived using income variables for the upper quintile later in this paper should be interpreted tentatively, since the topcoding problem affects more than 50% of those observations. We assume a mid-point for the top income band of £85,000 per year.

IPM				N		42600
				Wald chi2(32) =		2526.2
Log likelihood	-147824.6			Prob > chi2 =		0
consumption	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
h150	-0.01	0.01	-0.97	0.33	-0.02	0.01
h14	-1.03	0.15	-6.85	0.00	-1.33	-0.74
h15	0.15	0.10	1.44	0.15	-0.05	0.35
h20	-0.31	0.21	-1.47	0.14	-0.71	0.10
h24	0.06	0.10	0.64	0.52	-0.13	0.25
h26	0.02	0.25	0.07	0.95	-0.48	0.52
h29	-0.12	0.12	-1.02	0.31	-0.36	0.11
h63	1.71	0.29	5.95	0.00	1.15	2.28
Rural	1.09	0.38	2.84	0.01	0.34	1.85
adult2	0.94	0.37	2.54	0.01	0.21	1.66
adult3	-0.32	0.38	-0.85	0.40	-1.06	0.42
child1	0.83	0.42	1.98	0.05	0.01	1.66
child2	0.42	0.49	0.86	0.39	-0.54	1.39
child3	0.77	0.71	1.09	0.28	-0.62	2.17
bike1	1.07	0.32	3.37	0.00	0.45	1.69
bike2	-0.19	0.35	-0.54	0.59	-0.87	0.50
no_bikes	-2.03	0.55	-3.71	0.00	-3.10	-0.96
no_veh	2.51	0.25	9.97	0.00	2.01	3.00
Bcar	4.21	0.29	14.60	0.00	3.65	4.78
Pensionerhh	-2.17	0.47	-4.64	0.00	-3.09	-1.25
Working	-0.18	1.18	-0.15	0.88	-2.50	2.14
Renters	-0.69	0.35	-1.96	0.05	-1.38	0.00
professional	3.79	0.42	9.07	0.00	2.97	4.61
Clerical	2.82	0.38	7.46	0.00	2.08	3.56
Othermanan~s	-0.09	0.47	-0.18	0.86	-1.00	0.83
Retired	-2.20	1.26	-1.75	0.08	-4.66	0.27
econinactive	-1.97	1.36	-1.44	0.15	-4.64	0.70
Detatched	1.05	0.36	2.88	0.00	0.33	1.76
Semi	0.50	0.31	1.59	0.11	-0.12	1.11
Flat	0.14	0.52	0.28	0.78	-0.87	1.16
convertedf~t	-0.33	0.92	-0.36	0.72	-2.14	1.48
estimated income (£)	5.7E-05	7.55E-06	7.52	0	4.19E-05	7.15E-05
_cons	14.34	1.76	8.14	0.00	10.89	17.80
Sigma	15.32	0.11	139.65	0.00	15.11	15.54

Purchase	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
Spring	0.05	0.01	5.58	0.00	0.04	0.07
Autumn	0.00	0.01	-0.04	0.97	-0.02	0.02
Summer	-0.02	0.01	-1.55	0.12	-0.04	0.00
h150	0.00	0.00	0.04	0.97	0.00	0.00
h14	-0.03	0.01	-6.66	0.00	-0.04	-0.02
h15	0.00	0.00	0.31	0.76	-0.01	0.01
h20	-0.02	0.01	-2.50	0.01	-0.03	0.00
h24	0.01	0.00	1.99	0.05	0.00	0.01
h26	0.01	0.01	1.41	0.16	-0.01	0.03
h29	0.00	0.00	1.13	0.26	0.00	0.01
h63	0.06	0.01	5.43	0.00	0.04	0.07
Rural	0.01	0.01	0.57	0.57	-0.02	0.03
adult2	0.14	0.01	10.12	0.00	0.11	0.17
adult3	0.04	0.01	3.38	0.00	0.02	0.07
child1	0.03	0.01	2.24	0.03	0.00	0.06
child2	0.01	0.02	0.78	0.44	-0.02	0.05
child3	-0.02	0.02	-1.00	0.32	-0.07	0.02
bike1	0.03	0.01	2.99	0.00	0.01	0.06
bike2	-0.02	0.01	-1.90	0.06	-0.05	0.00
no_bikes	0.23	0.02	11.96	0.00	0.19	0.26
no_veh	-0.17	0.01	-20.97	0.00	-0.18	-0.15
Bcar	-0.11	0.01	-9.47	0.00	-0.13	-0.09
Pensionerhh	-0.08	0.02	-4.20	0.00	-0.11	-0.04
Working	-0.13	0.05	-2.45	0.01	-0.23	-0.03
Renters	-0.06	0.01	-4.92	0.00	-0.09	-0.04
professional	0.09	0.01	6.43	0.00	0.06	0.11
Clerical	0.15	0.01	11.88	0.00	0.13	0.18
Othermanan~s	0.17	0.02	9.61	0.00	0.14	0.20
Retired	-0.07	0.05	-1.31	0.19	-0.18	0.04
econinactive	-0.08	0.06	-1.35	0.18	-0.19	0.04
Detatched	-0.07	0.01	-5.66	0.00	-0.10	-0.05
Semi	0.01	0.01	1.01	0.31	-0.01	0.03
Flat	-0.06	0.02	-3.12	0.00	-0.10	-0.02
convertedf~t	-0.10	0.03	-2.95	0.00	-0.16	-0.03
estimated income (£)	-3.28E-06	2.36E-07	-13.87	0	-3.74E-06	-2.81E-06
_cons	0.55	0.07	7.98	0.00	0.41	0.68

Table 2 Estimation of the IPM, DV = NTS Fuel Purchases, motoring households only; upper panel showing the estimated consumption quantity equation, lower panel showing the estimated purchase decision equation

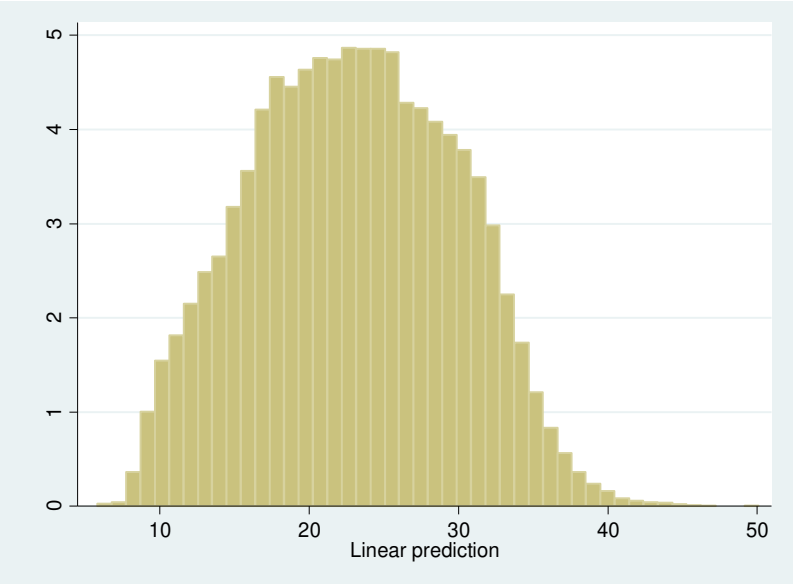


Figure 4 Histogram of fitted values from the IPM consumption equation. Uncensored.

