Quasi-Experimental Evidence on Tobacco Tax Regressivity: Additional Information for Review Purposes

Steven F. Koch

October 2017

A Pre-Match and Post-Match Balance

Underlying matching, there are really just two assumptions. For balance, we focus only on one of them. We begin by defining a binary outcome, denoted by $\mathbb{Y} = 2010$, for the 2010/11 IES, and $\mathbb{Y} = 2005$, for the 2005/06 IES. We continue by defining Z, which corresponds to observable information in the surveys, and U refers to unobserved information. One of the underlying assumptions is referred to as strong ignorability (or unconfoundedness). It assumes $\mathbb{Y} \perp\!\!\!\perp U|Z$. Following Rosenbaum & Rubin (1983), given a propensity score $e(Z_i) = \text{prob}(\mathbb{Y}_i = 2010|Z_i) = E[\mathbb{Y}_i|Z_i]$, $Z \perp\!\!\!\perp \mathbb{Y}|e(Z)$. Thus, matching on e yields distributions of Z across the two surveys that are asymptotically equivalent. However, matches are not always entirely balanced, because, in practice, we do not have the true propensity score; relatedly, matches may not be possible, given the data that is available.

Genetic matching (Diamond & Sekhon 2013) is an alternative to the standard approach. It is an iterative matching algorithm that generalizes propensity score and Mahalanobis distance matching. It makes use of an evolutionary search algorithm (Sekhon & Mebane, Jr. 1998) that optimizes balance between the observed covariates. In their Monte Carlo experiments, Diamond

& Sekhon (2013) show that it performs better than common alternatives for matching, such as binary probability propensity scores, and machine learning algorithms that have been shown to operate well for matching (Lee et al. 2010), such as boosted Classification and Regression Trees (Brieman et al. 1984, Bühlman & Yu 2003) and random forests (Brieman 2001).

In order to balance optimally, one needs an 'optimal balance' metric or loss function. The loss function used minimizes the maximum $p-$values from KS-test statistics (or paired $t-$tests in the case of discrete variables).

$$G = \sqrt{(Z_i - Z)' \, (S^{-1/2})' W S^{-1/2} \, (Z_i - Z)} \tag{1}$$

In (1), $S^{-1/2}$ is the Cholesky decomposition of the sample covariance matrix. The decomposition is most easily understood as the square root of S (although there are a few technical wrinkles underscoring its existence and its form). Optimal results are expected to be associated with the best balance that is possible, given the data used, the underlying loss function, and the variables included in Z_i in (1). In the following subsections, we provide a comparison of the outcomes for the reviewers.

A.1 All Households 2005/06 and 2010/11

We begin by considering all households in the dataset. Technically, given that we held household head race, gender and urban/rural locale fixed, and matched within those cells, we do not actually include all households. In particular, there are two few Asian households in the dataset to allow for them to be included, while rural households headed by white men or women are also few; the same can be said for rural households headed by coloured women.

The first table, Table A.1, provides evidence of the poor matches across surveys, generally.

Table A.1: All Household Descriptive Statistics Before Matching

Variable	2005/06 Mean	2010/11 Mean	Pr(> |t|) $t-$test	Pr(> |ks|) ks−test
HH Head Age Group	10.234	10.222	0.699	0.002
HH Head Schooling	1.809	1.561	0.000	0.000
Real Income	7194.220	5773.147	0.000	0.000
Real Expenditure	5723.574	4589.567	0.000	0.000
HH Male Ratio	0.465	0.462	0.271	0.396
HH Adult Ratio	0.773	0.756	0.000	0.000
Girls (0-4) in HH	0.198	0.201	0.473	0.764
Boys (0-4) in HH	0.201	0.206	0.242	0.246
Girls (5-14) in HH	0.376	0.452	0.000	0.000
Boys (5-14) in HH	0.380	0.457	0.000	0.000
Men (15-64) in HH	1.070	1.091	0.019	0.053
Women (15-64) in HH	1.192	1.243	0.000	0.003
Men (65+) in HH	0.123	0.129	0.070	0.058
Women (65+) in HH	0.209	0.208	0.806	0.841
Eastern Cape	0.117	0.114	0.248	
Western Cape	0.132	0.134	0.465	
Northern Cape	0.048	0.082	0.000	
Free State	0.086	0.083	0.242	
Kwa-Zulu Natal	0.143	0.223	0.000	
Northwest Province	0.100	0.074	0.000	
Gauteng Province	0.153	0.117	0.000	
Mpumulanga Province	0.091	0.080	0.000	
Observations	24974	20898		

Means for 2005/06 and 2010/11 Income and Expenditure Surveys along with $t-$tests of mean differences and ks−tests for distributional differences. Note: for discrete variables, no KS-test has been performed, which is why those cells are blank.

