The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

CESifo Working Papers
www.cesifo.org/wp

Non-Parametric Preprocessing for the Estimation of Equivalence Scales

Christian Dudel
Jan Marvin Garbuszus
Notburga Ott
Martin Werding

CESIFO WORKING PAPER NO. 5103
CATEGORY 12: EMPIRICAL AND THEORETICAL METHODS
DECEMBER 2014

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 5103*

Non-Parametric Preprocessing for the Estimation of Equivalence Scales

Abstract

Empirically analyzing household behavior usually relies on informal data preprocessing. That is, before an econometric model is estimated, observations are selected in such a way that the resulting subset of data can be assumed to be sufficiently homogeneous with respect to the specific research question pursued. For example, households with members above retirement age may be excluded where it seems important that they differ from other households with respect to time use and home production. We propose the use of matching techniques and balance checking at this initial stage. This can be interpreted as a non-parametric approach to preprocessing data and as a way to formalize informal procedures. To illustrate this, we use German micro-data on household expenditure to estimate equivalence scales as a specific example. Our results show that matching leads to results which are more stable with respect to model specification and that this type of formal preprocessing is especially useful if one is mainly interested in results for specific subgroups, such as low-income households.

JEL-Code: C180, D100, D120.

Keywords: equivalence scales, matching, balancing, balance checking, non-parametric preprocessing, household expenditure, household behavior.

Christian Dudel
Ruhr University Bochum / Germany
christian.dudel@ruhr-uni-bochum.de

Jan Marvin Garbuszus
Ruhr University Bochum / Germany
jan.garbuszus@ruhr-uni-bochum.de

Notburga Ott
Ruhr University Bochum / Germany
notburga.ott@ruhr-uni-bochum.de

Martin Werding
Ruhr University Bochum / Germany
martin.werding@ruhr-uni-bochum.de

November 2014

1 Introduction

Equivalence scales are routinely applied in research on inequality and poverty. They are used to adjust household income of households that differ in size and composition to make them comparable in terms of their welfare level. A popular example is the so-called modified OECD scale (first proposed by Hagenaars et al, 1994). A single adult is taken as a reference household with scale weight 1.0. Each additional adult increases the scale weight by 0.5, and each child below age 15 increases it by 0.3. As a result, a couple with one child has a total scale weight of 1.8. This means that the households needs 1.8 times as much income to reach the same welfare level as the reference household. Apart from use in scientific research, equivalence scales are also used for the design of policies. Examples are welfare benefits in Germany (Dudel et al, 2013) or old-age pensions in Great Britain (Stewart, 2009).

Because of their significance for research and applied work, equivalence scales have long been of interest to economists, and many different methods for estimating them have been devised (for an overview, see Coulter et al, 1992). Results heavily depend on the choice of models and model specifications, as shown by wide ranges of results derived from the same data or methods (see, e.g., Buhmann et al, 1988; Bellemare et al, 2002; Schröder, 2009). Even if a certain model is deemed most appropriate (or is the only one which can be estimated with a given data set), adding just one covariate may substantially alter the results. Furthermore, equivalence scales are often estimated based on data for rather different groups of households, e.g., households with and without children. If these groups are not fully comparable with respect to covariates included in the model, results may heavily depend on hidden extrapolations that are not based on observed data (King and Zeng, 2006).

A related problem is the estimation of equivalence scales for specific subgroups of the population, e.g., households that differ in size and fall into the lowest income quintile. Estimating differentiated scale weights for subgroups like these can be considered important due to heterogeneity in relevant household behavior. Taking into account all households that are available increases sample size and thus efficiency, but it implies the assumption that model parameters are identical for households across all welfare levels. This assumption is highly questionable if household behavior is informally observed to differ considerably. At the same time, it is far from clear which subgroups of households could be taken to be sufficiently homogeneous, if their comparability is to be established w.r.t. to income. This would require equivalence scales to be known *a priori*, which renders the whole task of estimating equivalence scales a circular process.

For instance, assume that scale weights are needed to advise policies aimed at children

in low-income households, relying on a large dataset and some method for estimating equivalence scales. As a starting point, the lowest income quintile, i.e., the poorest 20% of couples with one child shall be used. Households included in this subgroup are to be compared to childless couples as reference households. Now the question arises which of the childless couples should be included in the analysis. If the poorest 20% of childless couples were used, this would stipulate their welfare level to be comparable to those of the poorest 20% of couples with one child. If they are in fact better off (or worse off), results will again heavily depend on hidden extrapolations and may be severely biased.

To solve this problem, the paper builds on recent literature on the estimation of treatment effects using the so called potential outcomes framework. More specifically, we propose a two-step procedure that includes some matching procedure in the first step and standard approaches to estimating equivalence scales in the second step. In doing so, we follow Szulc (2009, 2011) and include welfare indicators in the matching step. Matching can be seen as non-parametric preprocessing (NPP), and we will argue that much of the empirical literature on household behavior effectively uses informal preprocessing (IP) of data before applying more elaborate methods, where IP serves the same goals as NPP. The approach proposed here can therefore be seen as a more formal way of implementing data preprocessing which is conventionally used elsewhere.

The remainder of this paper is organized as follows. Section 2 explains the general reasoning behind the combination of matching and regression models which can be found in the literature on estimating treatment effects. Its application to estimating equivalence scales compared to informal preprocessing is described in section 3. Results of an empirical application using German micro-data are provided in section 4. Section 5 concludes.

2 Combining matching and regression techniques

2.1 Potential outcomes framework

Starting with work by Rubin (1973, 1979), there is now a large literature on the combination of matching and regression models (Rosenbaum and Rubin, 1984; Rubin and Thomas, 2000; Imai and Van Dyk, 2004; King and Zeng, 2006; Ho et al, 2007; Morgan and Winship, 2010; Iacus et al, 2011; Abadie and Imbens, 2011; Iacus et al, 2012; Hainmueller, 2012). A common starting point is the potential outcomes framework, also called Rubin causal model, Neyman-Rubin causal model, or Roy causal model (see Holland, 1986; Sekhon, 2008; Heckman, 2008).

The basic reasoning stems from, and uses the language of, experimental designs. Given a sample of n units, let D denote a binary indicator which equals 1 if unit $i = 1, \ldots, n$ received a certain treatment. Otherwise D equals 0. Units who received the treatment

form the "treatment group" and the remaining units are called "control group". The effect of treatment on some outcome Y is to be estimated.

Specifically, $y_i(1)$ denotes the outcome which would be observed if unit i receives the treatment and $y_i(0)$ denotes the outcome without treatment. If both values were known, the average treatment effect (ATE) could simply be determined as expectation of the difference $y_i(1) - y_i(0)$, $\delta = \mathrm{E}(Y(1) - Y(0))$. Although it is assumed that both potential outcomes exist for all units i, either $y_i(0)$ or $y_i(1)$ can be observed in practice and never both. Essentially, this amounts to a missing data problem with either $y_i(0)$ or $y_i(1)$ missing for each i.

