Johannes Kepler Universität Linz

Department of Economics

Replication Project:

INTRAHOUSEHOLD RESOURCE ALLOCATION IN RURAL PAKISTAN: A SEMIPARAMETRIC ANALYSIS

Elaborated from

k0355435 | Rudolf Gruber

Course guidance:

Dr. Rudolf Winter-Ebmer

Mag. Alexander Ahammer

Table of Content

List of Tables

Table of Figures

1. Summary

The article “Intrahousehold Resource Allocation in Rural Pakistan: A Semiparametric Analysis”, published by Sonia Bhalotra & Cliff Attfield in 1998 in the Journal of Applied Econometrics, investigates the shape of Engel Curves for several goods in Pakistani household consumption. A large household survey enables the authors to derive consistent estimates for these Engel Curves.

1.1. Research Question

Driven by the interest in the shape of Engel Curves for rural demand in Pakistan, the authors are also investigating different patterns of consumption for age and gender. In the authors point of view, the shape of the Engel Curves effects demand and taxation and allows to infer underlying demand while identified consumption patterns contribute to existing literature describing gender discrimination in Asia. In addition to that, further insights are provided to support welfare comparisons among varying household compositions. A special question of interest is the alteration of growing income on welfare for the poor. Existing literature suggests a linear relationship between the logarithm of food share and expenditure per capita. (Bhalotra & Attfield, 1998) are doubting this linear relationship arguing, for low income households, the amount of food share declines more slowly with an increase in income than for high income households. Determining the shape of the Engel Curve (Bhalotra & Attfield, 1998) are exclusively excluding the motivation for heterogenous household consumption in their research.

1.2. Methodology

Various estimation methods exist in econometric analysis such as parametric and non-parametric regression models. If the distribution of variables can be described by a finite number of parameters and the probability distribution of variables is known, parametric models are the methodology of choice. Contrary to parametric estimation procedures, non-parametric models are used, if assumptions for the probability distribution of variables cannot be made or does not make much sense. Semiparametric estimation procedures combine parametric and non-parametric estimation methodology. Because of doubting the linear relationship between the logarithm of food share and per capita expenditure, as literature suggests, the functional form of food share on expenditure per capita is not known. Therefore, a semiparametric regression model is the methodology of choice for (Bhalotra & Attfield, 1998) to achieve consistent coefficients.

Adding various controls for household heterogeneity, the empirical model denotes as:

$$\omega_i = F(y_i) + \alpha \ln N_i + \Sigma_k \gamma_k (N_{ki}/N_i) + \varphi^T z_i + v_i$$

Figure 1: Empirical Model - (Bhalotra und Attfield 1998, S. 466)

The dependent variable $\boldsymbol{\omega_i}$ is the proportion of different shares like food, milk, adult goods and child goods of the household budget. The proportion of food share does not only include expenditures but also imputed values for home grown produce. $\boldsymbol{y_i}$ is the logarithm of total expenditures per household with F (*) being the unknown functional relationship between the logarithm of total expenditures and the proportion of a specific share. $\boldsymbol{N_i}$ is the number of household members which appears in the empirical model in logarithmic form. The sum over $\boldsymbol{\frac{N_{ki}}{N_i}}$ is the heterogenous variation in household composition for which the authors control for. Besides work status and gender, (Bhalotra & Attfield, 1998) include a much narrower view on age than previous studies in literature. For the identification of age effects in consumption patterns, especially for children, the authors include an indicator for every year of age from 0 to 14. Elder household members are binned in groups including young (15-24 years), prime-age (25-59 years) and elderly (60 years of age and more). A vector of further controls, denoted as $\boldsymbol{z_i}$ in the empirical model controls for seasonal effects, different consumption behavior among provinces and a birth order dummy for each gender to identify privileges of first born children. The error term $\boldsymbol{v_i}$ catches unobserved stochastic variation for example in taste.