The second table, Table A.2, provides evidence that the matches have improved. However, the improvement is subtle. Roughly, the same number of variables had statistically significant differences across the surveys before and after matching. However, many of the mean differences, for example, lessened. Furthermore, it should also be kept in mind that in a sample of 40 000 observations, statistically significant differences can arise from rather small mean differences and distributional differences.

Table A.2: All Household Descriptive Statistics After Matching

	2005/06	2010/11	Pr(> |t|)	Pr(> |ks|)
Variable	Mean	Mean	t−test	ks−test
Propensity Score	0.469	0.469	0.751	0.998
HH Head Age Group	10.320	10.222	0.002	0.000
HH Head Schooling	1.519	1.535	0.144	0.113
Real Income	5397.684	5420.685	0.808	0.122
Real Expenditure	4283.081	4303.439	0.772	0.022
HH Male Ratio	0.462	0.461	0.630	0.008
HH Adult Ratio	0.761	0.754	0.008	0.000
Girls (0-4) in HH	0.192	0.205	0.007	0.052
Boys (0-4) in HH	0.195	0.208	0.004	0.002
Girls (5-14) in HH	0.444	0.459	0.052	0.128
Boys (5-14) in HH	0.438	0.462	0.001	0.013
Men (15-64) in HH	1.037	1.089	0.000	0.000
Women (15-64) in HH	1.190	1.245	0.000	0.000
Men (65+) in HH	0.130	0.127	0.351	0.357
Women (65+) in HH	0.214	0.208	0.153	0.140
Eastern Cape	0.116	0.109	0.019	
Western Cape	0.172	0.136	0.000	
Northern Cape	0.064	0.081	0.000	
Free State	0.093	0.084	0.001	

... continued on next page ...

Variable	2005/06 Mean	2010/11 Mean	Pr(> |t|) $t-$test	Pr(> |ks|) ks−test
Kwa-Zulu Natal	0.146	0.219	0.000	
Northwest Province	0.098	0.076	0.000	
Gauteng Province	0.116	0.118	0.566	
Mpumulanga Province	0.080	0.082	0.584	
Observations	20126	20126		

Means for 2005/06 and 2010/11 Income and Expenditure Surveys along with $t-$tests of mean differences and ks−tests for distributional differences. Note: for discrete variables, no KS-test has been performed, which is why those cells are blank.

A.2 Tobacco Consuming Households 2005/06 and 2010/11

The second set of tables focus on tobacco consuming households matched in the same way as described above, and in the main text. Those match balance test statistics are reported in Tables A.3 and A.4.

Table A.3: Tobacco Consuming Household Descriptive Statistics Before Matching

Variable	2005/06 Mean	2010/11 Mean	Pr(> |t|) $t-$test	Pr(> |ks|) ks−test
HH Head Age Group	10.207	10.420	0.000	0.001
HH Head Schooling	1.703	1.381	0.000	0.000
Real Income	7134.495	5182.330	0.000	0.000
Real Expenditure	5648.117	4092.177	0.000	0.000
HH Male Ratio	0.567	0.538	0.000	0.000

... continued on next page ...

Variable	2005/06 Mean	2010/11 Mean	Pr(> |t|) $t-$test	Pr(> |ks|) ks−test
HH Adult Ratio	0.818	0.782	0.000	0.000
Girls (0-4) in HH	0.170	0.194	0.002	0.008
Boys (0-4) in HH	0.170	0.192	0.003	0.000
Girls (5-14) in HH	0.302	0.414	0.000	0.000
Boys (5-14) in HH	0.314	0.426	0.000	0.000
Men (15-64) in HH	1.272	1.294	0.162	0.357
Women (15-64) in HH	1.065	1.188	0.000	0.000
Men (65+) in HH	0.141	0.148	0.277	0.256
Women (65+) in HH	0.185	0.210	0.000	0.001
Eastern Cape	0.205	0.148	0.000	
Western Cape	0.112	0.139	0.000	
Northern Cape	0.071	0.117	0.000	
Free State	0.134	0.111	0.000	
Kwa-Zulu Natal	0.091	0.156	0.000	
Northwest Province	0.099	0.084	0.003	
Gauteng Province	0.144	0.114	0.000	
Mpumulanga Province	0.078	0.066	0.005	
Observations	6032	8110		

Means for 2005/06 and 2010/11 Income and Expenditure Surveys along with $t-$tests of mean differences and ks−tests for distributional differences. Note: for discrete variables, no KS-test has been performed, which is why those cells are blank.