In case of a random assignment of units to treatment and control group, $Y(1)$ and $Y(0)$ are independent of D,

$$Y(0), Y(1) \perp D. \tag{1}$$

This assumption implies that $\mathrm{E}(Y(1)|D = 1) = \mathrm{E}(Y(1))$ and $\mathrm{E}(Y(0)|D = 0) = \mathrm{E}(Y(0))$, so that δ can be determined from

$$\mathrm{E}(Y(1)|D = 1) - \mathrm{E}(Y(0)|D = 0). \tag{2}$$

That is, the difference in means between treatment and control group can be used to estimate the effect of the treatment.

In empirical studies, randomization is often not possible and the assumption introduced above seems dubious because other variables may influence both the outcome and selection into treatment. Instead, a conditional variant is assumed to hold (e.g., Rosenbaum and Rubin, 1983):

$$Y(0), Y(1) \perp D|X, \tag{3}$$

where X is a vector of covariates. This so-called "unconfoundedness" assumption leads to conditional variants of the results above, i.e., $\mathrm{E}(Y(1)|D = 1, X) = \mathrm{E}(Y(1)|X)$ and $\mathrm{E}(Y(0)|D = 0, X) = \mathrm{E}(Y(0)|X)$. This means that conditional on covariates X, there is no selection into treatment (for another interpretation, see Morgan and Harding, 2006). Thus, treatment group and control group are not directly comparable, but unconfoundedness implies that observations with the same characteristics X would be.

Now, δ can be determined from

$$\delta = \mathrm{E}(\mathrm{E}(Y(1)|X) - \mathrm{E}(Y(0)|X)), \tag{4}$$

if the additional assumption $0 < \Pr(D = 1|X) < 1$ is fulfilled, ensuring that conditional expectations are defined for all X. If both assumptions hold, treatment assignment is called strongly ignorable (Imbens, 2004). A third assumption which is required for observational studies as well as for experimental studies is called "stable-unit-treatment-value" assumption. It requires outcomes for observation i to be independent of a treatment of unit j.

2.2 Matching estimators

Given the conditions introduced above the difference in means (2) is not a valid estimator because treatment and control group may differ with respect to X. Yet, if $\Pr(X|D = 1) = \Pr(X|D = 0) = \Pr(X)$, treatment and control group would be called "balanced" with respect to the relevant covariates, and (2) would still lead to valid estimation of δ because it effectively equals (4). This is what is achieved through matching techniques of which many different variants have been proposed in the literature (for an overview, see Imbens, 2004; Imbens and Wooldridge, 2009; Stuart, 2010). In what follows, nearest neighbor matching will be described.

A simple variant works as follows. For each unit $y_i = d_i y_i(1) + (1 - d_i)y_i(0)$, d_i and x_i are observed; δ is estimated through

$$\hat{\delta} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i(1) - \hat{y}_i(0), \tag{5}$$

where

$$\hat{y}_i(1) = \begin{cases} y_i^m & \text{if} \quad d_i = 0 \\ y_i & \text{if} \quad d_i = 1 \end{cases} \tag{6}$$

and

$$\hat{y}_i(0) = \begin{cases} y_i & \text{if} \quad d_i = 0 \\ y_i^m & \text{if} \quad d_i = 1. \end{cases} \tag{7}$$

Missing information on y_i^m is found via matching. If unit i received the treatment ($D = 1$), y_i^m is based on one or more observations for which $D = 0$ and $X = x_i$. If unit i belongs to the control group ($D = 0$), units for which $D = 1$ are used to determine y_i^m instead. That is, units with identical values of covariates are "matched" and used to impute the missing data. In a sense, each observation then has two results for the outcome variable, one under treatment and one for being in the control group. This, in turn, leads to balanced treatment and control groups.

In practice it will usually not be possible to perform exact matching as outlined above, because it will not be possible to find observations which are perfectly identical with respect to X. At least, this is less and less likely to happen if X includes continuous variables and the number of covariates included in X is large. Therefore, some distance function $d(x_i, x_j)$ can be used which assesses how similar observations are to each other (for examples and discussions of possible distance functions, see Zhao 2004; Dettmann et al 2011). Then, y_i^m is found by matching the most similar observations. This does not produce a fully balanced data set, though, and differences between treatment and control group may still lead to biased estimates of δ.

Matching is thus not guaranteed to achieve balanced treatment and control groups. Because of this, one has to carefully check balancing, i.e., compare the distributions of X in both treatment group and control group before and after matching. This may lead to dropping observations for which no reasonable match can be found. If observations are dropped, this affects the final result of the estimation. For example, assume that the treatment group includes individuals with values of a continuous covariate X between g and g', with $g < g'$. In the control group, the covariate ranges from h to h', with $h < g < h' < g'$. Treatment and control group overlap in the region $g < X < h'$. For observations in the treatment group with $X > h'$, no comparable observations can be found in the control group. The same holds for control units with $X < g$. One can then restrict attention to the region of common support, $g < X < h'$, or a slightly larger version, $g - c < X < h' + c$, where c is some constant which is small compared to g and h'. Proceeding in such a fashion will recover $\mathrm{E}(Y(1) - Y(0)|g < X < h')$ but not $\mathrm{E}(Y(1) - Y(0))$.

2.3 Combining matching and regressions

Another possibility for dealing with the problem of comparability is the use of regression techniques. In this case, $\mathrm{E}(Y|D, X)$ is usually estimated parametrically introducing D as a dummy variable. To recover an unbiased estimate of δ, the model has to be specified appropriately. Moreover, balancing is still of importance. To fix ideas, suppose that X is a continuous variable. Let g be some constant for which $\Pr(X > g|D = 1) = 0$ and $\Pr(X > g|D = 0) = p$. Estimating a simple regression of the form

$$Y = a + b_X X + b_D D + \epsilon \tag{8}$$

can be seen as imputation of $\hat{y}_i(1)$ and $\hat{y}_i(0)$ for each unit i, and the treatment effect which equals b_D is simply given by (5) (Imbens, 2004). Imputed values $\hat{y}_i(1)$ for observations with $X > g$ and $D = 0$ will be based on extrapolation and lack an empirical basis. This

is essentially the overlap problem discussed in the preceding subsection.