The semiparametric approach which (Bhalotra & Attfield, 1998) used in their research is based on the *"Root N Consistent"* estimation procedure introduced by Robinson in 1988. Semiparametric estimation procedures require sophisticated statistical knowledge and nonparametric estimation results deviated larger than the square root of the sample size. Robinsons breakthrough in 1988 was to limit these deviations for semiparametric estimates by generalizing OLS estimates with the insertion of a nonparametric estimation procedure which consistently does not deviate more than the square root of the sample size. Not having much assumptions about a distribution, the numerical calculation for the right bandwidth selection was not easy to compute. To overcome this restriction, the kernel smoothing method via Fast Fourier Transformation, introduced by Haerdle 1987 was part of the methodological toolset which (Bhalotra & Attfield, 1998) used in their research.

1.3. Findings

Addressing their research, (Bhalotra & Attfield, 1998) had special interest in age effects because of assuming relative neglect for children and elder family members. In addition to that, more privileges for first born children are not uncommon too in the Pakistani line of latitude. These formulated assumptions by the authors cannot be validated by the estimation results. Although the study does show many interesting findings:

The proportion of food share is on average 52 percent of the expenditure per capita and food share does show economies of size as well as the proportion for adult good share. Diseconomies of size are found in the proportions for milk share and child good share. Contrary to the assumption of privileges for first born children, no empirical evidence has been found to verify this situation. In addition to the declination of birth order effects, there are not any gender differences among children either. The most influential group on food share are working males. Estimation results do provide evidence that working men show more command on expenditure than their dependent counterparts. Working man also have more impact on the proportion of food share than working females. The identified effects between working and dependent male household members cannot be transferred to females. The data does not contain any empirical proof for differences among working and nonworking females. Furthermore, the assumption of relative neglect cannot be verified either. Elder household members do not consume less than younger household members, independent of their work status.

2. Replication

Section 3 describes replication results and graphs from (Bhalotra & Attfield, 1998) which includes statements of problems which appeared during the replication process and deviations in results with possible explanations.

2.1. Data Preparation

The companion dataset to the Article “Intrahousehold Resource Allocation in Rural Pakistan: A Semiparametric Analysis” is available for download on http://qed.econ.queensu.ca/jae/1998-v13.5/bhalotra-attfield/ (Last access on 10.8.2017). A zip file contains the subset out of a stratified Pakistani household survey of about 18000 households with no more than 20 household members (Bhalotra & Attfield, 1998, S. 465) and a text file which contains a list of variables and their definitions. The format of the data file has the .dat file extension which, after some research, dat is no more than a text file which can be opened in any text editor on a computer. Looking at the data in an editor, the dataset is organized as comma separated values.

Much more concern than the unknown .dat file extension was the structure of the dataset. A single observation is not a tuple what a professional would expect in relational data but rather contained 14 lines in the dataset without any break or indicator for the next observation. Furthermore, a header containing variable names was missing either. Tupel validation took place via counting the number of variables in the definition and guessing that this structure continuous in the data file.

The key for reading the data file and building a relational table in the econometric software package, which was for this replication R (https://cran.r-project.org/, last accessed on 10.8.2017), is a read in procedure line by line building a vector object for every observation. After 14 lines read in, which equals one observation, a conditional branch, identifying the condition, number of the actual line modulo 14 equal to zero, created a new vector object and the next observation was read in to the statistical software. After completing the reading process, a typecast to the data type data frame was performed and the dimensions of the data frame equaled the number of variables and observations in the description file. The script dataPreparation.R contains the code for the reading in process. Unfortunately to the unavailability of summery statistics, I was not able to validate if the read in values for variables equaled the dataset of Bhalotra and Attfield.

2.2. Parametric Regression Results

Parametric Engel curvature was estimated with OLS with an adjusted proportion share of various demanding goods in households as depended variable and the logarithm of expenditure as well as the logarithm of the square value for expenditure as independent variables. Besides the usage of the simple term "adjusted for", which caused some confusion for me, being a novice in econometric analysis, a slightly different model specification for the estimation of the Engel curvature than for the regression tables leaves unanswered questions in this research.

Figure 2 displays parametric Engel curves from my replication with confidence interval. A vertical dotted line indicates the maximum of the assumed quadratic fit. Unfortunately, I do not get the same domain and range for the Engel Curve as (Bhalotra & Attfield, 1998) but the maximum points of the curvatures are identical to the authors.