Variable	Obs	Mean ($\bar{x}$)	Median	Std. Dev. (ssd)	Min	Max	Cov (ssd/$\bar{x}$)
NTS Mileage (miles)	42707	13708	10000	11296	0	153000	0.8
NTS Fuel (litres)	42600	26.0	18.0	33.1	0	721	1.3
PSM Fuel (litres)	42600	25.6	18.9	22.8	0.4	493	0.9
IPM Fuel (litres)	42600	23.3	23.3	7.0	5.8	51	0.3

Table 3 Descriptive statistics of NTS variables and imputed fuel purchases derived from PSM and the IPM
Note: the lower value for N for fuel purchases occurs because of missing fuel purchase diary entries.

For further exploration of the relationship between the imputed fuel purchases and the mileage data we use a Q-Q plot. This consists of a scatterplot of sorted values of each variable, so that each point represents the same quantile of each distribution. Q-Q plots for the PSM and IPM estimates, against mileage, are shown in Figure 5 below. Excepting around the top dozen paired observations, inspection of the upper plot reveals a roughly linear relationship between the quantiles of the two variables, whereas the lower plot forms a pronounced arc with a central deflection towards the x-axis. This reflects the contrast between the heavy positive skew of the mileage distribution and relative absence of skew in the IPM estimates. The different scale on the x-axis between the upper and lower plots also reflects this difference. The points in the upper figure in fact also form a slight arc towards the y-axis, reflecting higher skewness in the imputed fuel consumption values than in the mileage data.

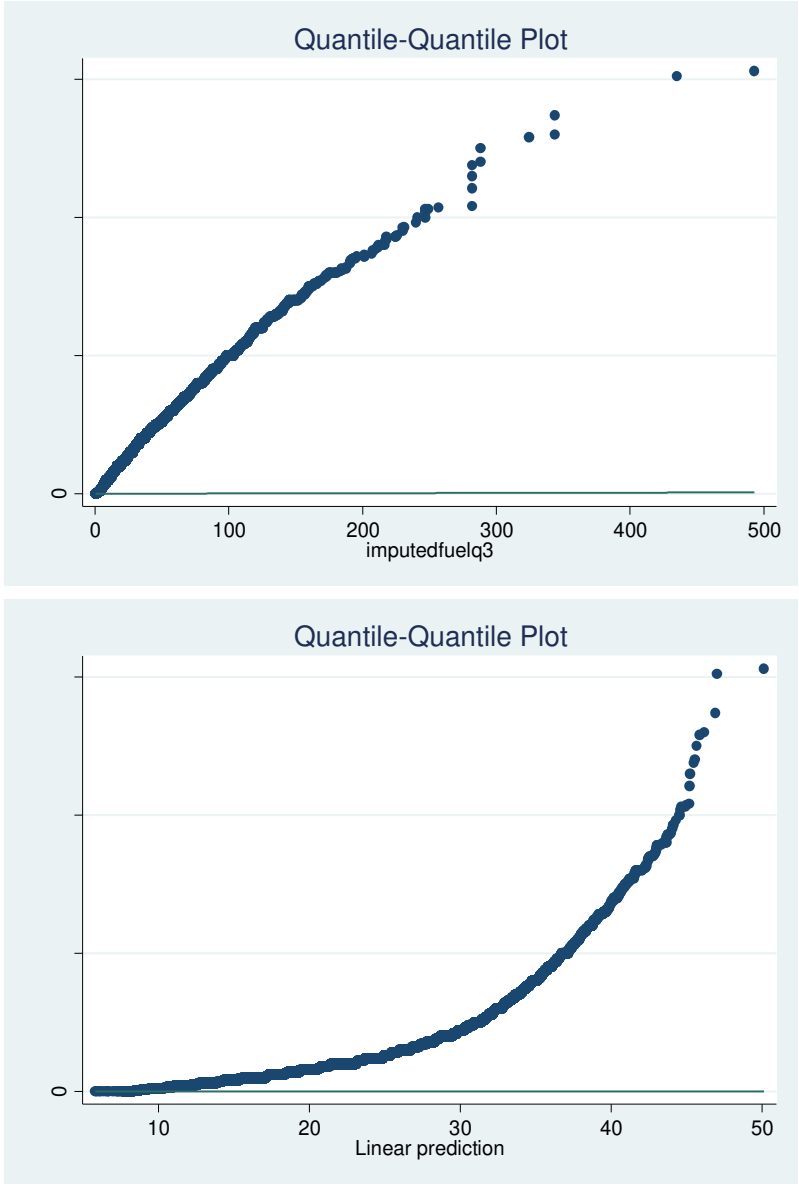


Figure 5 Q-Q plots of NTS mileage data against imputed Fuel purchases derived from PSM (upper) and the IPM (lower)

A plausible reason for the poorer performance of the IPM in respect of isomorphism to the mileage data is the stronger set of assumptions imposed using the IPM model. The PSM method uses similar assumptions to the IPM concerning the purchase decision, but imposes no specific structure on estimated quantities.

However, the pimary application of the IPM is not to simulate distributions of choices, as opposed to obtaining improved regression coefficients and standard errors. An important distinction between regression-based methods and applications of PSM is that the former, and not the latter, provide adjustment by *controlling for* **X**. PSM methods in contrast provides adjustment by *balancing on* **X** (Rosenbaum, 1998, p3553-4). Thus, the results from the IPM estimation, but not the PSM imputation, are informative about the relationship between consumption and the independent variables, and so insights are available through the former that are not through the latter.

In addition, for explanatory purposes the coefficients on the logistic regression in Table 2 should be preferred to those in Table 1, since they are estimated jointly with the coefficients for the consumption equation. It thereby takes into account the cardinal information contained in purchase quantities, whereas this is transformed into binary data for the PSM's probit model. An example of a

clear insight from the IPM in Table 2 above is that richer households, other things being equal, are estimated to buy fuel less frequently than poorer ones, but to consume it in larger quantities. This implies that when they do purchase fuel they make purchases that are larger by an amount that more than offsets the lower frequency. They do not seem simply to purchase fuel more frequently. Both of the IPM coefficients on income are highly significant, whereas the coefficient on income is not significant in the probit model used for PSM.[6]

Because of the difference between balancing on and controlling for **X**, we cannot simply analyse the consumption estimates from the PSM by income, occupation, household composition and so on, or use the data to analyse correlations. One can, however, repeat the PSM imputation exercise for sub-populations of interest (Rosenbaum, 1998). In so doing, one does not control for confounding associations, so in looking for example at lower income households the association between the variable of interest with income *per se* is not separable from associations between income and education or between income, household composition and other socioeconomic factors. Such uncontrolled associations, however, are often of policy interest.