In addition, depending on the functional form of the relationship between Y and X, estimates for b_D and b_Z can heavily depend on observations for which $X > g$. For example, assume that for $X < g$ the relationship is $Y = a + b_X X + b_D D + \epsilon$ and for $X > g$ it is $Y = a + b'_X X + b_D D + \epsilon$, with $b'_X \neq b_X$. Using the correct specification

$$Y = a + b_X X \mathrm{I}(X < g) + b'_X X \mathrm{I}(X > g) + b_D D + \epsilon, \tag{9}$$

where $\mathrm{I}(\cdot)$ denotes an indicator function, would still lead to unbiased estimates, while using

$$Y = a + b^*_X X + b_D D + \epsilon \tag{10}$$

could lead to biased estimates, depending on the proportion of observations with $X > g$ and the difference between b'_X and b_X.

A combination of matching and regression techniques allows to deal with the drawbacks of each method. Matching and checking balance can be used to achieve balanced treatment and control groups, which reduces hidden extrapolations and the sensitivity of regression estimates with respect to model specification. In turn, applying regression methods after matching controls for remaining differences in balancing and yields more reliable estimates than the simple difference in means in equation (5). In the literature, combinations of both methods have been suggested from two perspectives, both of which will be discussed in the next section. Rubin (1973, 1979) and Abadie and Imbens (2011) stress the use of regression as a technique for correcting biases in matching, whereas King and Zeng (2006), Ho et al (2007) and Iacus et al (2011, 2012) describe matching methods and balance checking as non-parametric preprocessing of data to reduce the sensitivity of results with respect to model specification.

Informally, sensitivity with respect to model specification can be understood in the following sense. Let $\mathcal{X}$ be the set of all elements of X (i.e., $\{X_1, X_2, \dots\}$), where some X_j could represent transformations of other variables (e.g., X^2) or interactions between other variables (e.g., XX'). Let $\mathcal{M}$ be the power set of $\mathcal{X}$ or some subset thereof, i.e., $\mathcal{M} = \mathcal{P}(\mathcal{X})$ or $\mathcal{M} \subset \mathcal{P}(\mathcal{X})$; $\mathcal{M}^*$ includes all elements of $\mathcal{M}$ combined with D, i.e., $\mathcal{M}^* = \{m_k \cup \{D\} | m_k \in \mathcal{M}\}$. Let $\mathcal{D} = \{d_1, d_2, \dots\}$ be the set of estimates of b_D where d_k denotes the result for b_D if variables m_k and D are included in the regression. Sensitivity to model specification can then be expressed as some measure of spread of the elements of $\mathcal{D}$, e.g., the variance $\mathrm{Var}(d_k \in \mathcal{D})$ or the range $\max \mathcal{D} - \min \mathcal{D}$.

Against this background, the statement that matching and balance checking reduce

sensitivity means that, if $\mathcal{D}$ denotes the results for b_D without matching and $\mathcal{D}_M$ the results after matching and balance checking, then $S(\mathcal{D}_M) < S(\mathcal{D})$. King and Zeng (2006) provide several examples where non-balanced data lead to higher sensitivity of results with respect to specification than balanced data. Empirical illustrations regarding the estimation of equivalence scales will be given in section 4.

3 Matching and equivalence scales

3.1 Equivalence scales

In the literature, several types of equivalence scales are effectively being considered. What we have in mind as a fruitful application of our approach are equivalence scales based on econometric analyses of observed household behavior, mainly consumption (see, e.g., Pollak and Wales, 1978; Deaton and Muellbauer, 1980; Blundell et al, 2003, for influential contributions and a recent application). Alternatives that are also relevant in current debates are empirical scales based on subjective perceptions (see, e.g., Kapteyn and Van Praag, 1975; Koulovatianos et al, 2005) to which part of our reasoning may be applicable as well, or normative scales that are typically based on pre-defined baskets of goods (and, ultimately, on experts' or politicians' choice) and are lacking an elaborate empirical basis. Here, we will focus entirely on scales of the first type to exemplify our ideas.

To this end, let $Q(P, Y, X)$ denote the demand function for a household with income Y and socio-demographic characteristics X who is facing prices P. $U(Q, X)$ denotes utility derived from demand, and total expenditure is given by QP. In what follows, we will assume that $QP = Y$. A cost function can then be defined as $C(P, X, U) = \min_Q[QP|U(Q, X) \geq U]$. $\mathcal{H} = \{r, 1, \ldots, h, \ldots\}$ is a set of household types with a reference type r. An equivalence scale is then defined as a set $\mathcal{A} = \{A_r, A_1, \ldots, A_h, \ldots\}$ in which particular elements A_h are given by the equivalence scale function

$$A_h(P, U) = A(P, U, X_h) = \frac{C(P, X_h, U)}{C(P, X_r, U)}. \tag{11}$$

Elements A_h are called scale weights and can be interpreted as discussed in the introduction. By definition $A_r(P, U) = 1$. If $A_h(P, U)$ is assumed to be constant and thus independent of utility, so that $A_h(P, U) = A_h(P)$, equivalence scales are said to be "base independent" (Pendakur, 1999).

3.2 Informal preprocessing

Irrespective of the data used or the methods applied, most econometric analyses of equivalence scales apply some kind of informal data preprocessing. For example, Pendakur (1999) uses data from the 1990 Canadian Family Expenditure Survey and restricts attention to households from metropolitan areas with all members below age 65. The reasoning behind this is that households of pensioners and rural households exhibit expenditure patterns or opportunities for home production which differ from those of other households. Another example is given by Wilke (2006) who uses German microdata from the Survey of Income and Expenditure (*Einkommens- und Verbrauchsstichprobe*; EVS). He aims at estimating equivalence scales for specific subgroups of households, such as households with low income, and splits the data set into subsamples in order to achieve homogeneity with respect to other covariates, like age. Both Pendakur (1999) and Wilke (2006) apply semiparametric methods, but similar approaches can also be found in work using parametric methods (e.g., Nelson, 1988; Phipps, 1998; Donaldson and Pendakur, 2004) and in other recent research on household behavior (e.g., Chiappori et al, 2002; Lewbel and Pendakur, 2008; Dauphin et al, 2011; Lise and Seitz, 2011; Bargain et al, 2013; Browning et al, 2013).

IP (informal preprocessing) as just described is meant to reach three goals. First, from a technical point of view, preprocessing can be seen as a way to restrict attention to households in the common-support region, i.e., to values of X for which $\Pr(X|D=0)>0$ and $\Pr(X|D=1)>0$, where for example D could capture household type, some policy measure which affects some households but not all, and the like.[1] Second, it will typically also lead to better balancing. That is, if $d[P(X|D=1), P(X|D=0)]$ denotes some measure of dissimilarity of the conditional distributions of X without preprocessing, and $d_p[P(X|D=1), P(X|D=0)]$ denotes the same measure after preprocessing, then $d_p[P(X|D=1), P(X|D=0)] < d[P(X|D=1), P(X|D=0)]$. Third, as a material consequence of all this, the analysis will focus on households which are similar with respect to household behavior, implying that commonly used welfare indicators have the same interpretation for, and are comparable across, different households. As a classical example, consider the share of expenditure for food, as suggested by Engel (1857). If two households A and B have the same expenditure share s^*, but household B relies more on home production than household A, comparability of their welfare levels will be limited. But if A and B are comparable with respect to household production, we can assume their welfare level to be equal (if we assume the expenditure share for food to be a valid welfare indicator).