Table 1 displays the replicated OLS output with identical results. Further, the Wald tests for joint significance deviates from (Bhalotra & Attfield, 1998). Test statistics was cross validated with Stata, which showed identical results to R. For reasons of these deviations, I do not have any explanation for except a different model specification.

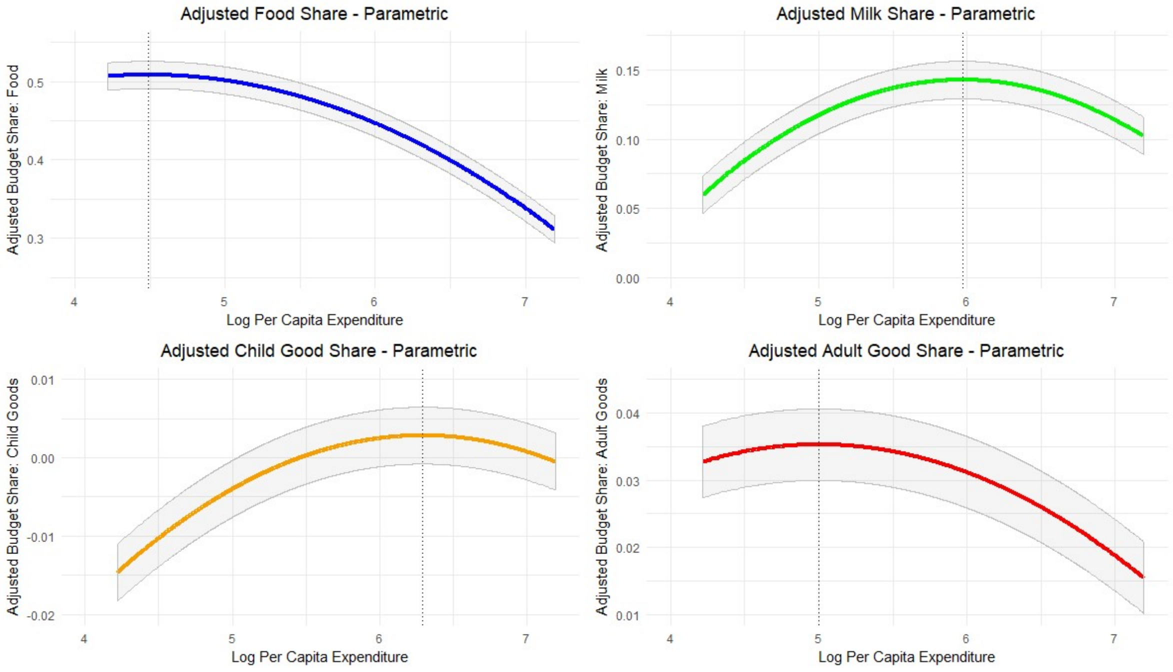

Figure 2: Parametric Engel Curves

Table 1: Parametric Regression Results

	Foodshare	Milkshare	Adult Good Share	Child Good Share
	(1)	(2)	(3)	(4)
Log Expenditure	0.164	0.367	0.014	0.044
	t = 5.860***	t = 15.540***	t = 1.467	t = 6.867***
Log Expenditure Square	-0.023	-0.030	-0.002	-0.003
	t = -9.566***	t = -14.890***	t = -2.622***	t = -6.066***
Log Household Size	-0.017	0.013	-0.002	0.006
	t = -6.160***	t = 5.545***	t = -2.559**	t = 9.244***
Birthorder Female	-0.001	0.003	-0.000	-0.000
	t = -0.171	t = 0.908	t = -0.230	t = -0.377
Birthorder Male	-0.004	0.003	-0.001	0.001
	t = -1.069	t = 1.104	t = -0.498	t = 1.398
------------------------------	--------------	---------------	---------------	---------------
Demographics	F = 6.24	F = 3.22	F = 12.61	F = 33.92
	#	#	#	#
Regions	F = 108.22	F = 132.53	F = 187.98	F = 98.43
	#	#	#	#
Seasons	F = 3.61	F = 0.31	F = 2.55	F = 0.91
	#	p = 0.82	p = 0.05	p = 0.43
Mean of dependent Variable	0.52	0.13	0.05	0.02
Y - Elasticity	-0.13	0.14	-0.16	41.04
------------------------------	--------------	---------------	---------------	---------------
Observations	9,643	9,643	9,643	9,643
R^2	0.221	0.078	0.124	0.198
Adjusted R^2	0.216	0.072	0.118	0.193
Residual Std. Error (df = 9576)	0.089	0.075	0.029	0.020
F Statistic (df = 66; 9576)	41.276***	12.305***	20.627***	35.854***