Before matching, we see that there are many statistically significant differences. All but two variables have statistically significant different means; those same two are the only two that appear (statistically) to have come from the same distributions. In other words, 20 variables have

statistically significant differences in means, while 12 have statistically significant differences in distributions. Following the matching algorithm, only 10 variables have mean differences large enough to matter, and only five have distribution differences large enough to matter. Furthermore, in most of the remaining variables, where differences matter, the mean differences have become smaller, as one would hope. Lastly, it should not be forgotten that there are nearly 4000 matched pairs; reasonably small differences can still be statistically significant in a sample of that size.

Table A.4: Tobacco Consuming Household Descriptive Statistics After Matching

	2005/06	2010/11	Pr(> |t|)	Pr(> |ks|)
Variable	Mean	Mean	$t-$test	ks−test
Propensity Score	0.590	0.591	0.626	0.823
HH Head Age Group	10.472	10.435	0.462	0.000
HH Head Schooling	1.365	1.358	0.660	0.010
Real Income	4861.899	4932.619	0.603	0.309
Real Expenditure	3853.554	3899.656	0.651	0.006
HH Male Ratio	0.545	0.540	0.202	0.123
HH Adult Ratio	0.787	0.782	0.151	0.229
Girls (0-4) in HH	0.200	0.197	0.667	0.136
Boys (0-4) in HH	0.175	0.194	0.008	0.004
Girls (5-14) in HH	0.403	0.417	0.220	0.262
Boys (5-14) in HH	0.386	0.427	0.000	0.005
Men (15-64) in HH	1.259	1.295	0.019	0.075
Women (15-64) in HH	1.126	1.183	0.001	0.066
Men (65+) in HH	0.143	0.147	0.441	0.375
Women (65+) in HH	0.202	0.211	0.159	0.060
Eastern Cape	0.164	0.142	0.000	
Western Cape	0.140	0.141	0.765	
Northern Cape	0.088	0.118	0.000	
Free State	0.152	0.113	0.000	
Kwa-Zulu Natal	0.097	0.150	0.000	

... continued on next page ...

Variable	2005/06 Mean	2010/11 Mean	Pr(> |t|) $t-$test	Pr(> |ks|) ks−test
Northwest Province	0.117	0.086	0.000	
Gauteng Province	0.106	0.115	0.055	
Mpumulanga Province	0.079	0.068	0.005	
Observations	7806	7806		

Means for 2005/06 and 2010/11 Income and Expenditure Surveys along with $t-$tests of mean differences and ks−tests for distributional differences. Note: for discrete variables, no KS-test has been performed, which is why those cells are blank.

A.3 Cigarette Consuming Households 2005/06 and 2010/11

In the final comparison, we consider cigarette consuming households. The match-balance statistics for this subset of households are contained in Tables A.5 and A.5.

Table A.5: Cigarette Consuming Household Descriptive Statistics Before Matching

Variable	2005/06 Mean	2010/11 Mean	Pr(> |t|) $t-$test	Pr(> |ks|) ks−test
HH Head Age Group	9.788	9.834	0.448	0.036
HH Head Schooling	1.941	1.687	0.000	0.000
Real Income	8730.060	6804.716	0.000	0.000
Real Expenditure	6818.902	5318.795	0.000	0.000
HH Male Ratio	0.590	0.579	0.073	0.003
HH Adult Ratio	0.830	0.801	0.000	0.000
Girls (0-4) in HH	0.161	0.172	0.206	0.235

... continued on next page ...