[1] The potential outcomes framework can also be used to analyze more than two groups, i.e., cases where D has more levels than 0 and 1.

3.3 Matching on welfare indicators

Thus, basic goals of IP and matching coincide. The use of matching methods for a two-stage estimation of equivalence scales has first been suggested by Szulc (2009, 2011). In his paper, treatment and control group are identified, for instance, as couples with one child ($D = 1$) and childless couples ($D = 0$), respectively. Matching these two groups based on some welfare indicator Z means trying to achieve balance of this indicator, so that $\Pr(Z|D = 1)$ tends to be the same as $\Pr(Z|D = 0)$. To fully understand the advantages involved, consider the extreme case where $\Pr(Z < g|D = 1) = 1$ and $\Pr(Z > g|D = 0) = 1$. In this case, the conditional distributions of Z do not overlap, and treatment and control group are not at all comparable with respect to household welfare. Applying some method for estimating equivalence scales will nevertheless lead to some scale estimate which is purely based on extrapolations that are somehow embedded in the empirical model. Now assume that $\Pr(Z > g|D = 0) = 0$ and $\Pr(Z > g|D = 1) > 0$. Again, using some standard method for estimating equivalence scales will yield results based on hidden extrapolations. Using matching techniques and controlling for the balance achieved as proposed in section 2 helps to avoid this.

Szulc (2009) follows the suggestions of Abadie and Imbens (2011) and combines matching with regression-adjustment in order to reduce potential biases (see subsection 2.3). Using household income as the relevant outcome, $y_i(1)$ denotes income of household i, a couple with one child, while $y_i(0)$ is the income of a couple without children. Conditioning both $y_i(1)$ and $y_i(0)$ on some welfare indicator Z would allow for directly calculating a household-specific equivalence scale $y_i(1)/y_i(0)$. But as only one of these values is observable, the ratio $Y(1)/Y(0)$ has to be estimated controlling for Z, making sure that treatment and control group are comparable with respect to their welfare level. This is achieved using socio-demographic characteristics and simple welfare indicators like the expenditure share of food as matching variables. Afterwards, $\hat{y}_i(1)$ and $\hat{y}_i(0)$ are adjusted via linear regression to remove potential biases due to remaining imbalance between the two groups.

Unfortunately, the procedures applied by Szulc (2009) involve a drawback. To arrive at an estimate of $Y(1)/Y(0)$, he uses log household income instead of household income. He first applies matching, then adjusts via regression, and finally uses adjusted values $\hat{y}_i^a(1)$ and $\hat{y}_i^a(0)$ to estimate the ATE captured by equation (5). Because of the logarithmic transformation, (5) equals the geometric mean $\mathrm{G}(\hat{Y}^a(1)/\hat{Y}^a(0))$, not the expectation value $\mathrm{E}(\hat{Y}^a(1)/\hat{Y}^a(0))$, and the latter can never be recovered using this approach.[2] As is well

[2]The reason is that the assumptions introduced in section 2 only allow for identifation of the marginal distributions of $Y(1)$ and $Y(0)$, not their joint distribution (Abbring and Heckman, 2007), which would be required to calculate $\mathrm{E}(\hat{Y}^a(1)/\hat{Y}^a(0))$.

known, G(·) will always be smaller than E(·). More specifically, any mean-preserving increase in the spread will further decrease G(·), while E(·) remains unchanged. That is, the larger the variance of $\hat{Y}^a(1)/\hat{Y}^a(0)$, the smaller G(·) will be compared to E(·).

3.4 Matching as non-parametric preprocessing

Because of this rather undesirable property, we follow King and Zeng (2006) and others, i.e., the second perspective described in subsection 2.3, where the use of matching techniques serves as a kind of formal, non-parametric data preprocessing. Also, this approach is more in line with usual procedures applied in the literature. The basic idea is simple. In a first step, households are matched using socio-demographic variables and possibly welfare indicators (other than income). The second step consists of the application of some standard approach to estimating equivalence scales. Note that this will generally affect which variables can be treated as dependent and independent ones in the second step. Szulc (2009) uses household income Y as outcome in both the matching estimation and the regression-adjustment step. In our context this would come close to the circular procedure described in the introduction. Instead, we will apply standard approaches to estimating equivalence scales using one or more welfare indicators as outcome and indirectly deriving the scales weights from parameter estimates (see below). This usually requires additional assumptions to identify $\mathrm{E}(Y(1)/Y(0))$.

As an example, consider the well-known approach by Engel (for a detailed discussion see Deaton and Muellbauer, 1986). A simple specification is given by

$$Z = a + \log Y b_Y + D b_D + X b_x + \epsilon, \tag{12}$$

where Z denotes the expenditure share of food, Y is household income, D is the treatment indicator as defined before, X is a vector of additional characteristics and ϵ is an error term. Estimates for this equation could be used to impute both $\hat{z}_i(1)$ and $\hat{z}_i(0)$ for each observation. Instead, we can also use two equations,

$$\begin{aligned} \hat{z}_i(d_i) &= \hat{a} + \log y_i(0)\hat{b}_Y + x_i\hat{b}_X \\ \hat{z}_i(d_i) &= \hat{a} + \log y_i(1)\hat{b}_Y + \hat{b}_D + x_i\hat{b}_X, \end{aligned} \tag{13}$$

where d_i denotes the treatment indicator for each unit i. If $d_i = 0$, $y_i(0)$ is observed and $y_i(1)$ is unknown. If, instead, $d_i = 1$, $y_i(1)$ is observed and $y_i(0)$ is not. Since both $\hat{z}_i(d_i)$ and x_i are observed as well, this leads to two equations with one unknown. Solving for

$y_i(1)/y_i(0)$ yields

$$\frac{y_i(1)}{y_i(0)} = \exp\left(-\frac{\hat{b}_D}{\hat{b}_Y}\right), \tag{14}$$

which in turn leads to

$$\mathrm{E}\left(\frac{Y(1)}{Y(0)}\right) = \exp\left(-\frac{\hat{b}_D}{\hat{b}_Y}\right), \tag{15}$$

because the right-hand side of (14) is constant.

Note that $\mathrm{E}\,(Y(1)/Y(0))$ is identified at the cost of assuming a specific functional form and that equivalence scales are independent of the welfare level (the latter assumption could be relaxed, though; see, e.g., Lancaster and Ray 1998 and section 4).