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.
# Denotes Significance at least at the 5 percent level

2.3. Nonparametric Regression Results

Performing nonparametric regressions, many procedures and packages are available in R. I didn’t find and package which implements (Robinsons 1988) semiparametric procedure in R and writing a function on my own, my statistical knowledge is not sophisticated enough for this effort, I chose semiparametric estimation with Generalized Additive Models (GAM) where one of many implementations can be found in the R package mgcv. GAM uses Splines for estimating functional forms with lambda as a smoothing parameter. Figure 3 displays the replication of the nonparametric Engel curves, again on adjusted depended variables, having a similar plot for the proportion of food share and milk share. Domain and Range deviates slightly from the original plot as well as the Engel curves for child goods and adult goods.

Table 2 displays the semiparametric regression table almost identical estimates but F statistics deviates again.

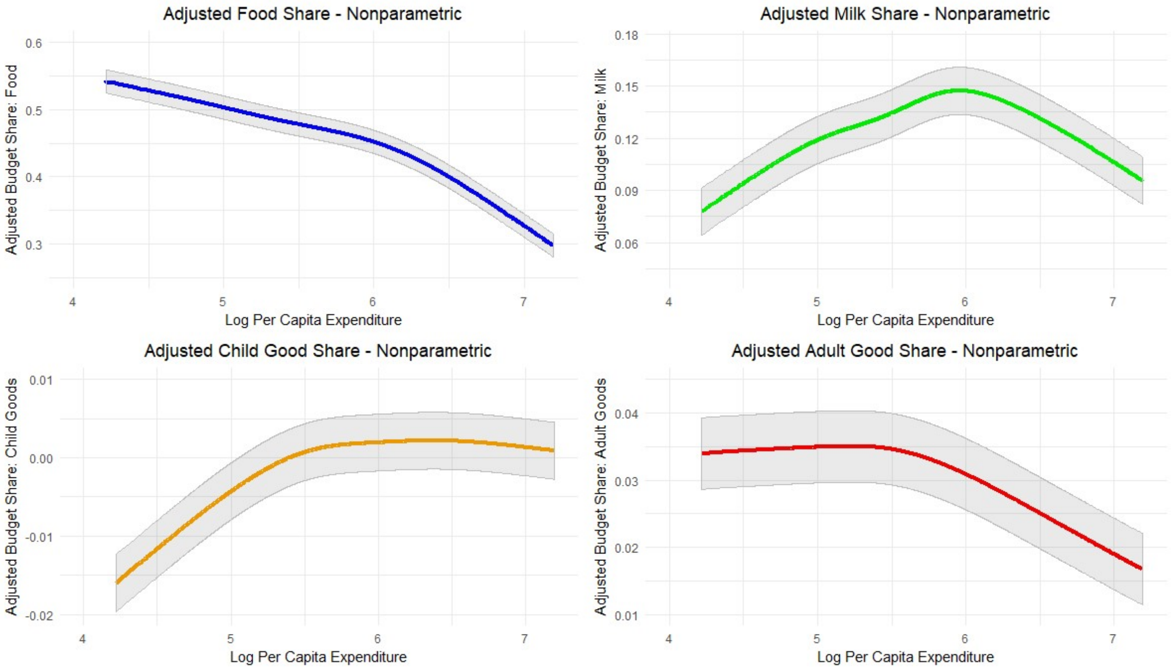

Figure 3: Nonparametric Engel Curves

Table 2: Nonparametric Regression Results

	Foodshare	Milkshare	Adult Good Share	Child Good Share
	(1)	(2)	(3)	(4)
Log Household Size	-0.016	0.013	-0.002	0.006
	t = -6.057***	t = 5.722***	t = -2.580***	t = 9.208***
Birthorder Female	-0.000	0.003	-0.000	-0.000
	t = -0.085	t = 0.994	t = -0.268	t = -0.419
Birthorder Male	-0.004	0.003	-0.001	0.001
	t = -1.035	t = 1.105	t = -0.515	t = 1.392
------------------------------	---------------	---------------	---------------	---------------
Demographics	F = 348.18	F = 179.77	F = 701.86	F = 1867.28
	#	#	#	#
Regions	F = 322.53	F = 396.93	F = 564.49	F = 296.08
	#	#	#	#
Seasons	F = 10.69	F = 0.88	F = 7.64	F = 2.73
	#	#	#	#
Mean of dependent Variable	0.52	0.13	0.05	0.02