Variable	2005/06 Mean	2010/11 Mean	Pr(> |t|) $t-$test	Pr(> |ks|) ks−test
Boys (0-4) in HH	0.160	0.176	0.068	0.055
Girls (5-14) in HH	0.269	0.372	0.000	0.000
Boys (5-14) in HH	0.282	0.387	0.000	0.000
Men (15-64) in HH	1.348	1.402	0.006	0.113
Women (15-64) in HH	1.058	1.159	0.000	0.000
Men (65+) in HH	0.118	0.122	0.535	0.504
Women (65+) in HH	0.151	0.160	0.242	0.140
Eastern Cape	0.251	0.180	0.000	
Western Cape	0.089	0.123	0.000	
Northern Cape	0.054	0.099	0.000	
Free State	0.108	0.094	0.021	
Kwa-Zulu Natal	0.105	0.153	0.000	
Northwest Province	0.081	0.073	0.188	
Gauteng Province	0.179	0.151	0.000	
Mpumulanga Province	0.081	0.070	0.044	
Observations	4414	5147		

Means for 2005/06 and 2010/11 Income and Expenditure Surveys along with $t-$tests of mean differences and ks−tests for distributional differences. Note: for discrete variables, no KS-test has been performed, which is why those cells are blank.

Initially, before matching, 15 variables had statistically significant mean differences, seven remained, following the matching execise. With respect to distributional differences, eight were statistically significant before matching, while three remained after matching.

Table A.6: Cigarette Consuming Household Descriptive Statistics After Matching

Variable	2005/06 Mean	2010/11 Mean	Pr(>|t|) t−test	Pr(>|ks|) ks−test
Propensity Score	0.556	0.557	0.570	0.592
HH Head Age Group	9.881	9.819	0.309	0.001
HH Head Schooling	1.648	1.659	0.582	0.492
Real Income	6396.054	6463.292	0.741	0.685
Real Expenditure	5014.314	5052.121	0.806	0.015
HH Male Ratio	0.587	0.583	0.494	0.698
HH Adult Ratio	0.802	0.801	0.808	0.030
Girls (0-4) in HH	0.174	0.175	0.963	0.627
Boys (0-4) in HH	0.172	0.178	0.507	0.314
Girls (5-14) in HH	0.348	0.375	0.044	0.259
Boys (5-14) in HH	0.361	0.386	0.072	0.349
Men (15-64) in HH	1.371	1.406	0.067	0.148
Women (15-64) in HH	1.093	1.150	0.006	0.078
Men (65+) in HH	0.116	0.120	0.502	0.479
Women (65+) in HH	0.154	0.159	0.497	0.617
Eastern Cape	0.209	0.179	0.000	
Western Cape	0.116	0.125	0.145	
Northern Cape	0.072	0.100	0.000	
Free State	0.123	0.096	0.000	
Kwa-Zulu Natal	0.108	0.141	0.000	
Northwest Province	0.089	0.075	0.011	
Gauteng Province	0.147	0.153	0.445	
Mpumulanga Province	0.083	0.072	0.059	
Observations	4909	4909		

Means for 2005/06 and 2010/11 Income and Expenditure Surveys along with t−tests of mean differences and ks−tests for distributional differences. Note: for discrete variables, no KS-test has been performed, which is why those cells are blank.

Although we see that matching is not perfect, we do see improvements in the balance of the data after matching, and this is especially clear for the tobacco and cigarette consuming households, who appear to more easily matched across suveys than the broad population.

B Income and Expenditure Densities

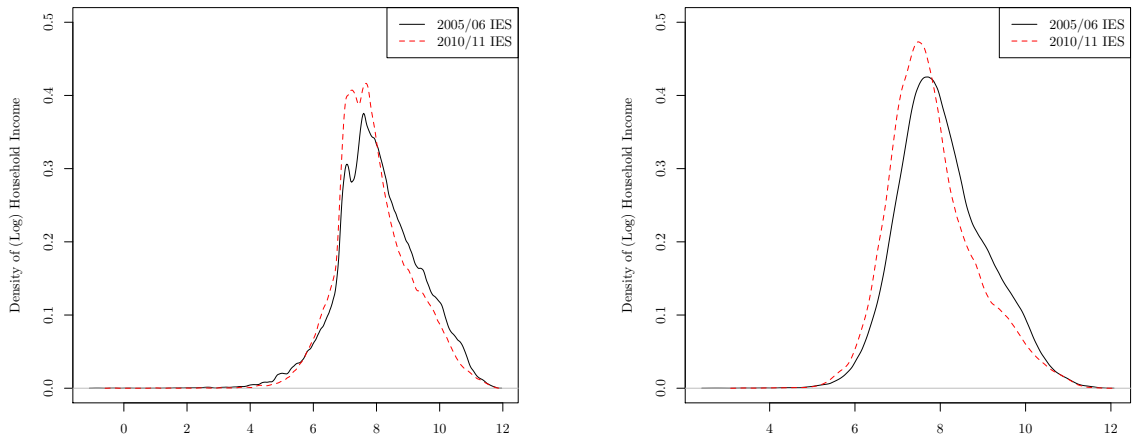
The improvement in match-balance is an important feature of a successful matching algorithm. However, because the analysis focuses on regressivity, which is always compared to living standards, it is especially important for the living standards of households be comparable across the surveys. Initially, we adjusted household income and household expenditure to match in real terms, which means they were adjusted to 2008, the last time the base was recalculated.[1]