Matching on characteristics X before applying the Engel approach (or some other method for estimating equivalence scales) will generally reduce the sensitivity of results with respect to model specification. A few more caveats are nevertheless required. In the matching step, it is not possible to include household income Y or the welfare indicator Z. Y can not be included, as some equivalence scale would then be required to make income of different types of households comparable – which is the ultimate goal of the analysis. Matching on Z will induce a tendency towards $b_D = 0$ and a scale value of 1. Ruling out Z as a candidate is problematic if one wishes to match on welfare indicators. This will turn out to be impossible, if only a single welfare indicator is available, in which case matching has to be restricted to socio-demographic characteristics. If two indicators Z_1 and Z_2 are available, one of them can be used in the matching step and the other in the regression step. Note that this requires Z_1 and Z_2 not to be perfectly correlated, as this would again imply a tendency towards $b_D = 0$. If more than two (independent) welfare indicators are available, one has to be used as the dependent variable in the regression step, while all others can be included in the matching step, allowing to introduce several welfare indicators into the analysis in a straightforward manner.

If two or more welfare indicators are available, another approach is to use simultaneous-equation models, e.g., demand systems for expenditure data. In this case, one can proceed in the following fashion. First, match observations using all welfare indicators except Z_1. Let M_1 denote the set of matched units. Proceed in a similar fashion for all other welfare indicators, in each case giving a set of matched units M_i. Analysis will then be carried out using the intersection of all sets, $M_1 \cap M_2 \cap M_3 \cap \ldots$ This guarantees that only households are included in the analysis that are appropriate with respect to each single equation. Note, however, that this may drastically reduce the number of observations

used for estimating equivalence scales, depending on the number of welfare indicators, the number of other variables, and the specific matching procedure used.

4 Empirical examples

4.1 Set-up

In what follows, the procedure outlined in this paper thus far will be applied to three examples. All examples are based on data from the German Survey of Income and Expenditure (*Einkommens- und Verbrauchsstichprobe*; EVS) conducted in 2008. The EVS is a quinquennial survey covering about 0.2 percent of the total population which includes detailed questions on household income, expenditure and socio-demographic characteristics of household members. All households record their income and expenditure in a certain quarter of the relevant year.

In each of the examples, we are interested in estimating the scale weight of couples with one child below age 14 compared to childless couples. The first example looks into the sensitivity to model specification with and without non-parametric preprocessing (NPP) without including welfare indicators in the matching step. The second example is a variant of the first one and turns to the estimation of income-dependent equivalence scales, achieving dependence on income through the way a model for the entire sample is specified. Again, welfare indicators are not included in the matching step. The third example focuses on the estimation of equivalence scales for a specific subgroup, *viz.* the poorest 20% of couples with one child, using welfare indicators in the matching step.

In all examples, Engel's approach is applied as described in the previous section. To assess sensitivity, about 8,000 different model specifications are estimated before and after NPP and balance checking, leading to a total of around 16,000 scale estimates for each example. In each specification, log household income and a dummy variable for the household type are included which are definitely required to arrive at a scale estimate; see equation (15). The differing specifications additionally consist of all possible combinations of: age of household head, age of household head squared, a dummy variable for the broad region (East Germany=1; West Germany=0), a set of dummies for the type of region (7 types), a set of dummy variables for employment status of the household head (5 possible states), a dummy variable for dual-earner couples (dual earner=1; otherwise=0), a set of dummy variables for education of the household head (5 possible levels), and a set of dummy variables for the quarter of year to capture seasonal effects. Age squared is only included if age also enters the estimate linearly. Also, interaction terms of the dual-earner dummy variable with all other variables just listed are included in some of the specifications.

Matching proceeds in the following fashion. Age, education and employment status of the household head, region and type of region, quarter and a dummy for dual-earner couples are always included in the matching step. As an additional welfare indicator, expenditure on clothing for adults is used, following the proposal by Rothbarth (1943). Regardless of whether this additional welfare indicator was included or not, one childless couple is matched to each of the couples with one child. This restricts results to the support of the variables for couples with one child. It is done because numbers of observations of couples with and without a child are very different (see below). Furthermore, basing estimates on households off the support of couples with children would mean that the resulting scale weights are partly based on households which are not actually observed with children.

To determine which observations to match, the distance function $d(x_i, x_j) = ||x_i - x_j||_V$ is used, where $||a||_V = \sqrt{a'Va}$ and V is a weighting matrix (Abadie and Imbens, 2011). V is set to the inverse of the diagonal variance matrix of the included variables. All analyses were performed using the freely available software R (R Core Team, 2014) and the Matching package (Sekhon, 2011).

Balancing before and after matching is assessed based on different measures. In the case of age, the difference in means between treatment and control group is used. The same holds for the proportions of East Germans, dual-earner households, and expenditure on clothing for adults. In the cases of educational attainment, type of region, employment status, and quarter of the year, the index of dissimilarity is used (e.g., Iacus et al, 2011), defined as

$$\mathrm{DI} = \frac{1}{2} \sum |\Pr(X = x|D = 1) - \Pr(X = x|D = 0)|. \tag{16}$$

DI equals zero if the distributions of covariates in treatment and control group coincide. A value of 1 results if the distributions are completely dissimilar.

Note that for most analyses IP as described in section 3 was applied, excluding households with members above age 65, with no employed members, or with at least one member receiving unemployment benefits. This means that results are compared which are either based on IP only or on both informal and formal preprocessing. In addition, some observations are dropped because of missing or implausible values. This leaves $2,314$ couples with one child as the tratment group and $7,054$ couples without children as potential controls for the following analyses.

4.2 Full sample

Table 1 includes results regarding the balance of treatment and control group before informal preprocessing (IP), after informal but before non-parametric preprocessing (NPP), and after matching.

[Table 1 about here.]

IP already reduces differences between treatment and control group considerably. For example, before IP mean age of the household head is almost 20 years higher for childless couples than for couples with one child. After preprocessing, the difference decreases to 8.6 years. After matching, virtually no differences are left for any of the variables considered apart from a small difference in mean age of household head (of 2.3 years).

While the effects of IP are clearly strong, one may wonder whether the smaller effects of NPP are worth the effort. Figure 1 shows the density of estimated scale weights which result with and without matching. In both cases, IP is being applied beforehand.

[Figure 1 about here.]

The range of results is 0.13 without matching and 0.04 after matching, the standard deviations of estimates are 0.04 and 0.01, respectively. This corresponds to a decrease of 70% (range) and 80% (standard deviation) of the variation after matching. Further note that range and standard deviation in case of IP are only slighlty smaller than without any preprocessing (range: 0.14; standard deviation: 0.04). Thus, matching has a strong effect on sensitivity, and estimates are more stable, despite the reduction in the number of observations. This is mostly due to the imbalance in age which is left after IP. Specifically, the estimates between 1.18 and 1.20 which result when NPP is not applied are mostly derived from model specifications that do not include age.[3]

The results shown in figure 2 relate to our second example. They are based on the same data as the results in figure 1. The only difference is that each model includes $D/Y(1)$ as an additional explanatory variable, that is, the inverse of household income of couples with one child.