------------------------------	---------------	---------------	---------------	---------------
Observations	9,643	9,643	9,643	9,643
Adjusted R^2	0.217	0.074	0.119	0.193
Log Likelihood	9,658.826	11,317.520	20,372.830	23,946.270
UBRE	0.008	0.006	0.001	0.000

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.
# Denotes Significance at least at the 5 percent level

2.4. Spurious Engel Curvature

To avoid spurious curvature in Engel curves, which may arise because of different slopes for heterogenous groups, if they are pooled into one sample. (Bhalotra & Attfield, 1998) determined semiparametric Engel curves for subsamples of the data. These subsamples were partitioned into regions, one earner-, two earner-, three earner or above households.

Besides indicator variables for regions, a stratification in number of earner households have not been available in the dataset. To overcome this restriction, I summed up the working proportion of a household and multiplied the sum with the exponential of household size. An indicator for the number of earners per household was created equal to (Bhalotra & Attfield, 1998). Subsequently, predictions for Engel curves with a prior defined regression model has been made. Figure 4 displays Engel curves for multiple earner households while Figure 5 visualizes the Engel curves for the different regions in Pakistan.

I can replicate the Engel curve for the regions as displayed in (Bhalotra & Attfield, 1998) but the replication for the number of household's curves deviates heavily. I assume an error in the variable creation process as causation. It is even possible, that the model for identifying the Engel curves in Figure 2 is not appropriate for predicting the Engel curves for number of earners because this information was implicitly coded in the investigation of the Engel curves. Adding an extra control for number of earners might be the better choice for this purpose.