7. Illustrative Application to Climate Change Policy

We now estimate a simple microsimulation of a carbon tax on motor fuels, using the NTS diary data and PSM imputation method. Mean effects of policies can be estimated without any such imputation. For policy analysis, though, other aspects of estimated impacts matter than mean effects. The range of estimated outcomes is also important, particularly amongst vulnerable groups. Measures which impact heavily on large numbers of disadvantaged households, or that are expected to benefit many affluent households will be difficult to justify politically, even if on balance they are progressive. If, then, one were interested to estimate effects of Carbon taxes or rations on motor fuels, as analysed for example by Comhar (2008), the zero-inflated nature of the data poses considerable limitations. The modelling used to estimate effects of such policies has achieved considerale technical sophisitication, as is evident for example in coupled Energy, Environment and Economy models (Barker, 1998). However, the zero inflation of the data, if not adjusted for, will restrict the insights available through the models since they are estimated using consumption data as a key input. The infrequent purchase problem is seldom discussed in the climate policy literature, however, despite its relatively heavy reliance on consumption data.

For illustrative purposes it is appropriate to use a simple model. We restrict our attention to what is perhaps the simplest available approach, namely static microsimulation. This technique estimates policy outcomes on the assumption that behaviour does not change, yielding estimates which are usually interpreted as ones of initial effects. The results of the PSM for fuel consumption can be transformed into the estimated payments of a £100/tCO_2 emissions tax simply by a lateral translation of Figure 3, given by

$$\hat{t}_i = £100 \text{ x } \hat{c}_i \text{ x } 52(\text{weeks}) \text{ x } 2.49(\text{kgCO}_2/\text{ltr})/1000. \qquad (4)^7$$

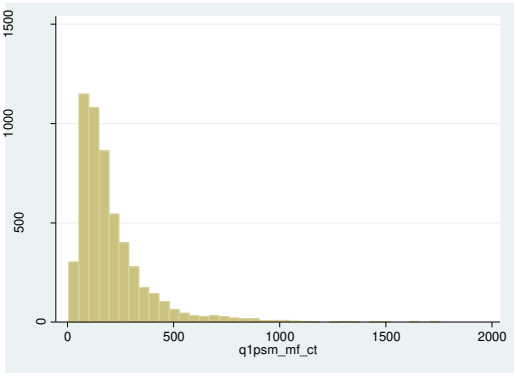
We then repeat the PSM estimation exercise of $\hat{t}_i$ for the 5 income quintiles reported in the NTS separately. The results are shown in Figure 6 and Figure 7, as sample frequencies for motoring households and as percentages for all households respectively. The shape of the distributions change markedly across the quintiles, becoming less skewed in higher quintiles. Thus, the higher mean tax payments amongst motorists known to obtain at higher incomes appear to be the product of a general shift in the distribution towards higher fuel consumption. This is also evident in the corresponding mileage distributions, and is not surprising. Note however, that the mileage distribution is not ideal for estimation of effects of a CO_2 tax because of heterogeneity in the fuel

[6] We have since corroborated this estimation result on data from the Living Cost and Food Survey. Details are available on request.

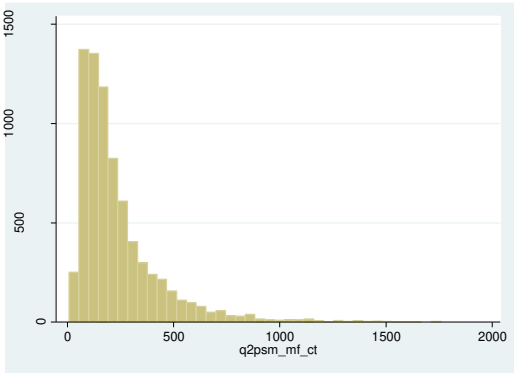
[7] Given that $\hat{c} \perp Z | e(\mathbf{X})$ it follows that a univariate function of $\hat{c}$ is also conditionally independent of Z given $e(\mathbf{X})$ (Dawid, 1979, lemma 4.2). $\hat{t}$ is such a function, so is underpinned by the same argument as given for $\hat{c}$.

efficiency of vehicles and in driver behaviour, both of which impact on fuel consumption and therefore emissions per mile driven.

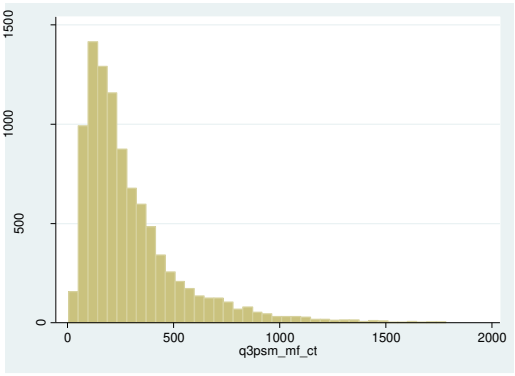
An estimate of any percentile of each distribution is now available. There are many reasons why this information is valuable. For example, the median is a better measure of a representative value in a distribution than the mean for many purposes, since the latter is influenced by extreme values.