[Figure 2 about here.]

As a result, equivalence scales now depend on income and are given by

$$\mathrm{E}\left(\frac{Y(1)}{Y(0)}|Y(1)\right) = \exp\left(\frac{-\hat{b}_D}{\hat{b}_Y} + \frac{-\hat{\gamma}\frac{1}{Y(1)}}{\hat{b}_Y}\right),$$

[3]Therefore, the example may seem somewhat artificial, since age is next to always included in parametric models. Note, however, that this does not always hold for non-parametric approaches which are in principle prone to the same problems due to imbalance as parametric approaches.

where γ is the coefficient of $D/Y(1)$. Note that this is not the only possibility to arrive at income-specific estimates, and others could have been used here instead (e.g., Lancaster and Ray, 1998).

Figure 2 shows the lowest and highest scale weights derived from all models as they vary by income $Y(1)$. The difference between highest and lowest scale weights ranges from 0.12 to 0.17 without matching and from 0.04 to 0.07 with matching. For example, this means that, depending on model specification and without matching, a couple with one child and household income of 2,000 Euro turns out to be comparable to childless couples with income between 1,317 and 1,467 Euro, whereas the corresponding figures with matching are 1,300 and 1,351 Euro. Again, matching considerably reduces sensitivity to model specification. Note, however, that specifying income-dependence based on $D/Y(1)$ leads to implausibly high scale weights for households with low incomes with as well as without matching.

4.3 Estimates for the lowest income quintile

The implausible results for low-income households showing up in figure 2 are possibly due to the restriction that b_Y and γ are the same for all households, irrespective of whether they have low or high income. It may thus be desirable to estimate equivalence scales for low-income households in a different way. One way to proceed is to use only the lowest income quintile of couples with one child, i.e., the poorest 20%, or any other "low" quantile. This leads to the problem which households of childless couples to include in the analysis. One could simply take the poorest 20%, but this raises the question whether the poorest childless couples and the poorest couples with one child are comparable to each other in terms of their welfare level. If for some reason the poorest 20% of childless couples were worse off than the poorest couples with one child, it might be appropriate to include childless couples with an income between the 5% and 25% quantile. However, this would require scale weights or some similar information to be known *ex ante*.

After informal preprocessing, the poorest childless couple in the EVS data has a monthly income of 742 Euro, and the poorest couple with one child has an income of 1,225 Euro. They would be equally well-off at a scale weight of about 1.65, which appears rather high. On the other hand, the "richest" household among the poorest 20% has a monthly income of 1,804 Euro in the case of childless couples and 1,989 Euro for couples with one child. These figures would be equivalent at a scale weight of 1.1. This implies that, among the poorest 20%, the "poorest poor" and the "richest poor" can only be comparable simultaneously, if equivalence scales exhibit an unrealistically steep decline as income increases. In other words, it is an indication that the lowest quintiles of both groups are

possibly not comparable.

Therefore, matching can be applied to identify those childless couples that are comparable to the poorest couples with one child. Here, expenditure on clothing for adults is included in the matching as a welfare indicator. In addition, childless couples with a monthly household income above 3,200 Euro were dropped to prevent matching of low-income couples with one child with high-income childless couples. The value of 3,200 Euro was derived from the (rounded) maximum monthly income of low-income couples with one child (2,000 Euro) times 1.6, where 1.6 can be safely assumed to be an upper bound of the scale weight *ex ante*. Using the resulting data set of matched households, Engel's approach was applied and scale weights were estimated using equation (15). All in all, this procedure only requires base independence to hold locally.

Table 2 shows results on balancing after informal preprocessing, but before matching, comparing the poorest couples with one child to the full sample of childless couples. It also includes results based on the naive approach of using the lowest income quintile of childless couples for comparison. Finally, it displays results after matching (including a welfare indicator) has been applied as formal, non-parametric preprocessing.

[Table 2 about here.]

In this case, informal preprocessing leaves substantial differences between treatment and control group, in contrast to the results shown in table 1. The naive variant leads to better balancing for most, but not all, variables, while some differences still remain. After matching most variables are balanced, except for the continuous variables age and expenditure on clothing which show rather small differences, though.

Final results of the scale weight estimations are shown in figure 3.

[Figure 3 about here.]

These results differ markedly in two respects. First, sensitivity to model specification is large both for unmatched results and the naive variant. The range is 0.15 (unmatched) and 0.14 (naive), and standard deviations amount to 0.05 and 0.04, respectively. Applying non-parametric preprocessing reduces both the range (0.06) and the standard deviation (0.02) and thus leads to more stable results. Second, the means of all estimates are 1.43 respectively 1.27 in the case of unmatched and naive results. Matching leads to a mean of 1.31, in between the two other figures. The same relations hold for median estimates as well as minimum and maximum estimates. Using matching in order to identify suitable control units thus turns out to have strong effects on both sensitivity and the level of estimates in this case, where we are seeking income-dependent scales weights for households with low income.

What can also be taken away from these last results is that scale weights for low-income households, i.e., those located in the lowest quintile of couples with one child, are indeed higher than those obtained for the full sample, if scale weights are assumed to be income-invariant (see figure 1). But they are lower than those suggested by figure 2 for households with income below 2,000 Euro.

5 Conclusion

In this paper, we have argued that most econometric analyses of household behavior employ informal data preprocessing to achieve comparability of observations across different household types. We propose matching and balance checking as useful elements of a two-stage procedure which is more firmly rooted in statistical theory and provide appropriate methods for assessing and establishing comparability. The procedure is open to combining different variants of matching with a variety of equivalence scale estimators (Dudel et al, 2013). Results building on this approach to formal, non-parametric preprocessing can be expected to be less sensitive with respect to model specification than results which only rely on informal preprocessing. In some cases, also the level of estimates may be affected.

As an example, we used German expenditure data to estimate equivalence scales for couples with one child below age 14 compared to childless couples. Apart from demonstrating the increased stability of estimates, a methodology for estimating equivalence scales for specific subpopulations – e.g., income groups – was developed and applied, following a proposal by Szulc (2009), improving on his approach, and making use of additional welfare indicators for matching. Results prove to be less sensitive and more plausible than results relying on other methods. Furthermore, the approach can be easily adopted using standard statistical software.

Possible extensions of our approach include the use of more diverse welfare indicators and the combination with more complex regression models. For example, welfare indicators could include indicators highlighted in Amartya Sen's capability approach (Lelli, 2005). Other methods for estimating equivalence scales which could possibly benefit from non-parametric preprocessing include semi- and non-parametric approaches. Especially for the latter, inclusion of additional control variables – next to some welfare indicator, income, and household type – is nontrivial.