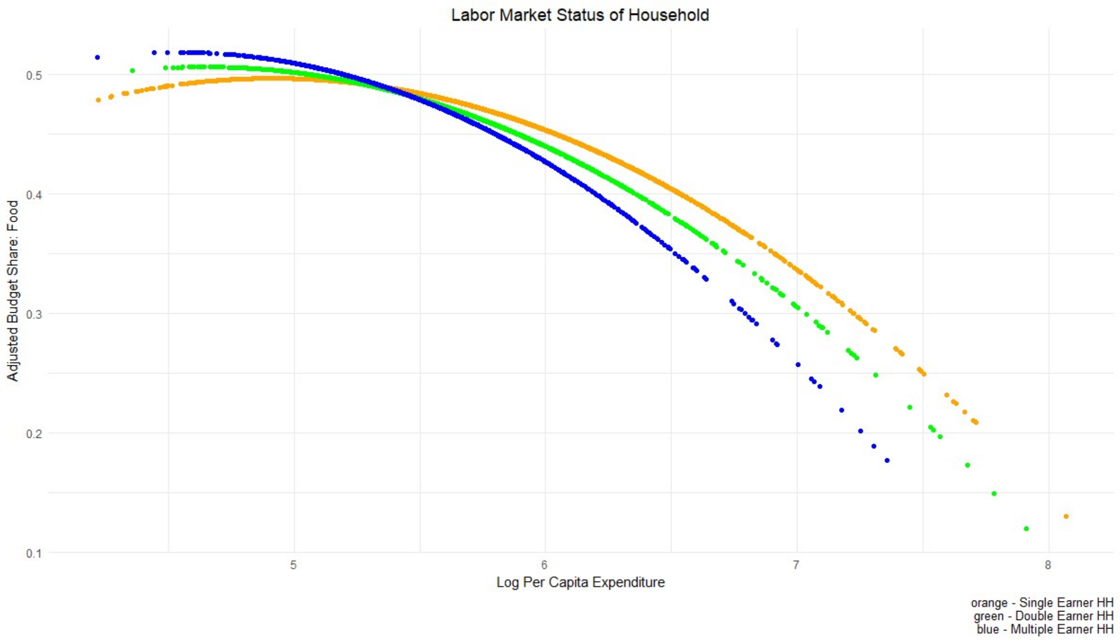

Figure 4: Labor Market Status of Household

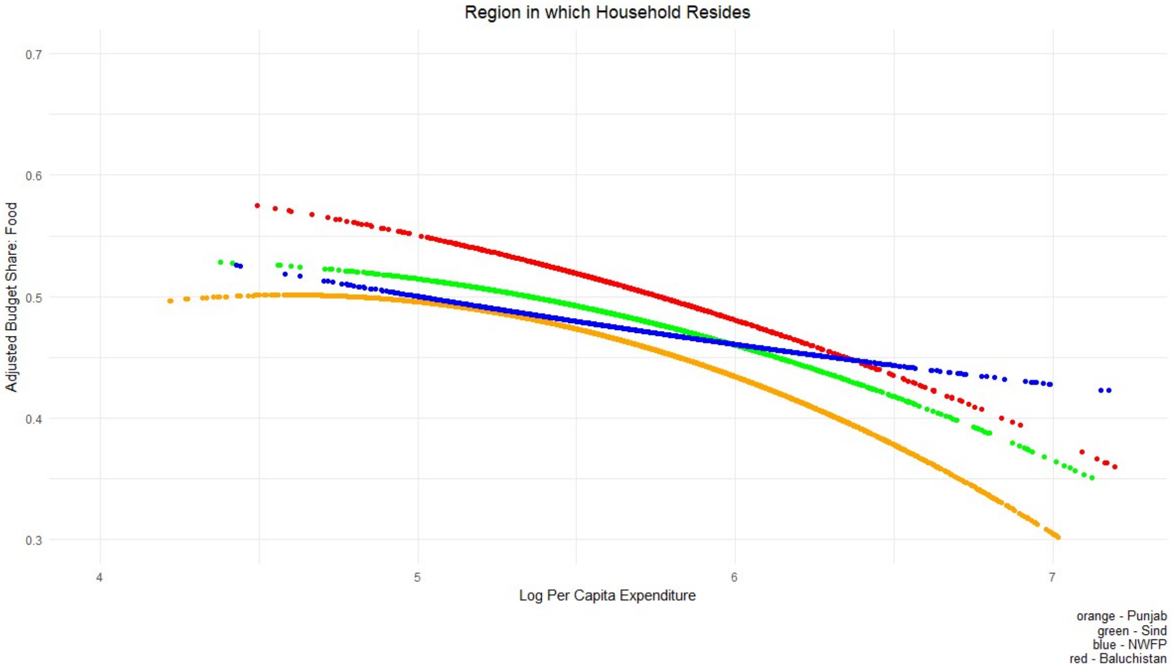

Figure 5: Region in which Household Resides

2.5. Nonparametric Elasticities

(Bhalotra & Attfield, 1998, S. 470) are using nonparametric elasticities over the full range of income to determine systematic measurement errors in food expenditure or total expenditure. The authors describe having nonparametric elasticities at every point of the Engel curve as “useful” if measurement errors appear to a specific part of the expenditure distribution. Parametric elasticities at a mean value are likely to be biased because of the data bias in the slope of the Engel Curve. The calculation of the elasticities is given by $1+\frac{\acute{m_y}}{m_y}$ where m_y is the nonparametric estimation estimate and m_y the estimation of it’s derivative.