B.1 Before Matching

Once the real adjustments were made, we pooled the data together, to see how different the underlying (real) income and expenditure data differed. Although we have estimates of the differences, based on ks−tests in the previous discussion, we felt it was useful to see the improvement in the distributions that arises from matching.

Initially, we present the densities for unmatched data, and across all three subsamples. We begin with all households. These results are illustrated in three sets of two panels. The first set is contained in Figure B.1. These illustrate the densities of (Log) of (Real) Household Income/Expenditure for all households in the two surveys. Thus, the x−axis, not labelled, is either (Log) of (Real) Household Income or Expenditure.

[1] Although the IES was done in 2005/06, it took time for the data to be analysed, and for a new base to be created and put in place.



(a) Log of Real Income

(b) Log of Real Expenditure

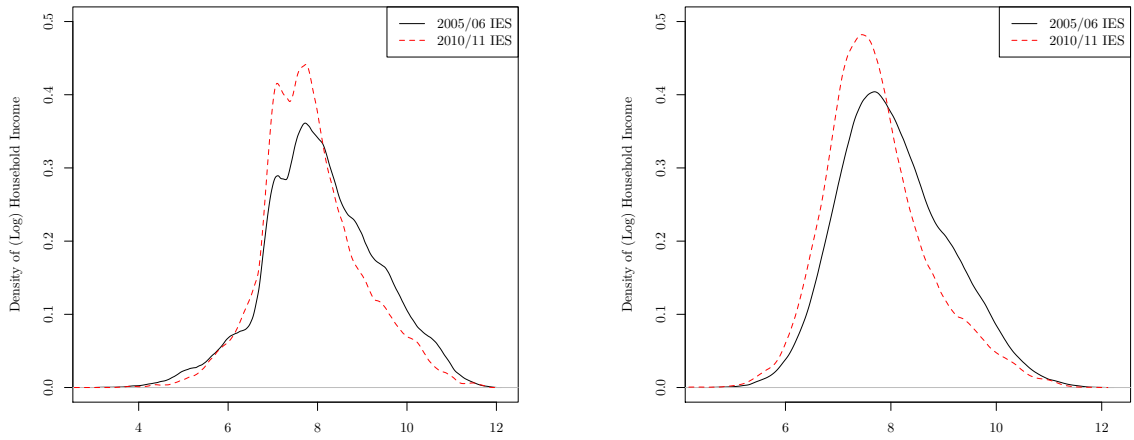
Figure B.1: Densities of the log of real household income, Panel B.1a, for tobacco consuming households, along with the log of real household expenditure, Panel B.1b, before matching.

There are interesting differences between the initial 2005/06 and follow-up 2010/11 IES income and expenditure data. Roughly speaking households in 2010/11 are poorer in real terms than they were in 2005/06. Plausibly, this could be a result of the 2008/09 global financial crisis, which negatively affected economic growth in South Africa. It is also possible that this shift represents a change in accessibility of households. Gated communities have become increasingly common; houses in those communities are likely to be owned by better-off households. Given that access to these communities is controlled, it is possible that surveyors have had a more difficult time accessing those households.

Those differences carryover into the tobacco consuming subset, which are illustrated in the two panels of Figure B.2, as well as those illustrated in the two panels of Figure B.3.