Bibliography

Abadie A, Imbens GW (2011) Bias-corrected matching estimators for average treatment effects. Journal of Business and Econonic Statistics 29:1–11

Abbring JH, Heckman JJ (2007) Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In: Handbook of Econometrics, Volume 6B, Elsevier, pp 5146–5303

Bargain O, Donni O, Kwenda P (2013) Intrahousehold distribution and poverty: Evidence from Cote d'Ivoire, Thema Working Paper No. 2013-23

Bellemare C, Melenberg B, van Soest A (2002) Semi-parametric models for satisfaction with income. Portuguese Economic Journal 1:181–203

Blundell R, Browning M, Crawford I (2003) Nonparametric engel curves and revealed preference. Econometrica 71:205–240

Browning M, Chiappori PA, Lewbel A (2013) Estimating consumption economies of scale, adult equivalence scales, and household bargaining power. Review of Economic Studies 80:1267–1303

Buhmann B, Rainwater L, Schmaus G, Smeeding TM (1988) Equivalence scales, well-being, inequality, and poverty: Sensitivity estimates across ten countries using the Luxembourg Income Study (LIS) database. Review of Income and Wealth 34:115–142

Chiappori PA, Fortin B, Lacroix G (2002) Marriage market, divorce legislation, and household labor supply. Journal of Political Economy 110:37–72

Coulter FAE, Cowell FA, Jenkins SP (1992) Equivalence scale relativities and the extent of inequality and poverty. Economic Journal 102:1067–1082

Dauphin A, El Lahga AR, Fortin B, Lacroix G (2011) Are children decision-makers within the household? Economic Journal 121:871–903

Deaton A, Muellbauer J (1980) Economics and consumer behavior. Cambridge University Press, Cambridge, MA

Deaton A, Muellbauer J (1986) On measuring child costs: With application to poor countries. Journal of Political Economy 94:720–744

Dettmann E, Becker C, Schmeißer C (2011) Distance functions for matching in small samples. Computational Statistics and Data Analysis 55:1942–1960

Donaldson D, Pendakur K (2004) Equivalent-expenditure functions and expenditure-dependent equivalence scales. Journal of Public Economics 88:175–208

Dudel C, Garbuszus JM, Ott N, Werding M (2013) Überprüfung der bestehenden und Entwicklung neuer Verteilungsschlüssel zur Ermittlung von Regelbedarfen auf Basis der Einkommens- und Verbrauchsstichprobe 2008, Research Report for the Federal Ministry of Labour and Social Affairs

Engel E (1857) Die Productions- und Consumptionsverhältnisse des Königsreichs Sachsen. Zeitschrift des Statistischen Bureaus des Königlich Sächsischen Ministeriums des Inneren 3:8+9

Hagenaars A, de Vos K, Zaidi M (1994) Poverty Statistics in the Late 1980s: Research Based on Micro-data, Office for Official Publications of the European Communities, Luxembourg

Hainmueller J (2012) Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political Analysis 20:25–46

Heckman JJ (2008) Econometric causality. International Statistical Review 76:1–27

Ho DE, Imai K, King G, Stuart EA (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis 15:199–236

Holland PW (1986) Statistics and causal inference. Journal of the American Statistical Association 81:945–960

Iacus SM, King G, Porro G (2011) Multivariate matching methods that are monotonic imbalance bounding. Journal of the American Statistical Association 106:345–361

Iacus SM, King G, Porro G (2012) Causal inference without banalce checking: Coarsened exact matching. Political Analysis 20:1–24

Imai K, Van Dyk DA (2004) Causal inference with general treatment regimes: Generalizing the propensity score. Journal of the American Statistical Association 99:854–866

Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. Review of Economics and Statistics 86:4–29

Imbens GW, Wooldridge JM (2009) Recent develpments in the econometrics of program evaluation.

Journal of Economic Literature 47:5–86
Kapteyn A, Van Praag B (1975) A new approach to the construction of family equivalence scales. European Economic Review 7:313–335
King G, Zeng L (2006) The dangers of extreme counterfactuals. Political Analysis 14:131–159
Koulovatianos C, Schröder C, Schmidt U (2005) On the income dependence of equivalence scales. Journal of Public Economics 89:967–996
Lancaster G, Ray R (1998) Comparison of alternative models of household eqivalence scales: The australian evidence on unit record data. Economic Record 74:1–14
Lelli S (2005) Using functionings to estimate equivalence scales. Review of Income and Wealth 51:255–284
Lewbel A, Pendakur K (2008) Estimation of collective household models with engel curves. Journal of Econometrics 147:350–358
Lise J, Seitz S (2011) Consumption inequality and intra-household allocations. Review of Economic Studies 78:328–355
Morgan SL, Harding DJ (2006) Matching estimators of causal effects: Prospects and pitfalls in theory and practice. Sociological Methods and Research 35:3–60
Morgan SL, Winship C (2010) Counterfactuals and Causal Inference. Methods and Principles for Social Research. Cambridge University Press, Cambridge, MA
Nelson JA (1988) Household economies of scale in consumption: Theory and evidence. Econometrica 56:1301–1314
Pendakur K (1999) Semiparametric estimates and tests of base-independent equivalence scales. Journal of Econometrics 88:1–40
Phipps SA (1998) What is the income "cost of a child"? Exact equivalence scales for canadian two-parent families. Review of Economics and Statistics 80:157–164
Pollak RA, Wales TJ (1978) Estimation of complete demand systems from household budget data: The linear and quadratic expenditure systems. American Economic Review 68:348–359
R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/
Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55
Rosenbaum PR, Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 79:516–524
Rothbarth E (1943) Note on a method of determining equivalent income for families of different composition. In: Madge C (ed) War Time Pattern of Saving and Spending, Cambridge University Press, Cambridge, app. 4
Rubin DB (1973) The use of matched sampling and regression adjustment to remove bias in observational studies. Biometrics 29:185–203
Rubin DB (1979) Using multivariate matched sampling and regression adjustment to control bias in obersvational studies. Journal of the American Statistical Association 74:318–328
Rubin DB, Thomas N (2000) Combining propensity score matching with additional adjustments for prognostic covariates. Journal of the American Statistical Association 95:573–585
Schröder C (2009) Variable Income Equivalence Scales: An Empirical Approach. Springer, New York
Sekhon JS (2008) The Neyman-Rubin model of causal inference and estimation via matching methods. In: Box-Steffensmeier JM, Brady HE, Collier D (eds) The Oxford Handbook of Political Methodology, Oxford University Press, Oxford, p 271–299
Sekhon JS (2011) Multivariate and propensity score matching software with automated balance optimization: The matching package for r. Journal of Statistical Software 42:1–52
Stewart MB (2009) The estimation of pensioner equivalence scales using subjective data. Review of Income and Wealth 55:907–929
Stuart EA (2010) Matching methods for causal inference: A review and a look forward. Statistical Science 25:1–21
Szulc A (2009) A matching estimator of household equivalence scales. Economics Letters 103:81–83
Szulc A (2011) Empirical versus policy equivalence scales: matching estimation, Warsaw School of Economics Working Papers 9/2011
Wilke RA (2006) Semi-parametric estimation of consumption-based equivalence scales: The case of Germany. Journal of Applied Econometrics 21:781–802