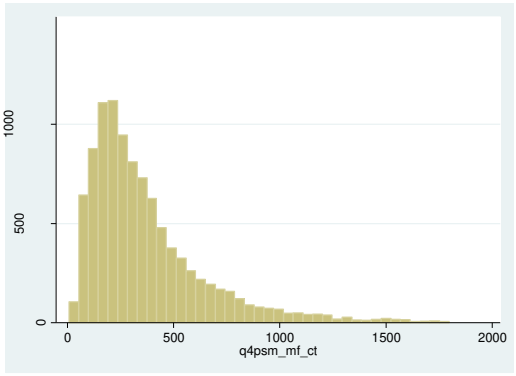
Quintile 1



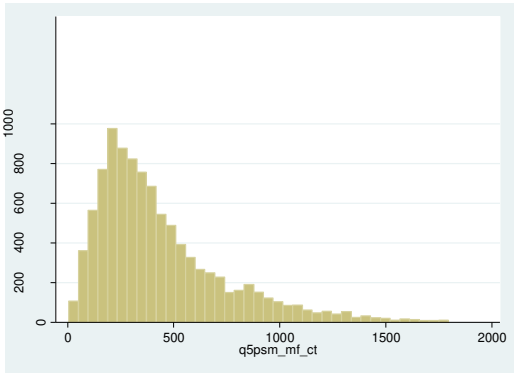
Quintile 2



Quintile 3

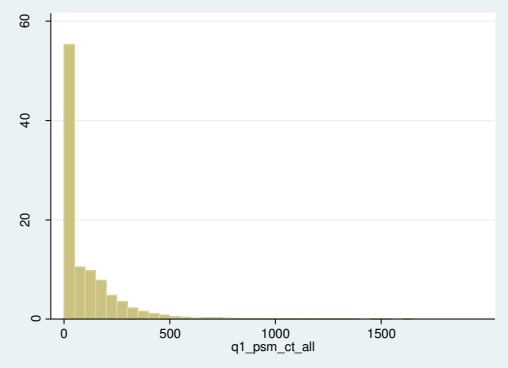


Quintile 4

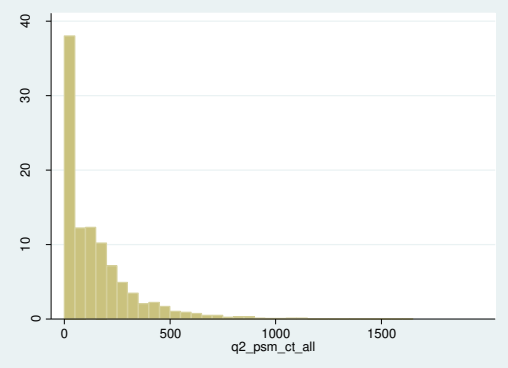


Quintile 5

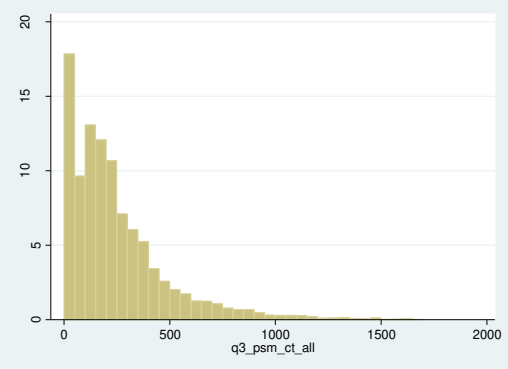
Figure 6. Estimated initial effects of a £100/tCO2/year Carbon Tax on Motor Fuels by income quintile, as sample frequencies, using PSM. Motoring households only. Histograms are Censored at £2000 (the 99th percentile of the distribution for quintile 4).



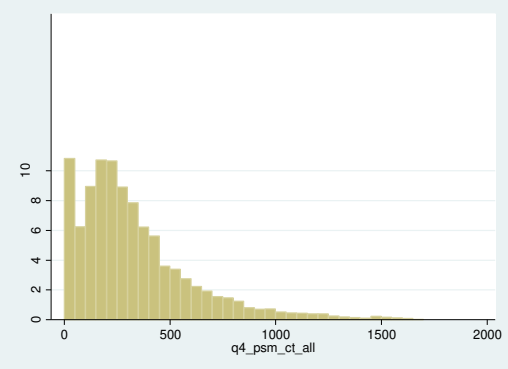
Quintile 1



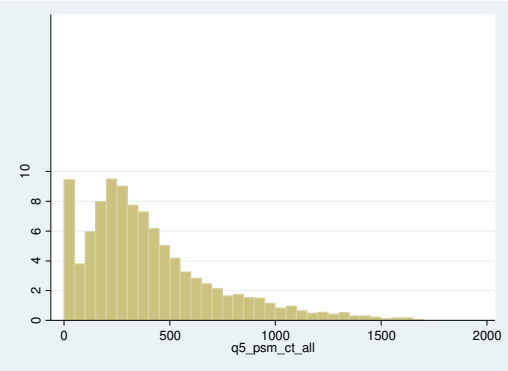
Quintile 2



Quintile 3



Quintile 4



Quintile 5

Figure 7: Estimated initial effects of a £100/tCO2/year Carbon Tax on Motor Fuels by income quintile, as percentages, using PSM. All households. Histograms are Censored at £2000 (the 99th percentile of the distribution for quintile 4).

The spread of effects around a mean value may also be politically sensitive. The mean CO_2 tax over the whole sample of motoring households is estimated as £342. An estimated 14% of the motoring households in the lowest income quintile pay more than this, judging by the figures represented in Figure 6. This is likely to be contentious because the mean charge represents a relatively large share of their income. On the other hand, less than 6% of households in this quintile, inclusive of non-motoring households, are estimated to pay this much.

Using the diary data represented in Figure 1, in contrast, 25% of motoring households in the lowest income quintile would have an estimated tax burden greater than the estimated mean charge. But there is no justification for using this figure as an estimate given the purchase

infrequency problem. Bias occurs predictably because zero inflation is balanced at the mean purchase by inflated values.

We next note that it may often be possible to go beyond the simple transformation of the estimated purchase quantity conducted above. Consider that, for each household that was observed to make a purchase, we can derive its estimated consumption according to (2), independently of any matching, using the logistic regression results. We can similarly observe for each such household the joint occurrence of estimated consumption with another component of **X**. For any household for whom r_1 is observed, not imputed, for example, the household's income and household size are also observed. So for these households we could also derive the estimated carbon tax as a proportion of income, or work out their estimated net payments under a tax and rebate scheme. Given parallel conditional independence assumptions to those underpinning the consumption estimates in (1), an estimated distribution for all households might also be inferred for these effects, via matching. Specifically, we require that:

$$Z \perp (s, r_1) \mid e(\mathbf{X}) \qquad (5)$$

Where s is the additional covariate used to calculate the policy outcome. (5) says that observation of purchase is conditionally independent of the joint distribution of potential purchased quantities and s, given the propensity score. Conditional independence of r_1 with Z is already assumed and conditional independence of s with Z is already examined if s is one of the components of **X** on which e(**X**) is estimated. The additional assumption required is that the association between s and r_1 is conditionally independent of Z given the propensity score. The plausibility of (5) will need to be considered case by case.

As indicated above, it would be of interest for policy analysis to estimate the financial results of any CO_2 policy as a proportion of income (t/y), to examine their possible regressivity. We calculate $\widehat{t/y}$ for households that purchased fuel, and then use PSM-matched values of $\widehat{t/y}$ for motoring households that did not. That is,

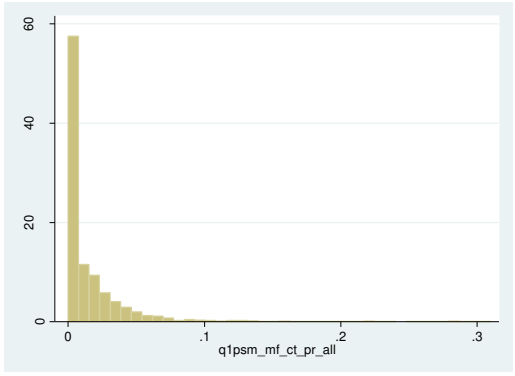
$$\widehat{t/y}_i = \begin{cases} \hat{t}/\mathrm{y} & if\ \mathrm{Z}_i = 1 \\ \widehat{t/y}_{j:\hat{e}_j(\mathbf{X}) \cong \hat{e}_i(\mathbf{X}),\, \mathrm{Z}_j = 1} & if\ \mathrm{Z}_i = 0 \\ 0 & \text{if non-motoring} \end{cases} \qquad (6)$$