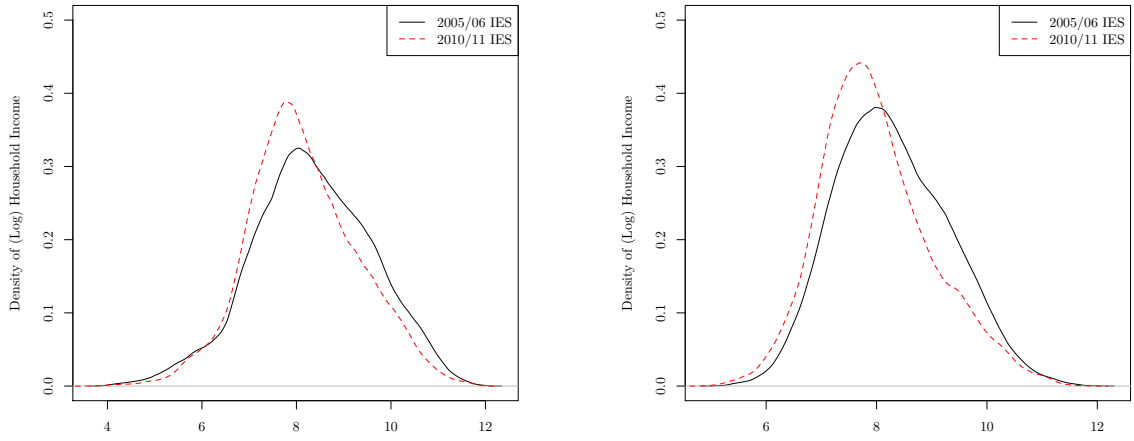
B.2 After Matching

We repeat the density estimate exercise following matching. The results are contained, again, in three sets of two-panel figures. Those are Figures B.4, B.5 and B.6. The effect of matching, as expected, given the results in the previous tables, is that income and expenditure are now much more comparable across all of the subsamples of matched data.



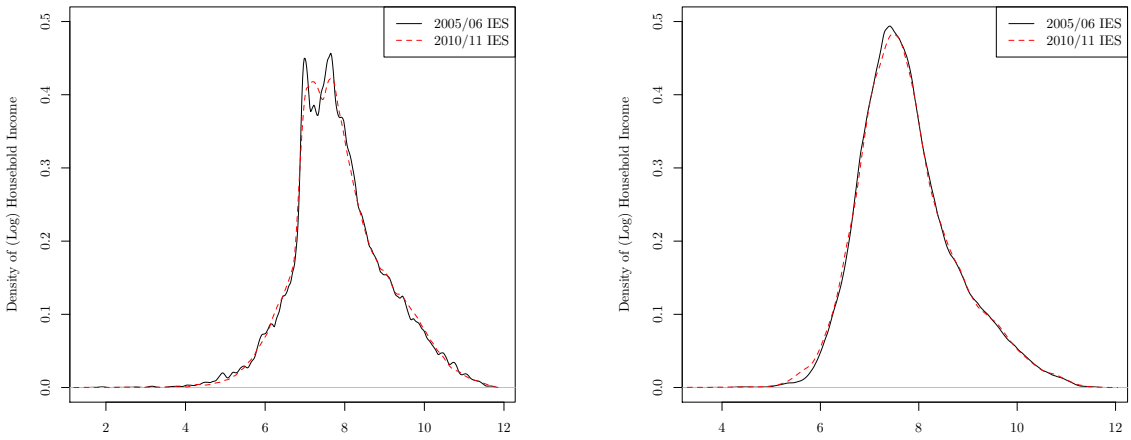
(a) Log of Real Income (b) Log Real Expenditure

Figure B.2: Densities of the log of real household income, Panel B.2a, for tobacco consuming households, along with the log of real household expenditure, Panel B.2b, before matching.