Zhao Z (2004) Using matching to estimate treatment effects: Data requirements, matching metrics, and monte carlo evidence. Review of Economics and Statistics 86:91–107

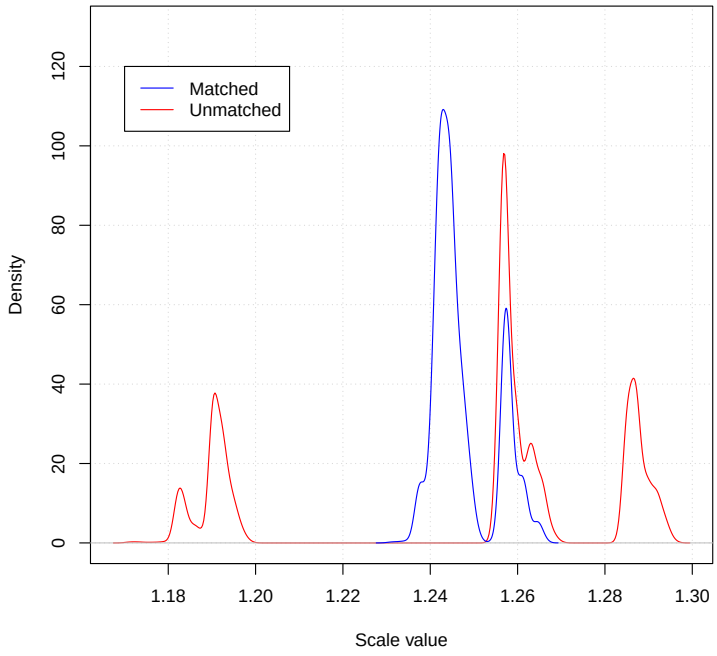

Figure 1: Scale estimates of all specifications with and without matching

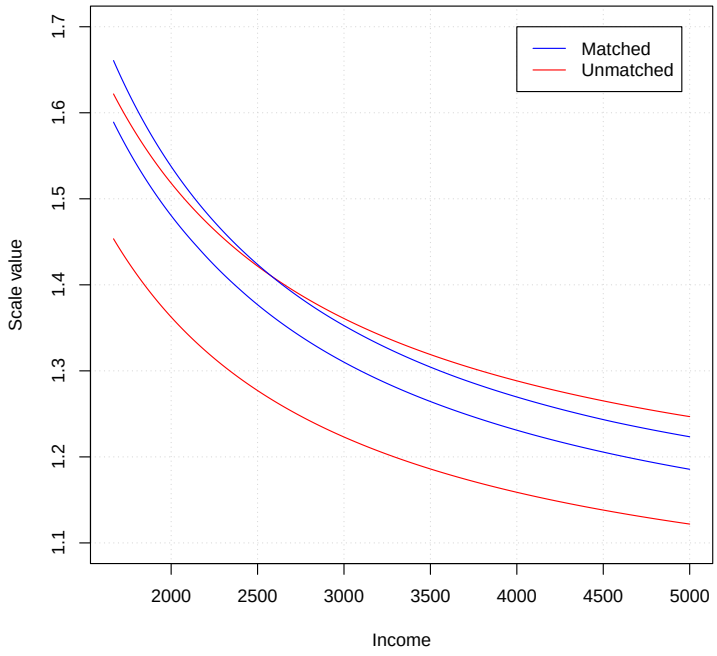

Figure 2: Income-dependent scale estimates of all specifications with and without matching

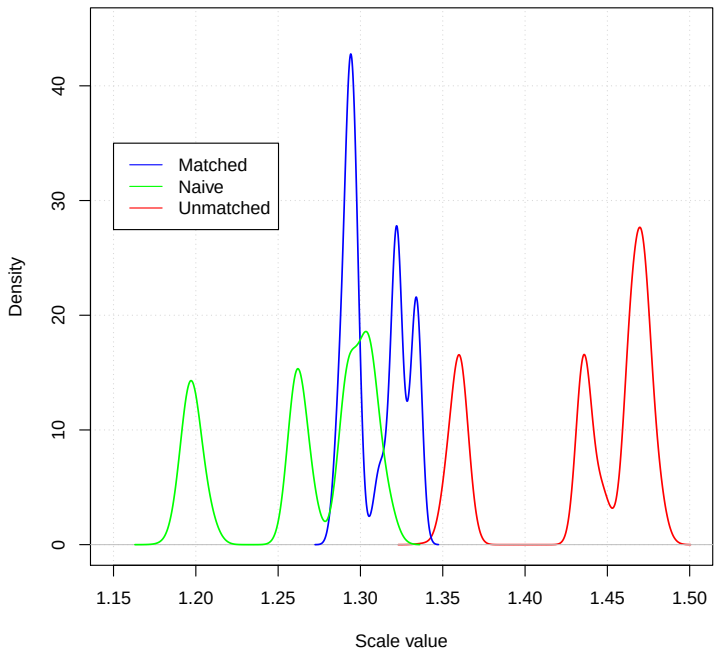

Figure 3: Scale estimates of all specifications with and without matching (poorest 20% of couples with one child)

Table 1: Balancing before informal preprocessing, before matching, and after matching

Variable	Before IP	Before NPP	After NPP
Mean Age (Difference)	19.96	8.62	2.26
Proportion East German (Difference)	0.01	0.03	0.00
Proportion Dual Earners (Difference)	-0.36	-0.05	0.00
Education (DI)	0.14	0.03	0.00
Region (DI)	0.01	0.03	0.00
Employment (DI)	0.43	0.03	0.00
Quarter (DI)	0.01	0.01	0.00
No of observations (Control)	15,533	7,054	2,314
No of observations (Treatment)	2,550	2,314	2,314

Table 2: Balancing before and after matching

Variable	Before NPP	Naive	After NPP
Mean Age (Difference)	11.03	8.36	3.71
Mean Expenditure Clothing/Month (Difference)	71.77	19.62	7.11
Proportion East German (Difference)	-0.07	0.06	0.00
Proportion Dual Earners (Difference)	0.14	-0.10	0.00
Education (DI)	0.17	0.05	0.01
Region (DI)	0.06	0.07	0.01
Employment (DI)	0.02	0.08	0.02
Quarter (DI)	0.07	0.03	0.01
No of observations (Control)	7,088	1,416	464
No of observations (Treatment)	464	464	464