We show the results in Figure 8 below. By the argument just given, the exercise makes the additional assumption that the association between *y* and *t* is independent of Z given **X**. However, we cannot observe this covariance for households that did not purchase fuel, so we cannot assess (5) directly. Using the NTS we can examine the likely association with recourse to the NTS mileage data, though, since mileage is assumed to be a function of *c*. Pearson's correlation coefficient between household mileage and income takes the value of 0.37 (95% c.i. $0.36 \leq \rho \leq 0.38$) for motoring households that did not purchase fuel, versus 0.44 (95% c.i. $0.43 \leq \rho \leq 0.46$) for the matched set of households that did. Thus, these results deserve somewhat more circumspection than those in Figure 7, given the evidence of a difference in the degree of association.

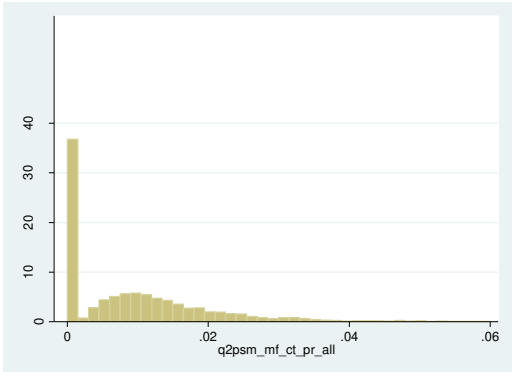
The estimates represented in Figure 8 enable an assessment of the progressivity of the policy. This is easier to ascertain numerically. Descriptive statistics from the estimated distributions of $\widehat{t/y}$ are therefore shown in Table 4 below. The table shows that the tax is estimated to be regressive evaluated at the mean. But it also shows that to be entirely attributable to the upper part of the estimated fuel purchase distribution, that is, to a small minority of low income households with unusually high fuel consumption. The policy is, in contrast, slightly progressive towards the lowest income quintile, if evaluated at the median, which is here a better indicator of typical effects. Amongst motorists the policy is more clearly regressive, but the effect is again exaggerated by positive skew if evaluated at the mean. It is clear that some low income households are estimated to

be quite adversely affected in the absence of revenue recycling or behaviour change, with 5% of the lowest income quintile estimated to pay more than 7% of their income as CO_2 tax on motor fuels.[8]

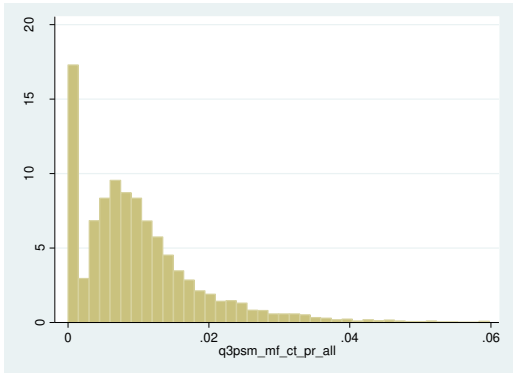
Our results here contrast with earlier reports in the climate change policy literature, also based on static microsimulation, that CO_2 taxes on motor fuels are progressive evaluated at the mean and only regressive amongst motorists (Dresner and Ekins, 2004). This difference in results is likely to be partly attributable to increasing car ownership over time. According to NTS estimates, 52% of households in the lowest income quintile owned or rented a car by 2012, up from just 34% in 1995/1997 (DFT, 2012). We do not offer the estimates in this paper primarily as policy analysis, however, rather than illustration of method, as the simulation approach is extremely simple. The



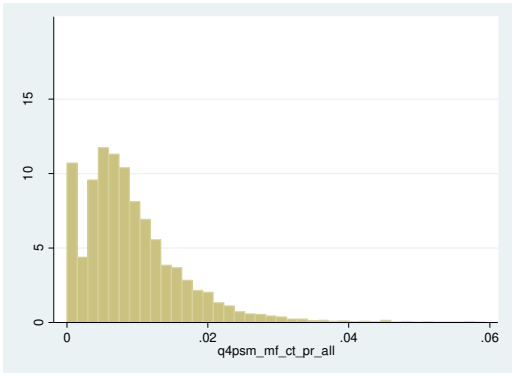
Quintile 1



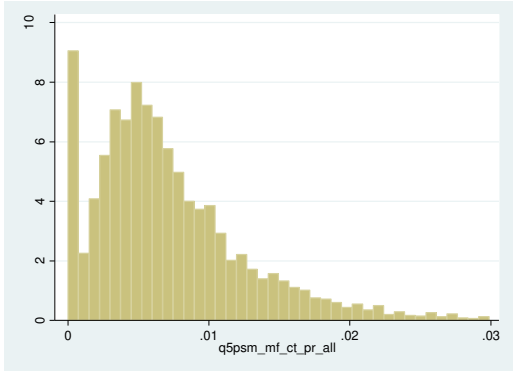
Quintile 2



Quintile 3



Quintile 4



Quintile 5

Figure 8. Estimated initial effects of a £100/tCO_2 tax on motor fuels by income quintile, as a proportion of income, using PSM. All households. Histograms are censored at the 99th percentile of each distribution

[8] Herein lies the main limitation of static microsimulation which assumes unchanging behaviour, whilst the point of CO_2 taxes is precisely to cause people to emit less. The results may nonetheless be informative about likely sources of resistance to the policy.

All households					
Quintile	mean	lower quartile	median	upper quartile	95th percentile
1	2.0	0.0	0.0	2.0	7.0
2	1.0	0.0	0.7	1.5	3.2
3	1.0	0.4	0.8	1.4	2.8
4	0.9	0.5	0.7	1.2	2.3
5	0.7	0.3	0.6	1.0	1.8
Motoring households					
Quintile	mean	Lower quartile	median	upper quartile	95th percentile
1	4.6	1.3	2.2	4.1	14.1
2	1.6	0.8	1.3	1.9	3.7
3	1.2	0.6	1.0	1.5	3.0
4	1.0	0.5	0.8	1.3	2.4
5	0.8	0.4	0.6	1.0	1.9

Table 4 Descriptive statistics on estimated initial effects of a £100/tCO_2 tax on motor fuels as a percentage of household income, using PSM

interaction of low income households with the tax and benefits system is not taken into account, for example, and so the income of the lowest income households is perhaps under-stated, which would exaggerate the impacts on these households. The type of analysis just conducted is not available without a method of imputing the distribution of consumption however.