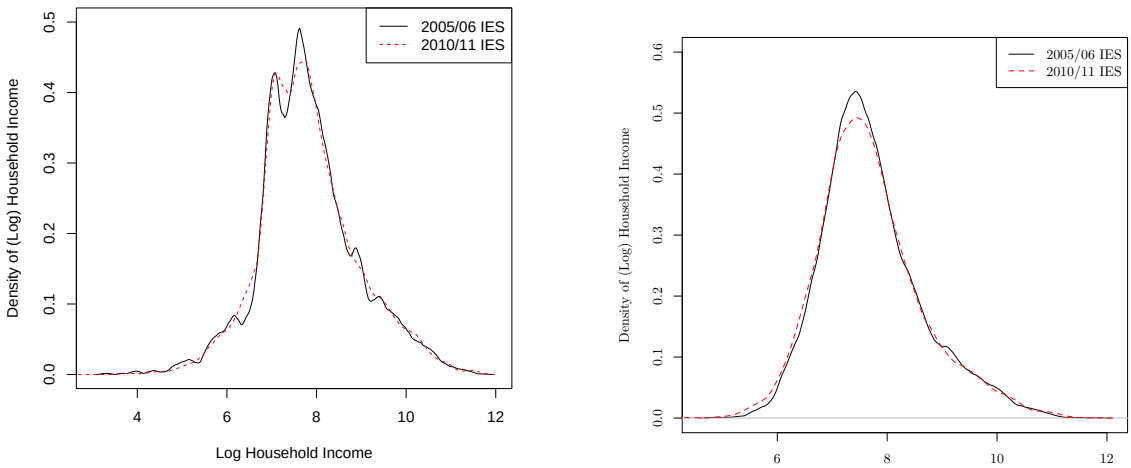
(a) Log of Real Income (b) Log Real Expenditure

Figure B.3: Densities of the log of real household income, Panel B.3a, for cigarette consuming households, along with the log of real household expenditure, Panel B.3b, before matching.



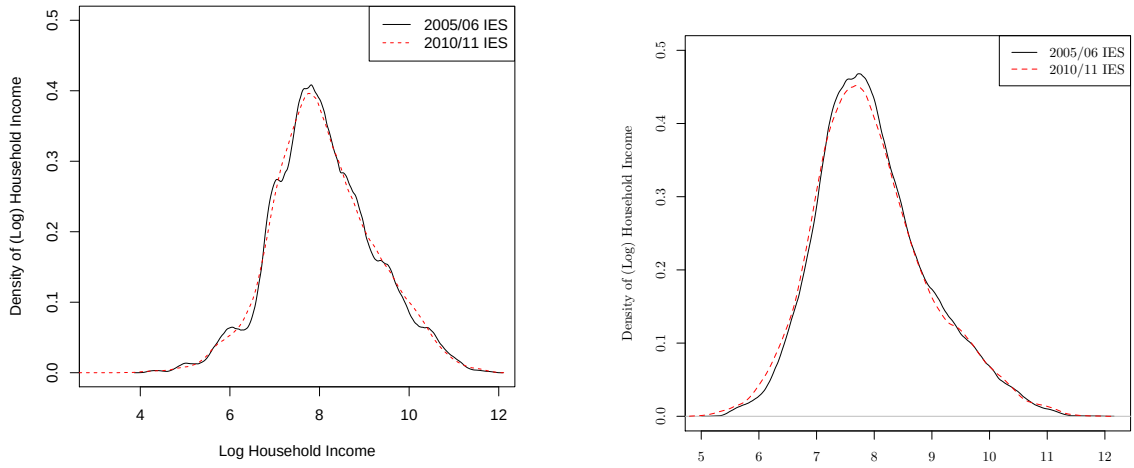
(a) Log of Real Income (b) Log of Real Expenditure

Figure B.4: Densities of the log of real household income, Panel B.4a, for tobacco consuming households, along with the log of real household expenditure, Panel B.4b, after matching.



(a) Log of Real Income (b) Log Real Expenditure

Figure B.5: Densities of the log of real household income, Panel B.5a, for tobacco consuming households, along with the log of real household expenditure, Panel B.5b, after matching.



(a) Log of Real Income

(b) Log Real Expenditure

Figure B.6: Densities of the log of real household income, Panel B.6a, for cigarette consuming households, along with the log of real household expenditure, Panel B.6b, after matching.

References

Brieman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Brieman, L., Friedman, J., Stone, C. J. & Olshen, R. A. (1984), *Classification and Regression Trees*, Chapman & Hall, New York.

Bühlman, P. & Yu, B. (2003), 'Boosting with the l_2−loss: Regression and classification', *Journal of the American Statistical Association* **98**(462), 324–339.

Diamond, A. & Sekhon, J. S. (2013), 'Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies', *The Review of Economics and Statistics* **95**(3), 932–945.

Lee, B., Lessler, J. & Stuart, E. A. (2010), 'Improving propensity score weighting using machine learning', *Statistics in Medicine* **29**(3), 337–346.

Rosenbaum, P. R. & Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**(1), 41–55.

Sekhon, J. S. & Mebane, Jr., W. R. (1998), 'Genetic optimization using derivatives: theory and application to nonlinear models', *Political Analysis* **7**, 189–203.