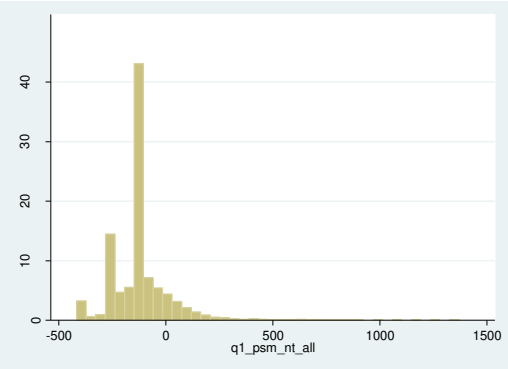
For a second example, in Figure 9 below we show estimated effects of a CO_2 tax and rebate scheme. To calculate this requires using the additional variable, *adults*, that is, the number of adults in the household. Each household is assumed to receive a lump sum, equal to the mean value of CO_2 contained in households' purchased motor fuel in a year (1.4 tons), multiplied by £100/t CO_2, multiplied by the number of adults in the household, from which the CO_2 tax is subtracted. One can think of this as a per-adult share of the tax revenue. The net tax payment, $\hat{k}$, is calculated from observed r_1 and s, and $\hat{e}(\mathbf{X})$, for each unit for which Z=1. Units of observation with Z=0 have their values imputed from a matched case, as in the previous PSM exercises. Finally we add the non-motoring households, for whom $\hat{k}$ is simply the product of *adults* and the per-adult permit value. To aid comparison with the other figures, *k*<0 indicates that a household benefits financially.

That is,

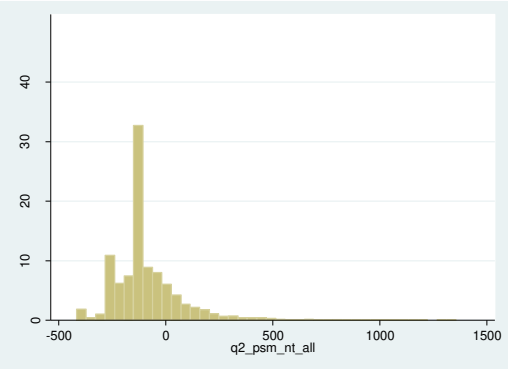
$$\hat{k}_i\,(£) = \begin{cases} adults_i \times 1.4 \times 100 - \hat{t}_i \times -1 & if\ \mathrm{Z}_i = 1 \\ \hat{k}_{j:\hat{e}_j(\mathbf{X}) \cong \hat{e}_i(\mathbf{X}),\ \mathrm{Z}_j = 1} & if\ \mathrm{Z}_i = 0 \\ 1.4 \times adults_i \times 1.4 \times 100 \times -1 & \text{if non-motoring} \end{cases} \quad (7)$$

We estimate $\hat{k}$ separately for each income quintile, with 5 separate PSM exercises, as before. From the estimated distributions we infer that a tax and rebate scheme would benefit the majority of households in quintiles 1, 2 and 3. The two upper quintiles would on average transfer income, thus the measure is broadly progressive. The spikes in the distribution constitute concentrations of the non-motoring households. Under this policy they benefit by a lump sum for each adult occupier. By ignoring these modes one can also visualise the that the rebated policy is estimated to be progressive amongst motoring households.

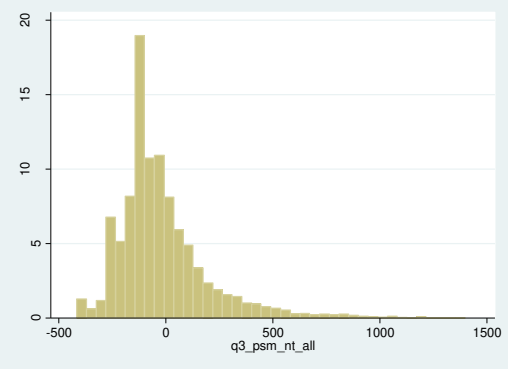
Is (5) plausible in this case? Here, s = *adults*, is again a component of **X** used to estimate e(**X**). The s component of the covariate balance assessment, which is conducted as part of the PSM exercise is therefore relevant to an assessment of (5). In this case we have already judged this to be satisfactory for all variables used in the PSM exercise, conducted for the sample as a whole. We report here that the balance is also satisfactory, albeit less so, on the variable *adults* in the separate estimations for each quintile, with a maximum standardised bias of 7% (quintile 1).



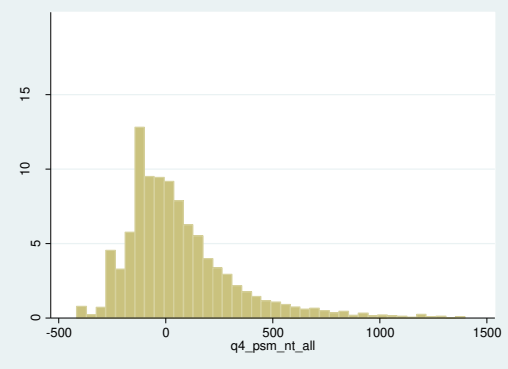
Quintile 1



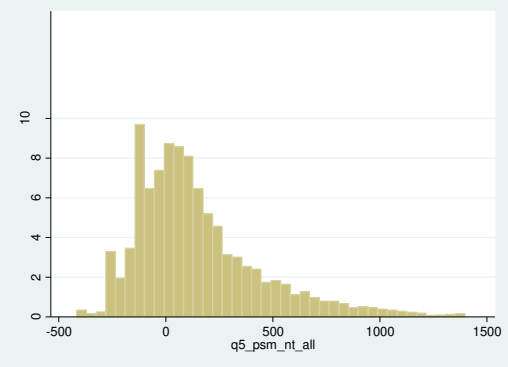
Quintile 2



Quintile 3



Quintile 4



Quintile 5

Figure 9. Estimated initial effects of a £100/tCO2/year Carbon Tax and Rebate Scheme on Motor Fuels by income quintile, using PSM. All households. Histograms are Censored at the 99th percentile of the quintile with the highest-valued 99th percentile.

We again use the NTS mileage data to assess the likely association between r_1 and *adults* for Z=0 and Z=1. Pearson's correlation coefficient between household mileage and *adults* takes the value of 0.25 (95% c.i. $0.24 \leq \rho \leq 0.27$) for motoring households that did not purchase fuel and 0.28

(95% c.i. $0.27 \leq \rho \leq 0.29$) for the matched motoring households that did.[9] Thus, we cannot be confident that the association is not different given **X**, but the estimated correlation coefficients are of the same sign and very similar magnitude. On the grounds that mileage is intimately related to fuel purchases, this provides some confidence that imputing the distribution of net tax payments after rebates does not introduce a new unobserved confound.

For completeness we now estimate the policy outcome of the rebated CO_2 tax as a proportion of income, d. We here assume

$$Z \perp (adults, income, r_1) \mid e(\mathbf{X}) \qquad (7)$$

and assess the quality of this additional assumption by examining the covariance of the interacted variable, *adults*income*, with household mileage. Pearson's correlation coefficient between household mileage and *adults*income* takes the value of 0.40 (95% c.i. $0.39 \leq \rho \leq 0.41$) for motoring households that did not buy fuel, versus 0.45 (95% c.i. $0.43 \leq \rho \leq 0.46$) for the matched households that did. So again there is evidence this association is not identical between the two groups, but that it is the same sign and similar in magnitude. The estimates are shown in Figure 10 and Table 5 below.

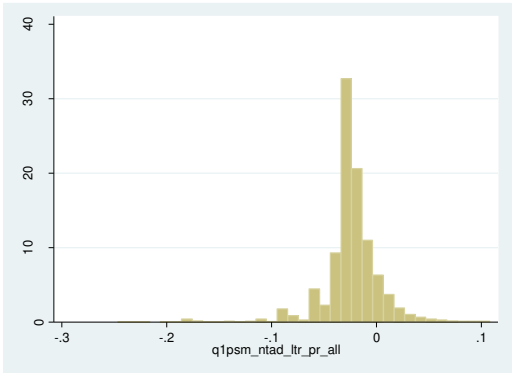
The estimates suggest that a rebated motor fuel carbon tax, or equivalently, 'cap and share' tradeable permit scheme, would be strongly progressive. A large majority of households are estimated to benefit overall, financed by a transfer from the upper two quintiles of the income distribution which is relatively small as a percentage of their income. Further, the rebated policy also appears to be progressive amongst motorists, though just over 30% of lower income motoring households are predicted to lose financially.

For comparison we show also in Table 6 the same statistics for this policy estimated, naively, using the fuel purchase diary data. There is close agreement at the mean between the PSM- and NTS diary-based estimates. But, as would be expected, they diverge at other points of the distributions because of the over-dispersion in the diary data. What is less obvious is that the distorting effects of purchase infrequency on these estimates are not constant across income quintiles. One reason for this is that vehicle ownership is concentrated at higher incomes. We label a zero purchase which occurs because of purchase infrequency rather than zero mileage a "false zero". A zero purchase occurring in a higher income quintile is more likely to be a false zero, reflecting the gradient in vehicle ownership rates. Secondly, lower income drivers tend to have lower mileage. So the balancing effect of false zeros against inflated values tends to misclassify households as having high per-adult mileage more strongly here than at higher incomes. These two effects, which tend to understate the progressivity of the policy, are evident in Figure 11 below. The diagram is generated by counting instances of households which are predicted to lose (win) using the NTS fuel purchase diary, but which are predicted to win (lose) using the NTS mileage data. A household's predicted outcome is calculated from the mileage data as follows:

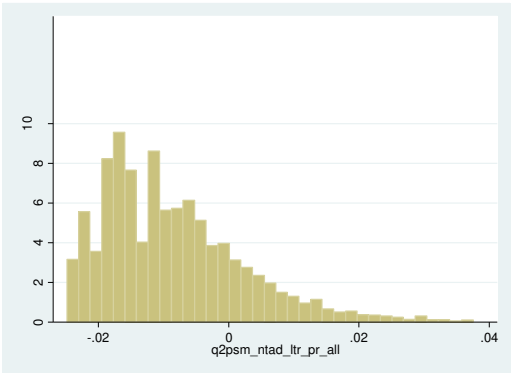
$$\text{household } i \quad \begin{matrix} wins \\ loses \end{matrix} \quad \leftrightarrow \frac{mileage_i}{(\sum_i mileage)/(\sum_i adults)} - adults_i \begin{matrix} < \\ > \end{matrix} 0 \qquad (8)$$

Figure 11 shows that the naïve estimates derived from the diary data will underestimate numbers of beneficiaries of the policy at low incomes and overestimate them at high incomes. It is therefore encouraging for the PSM method that the estimates in Table 5 posit higher numbers of beneficiaries at low incomes and lower ones at high incomes than those in Table 6. We also show the counts of winners and losers generated using definition (8) as percentages of each income quintile, using the NTS mileage data, in the rightmost column of Table 6. These accord closely with the counts in the rightmost column of Table 5, providing additional support for the PSM estimates.

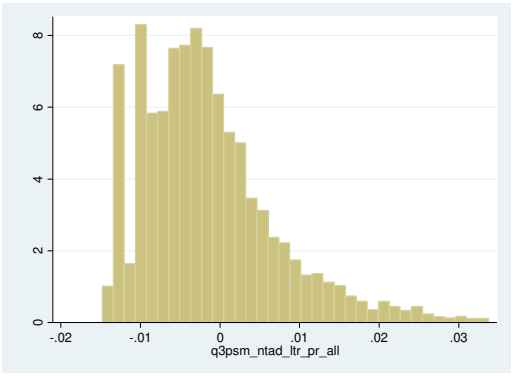
[9] Confidence intervals for ρ are calculated using the corrci routine in Stata (Cox, 2008).



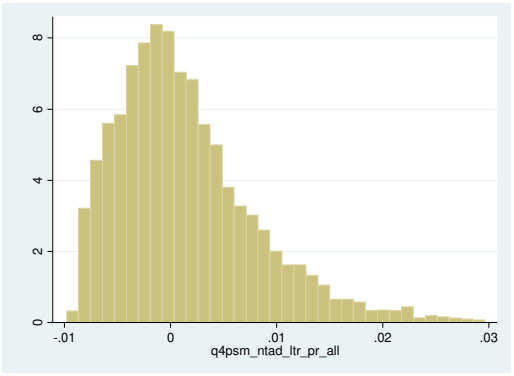
Quintile 1



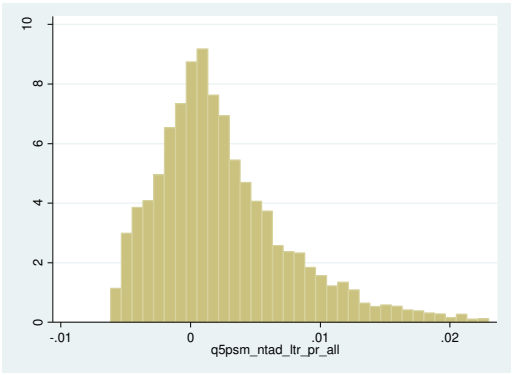
Quintile 2



Quintile 3



Quintile 4



Quintile 5

Figure 10. Estimated initial effects of a £100/tCO_2 tax and rebate scheme, as a proportion of income, using PSM. All households. Histograms are censored at the 1st and 99th percentile of each distribution.

All households quintile	mean	Lower quartile	median	upper quartile	95th percentile	% that gain
1	-2.7	-3.2	-2.5	-1.3	2.1	86.7
2	-0.8	-1.7	-1.1	-0.3	1.4	80.3
3	-0.1	-0.8	-0.3	0.3	1.6	65.3
4	0.2	-0.3	0.0	0.5	1.5	47.8
5	0.3	-0.1	0.2	0.5	1.3	34.2
Motoring households quintile	mean	Lower quartile	median	upper quartile	95th percentile	% that gain
1	-0.5	-2.2	-1.2	0.3	4.6	71.5
2	-0.2	-1.0	-0.5	0.2	1.9	69.1
3	0.1	-0.5	-0.1	0.4	1.8	58.3
4	0.3	-0.2	0.1	0.6	1.6	43.0
5	0.3	-0.1	0.2	0.5	1.3	28.1

Table 5. Estimated initial effects of a £100/tCO_2 tax and rebate scheme, as a proportion of income, using PSM, by income quintile

All households quintile	mean	Lower quartile	median	Diary Data upper quartile	95th percentile	% that gain	Interview data % that gain
1	-2.7	-3.7	-2.5	-1.8	4.5	83.9	86.7
2	-0.8	-2.0	-1.5	-0.1	2.7	75.9	79.9
3	-0.1	-1.2	-0.8	0.6	2.6	63.1	62.2
4	0.2	-0.7	-0.1	0.8	2.3	52.3	45.3
5	0.3	-0.4	0.1	0.7	1.9	47.4	32.8
Motoring households quintile	mean	Lower quartile	median	upper quartile	95th percentile	% that gain	% that gain
1	-0.6	-3.3	-1.9	1.0	8.4	65.8	71.6
2	-0.2	-1.7	-0.9	0.8	3.6	62.2	68.4
3	0.1	-1.0	-0.3	0.8	2.9	55.7	54.6
4	0.3	-0.7	0.1	0.9	2.4	48.1	39.3
5	0.3	-0.4	0.2	0.8	2.0	42.5	26.6

Table 6. Estimated initial effects of a £100/tCO_2 tax and rebate scheme, as a proportion of income, using NTS data, by income quintile

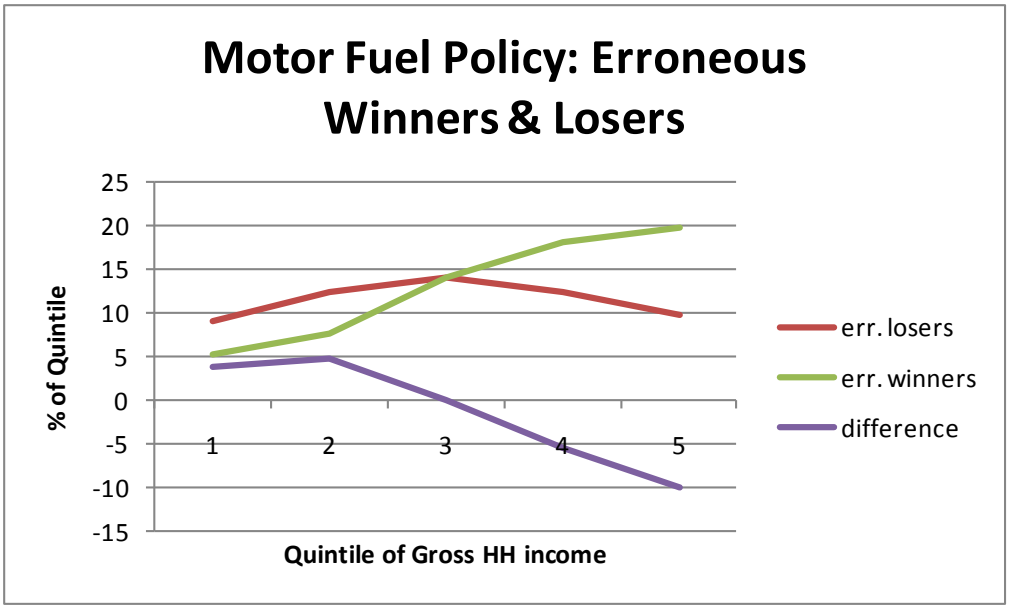


Figure 11. Erroneous classifications of households under a CO_2 tax and rebate on motor fuels, if derived using the NTS fuel purchase diary, by quintile of household income

Conclusions

PSM seems to provide a promising strategy for dealing with problems posed by purchase infrequency for specific research questions. In particular, there appears to be potential to estimate distributions of consumption rates and of effects of policies which are contingent on those rates. Of particular value is the potential to estimate quantiles of the distribution, rather than the mean. The use of PSM in this context should be seen as complementary to the IPM and seems to be more robust for estimating quantiles of the distribution of consumption and of related variables. But PSM is unsuitable for other types of research question. In particular, owing to the fact that PSM balances on covariates but does not control for them, the PSM-derived consumption estimates do not provide a basis for prediction conditional on particular values of covariate vectors, or for estimating regression coefficients.

To illustrate its potential we applied the PSM imputation technique to a simple static microsimulation problem. The results suggest that a CO_2 tax on motor fuels would be regressive, but that a rebated tax or cap and share scheme would be strongly progressive, even amongst motorists. The picture concerning regressiveness also appears to be complicated, however, by the strong positive skew of transport consumption, in ways that cannot be ascertained using estimates of means. We conclude that PSM merits further consideration in the context of purchase infrequency.

References

Austin, P.C. (2011). An introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies.

Barker, T.S. (1998). Use of Energy-Environment-Economy Models to Inform Greenhouse Gas Mitigation Policy. *Impact Assessment and Project Appraisal*, 16, 123-131.

Blundell, R. and Meghir, C. (1987). Bivariate Alternatives to the Tobit Model. *Journal of Econometrics*, 34, 179-200.

Cohmar Sustainable Development Council (2008). A study in personal carbon allocation: cap and share. Consultancy report prepared by AEA consulting and Cambridge Econometrics. Comhar, Dublin.

Cox, N.J. (2008). Speaking Stata: Correlation with confidence, or Fisher's z revisited. *The Stata Journal*, 8, 413-439.

Dawid, A.P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41, 1-31.

Deaton, A.S. and Irish, K. (1984). Statistical models for zero expenditures in household budgets. *Journal of Public Economics*, 23, 59-80.

DFT (Department for Transport) (2012). Table NTS0707: Adult personal car availability by ethnic group: Great Britain. Online document accessed 25.07.2013 at:
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/9972/nts0707.xls

Gibson, J. and Kim, B (2012). Testing the infrequent purchase model using direct measurement of hidden consumption from food stocks. *American Journal of Agricultural Economics*, 94, 257-270.

Kimhi, A. (1999). Double-hurdle and purchase-infrequency demand analysis: a feasible integrated approach. *European Review of Agricultural Economics* 26, 425–442.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistics Review*, 54, 139-157

Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Rosenbaum, P.R. and Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39, 33-38.

Rosenbaum, P.R. (1998). Propensity score. *Encyclopedia of Biostatistics.* Wiley: Chichester.