One determinant of the shape for elasticities is the choice of the correct bandwidth for the density function. Is the chosen bandwidth for a kernel too narrow, not much smoothing can be identified. In the opposite, a large bandwidth selection leads to over smoothing and no form of the kernel density can be identified. Besides, estimating the right bandwidth selection can be computationally intense. If a kernel appears in a Gaussian form, Silverman’s rule of thumb is an appropriate estimate for bandwidth selection but when calculating the bandwidth as Silverman proposes, the nonparametric elasticities did not look in any form like the identifications of (Bhalotra & Attfield, 1998). For the replication, I used R's built in density function which also returns a bandwidth. Slight adoptions and trial and error led to these visuals. It is noticeable, that food is a necessary good for the whole sample. The horizontal line at the value of 1 marks the distinction between necessary and luxury goods. The shapes of curvature are somehow like the original identifications except child good share where extreme outliers appear positive as well as negative. I assume a data issue because the same procedure was used for every good when replication elasticicities.

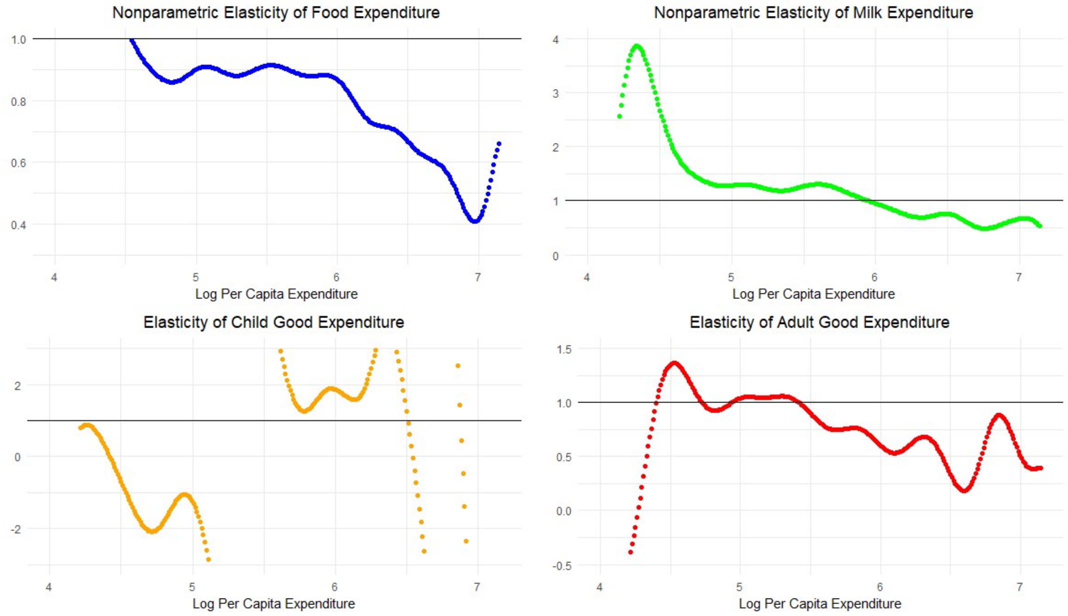

Figure 6: Nonparametric Elasticities

2.6. Extension

Section 3.6 extends the research from (Bhalotra & Attfield, 1998) by performing an outlier analysis and validating their findings in a subsample where outliers are deleted. Outliers are identified with the squared robust Mahalanobis distance. The algorithm is provided in the package aq.plot (Adjusted Quantile Plot).

Figure 7 visualizes the outliers in the dataset. Figure 8 and Figure 9 provide parametric and semiparametric Engel curves for the subsets where outlying values are removed. Table 3 to table 6 provide regression results for the parametric and nonparametric regression results.

(Bhalotra & Attfield, 1998, S. 463) are questioning the linearity in logarithmic form for food Engel curves which previous literature has identified. Correcting the dataset for outliers, the identified quadratic relationship loses its significance and the effect diminishes. This effect is contrary to the assumptions from

(Bhalotra & Attfield, 1998) and does verify previous research. Outliers are without any doubt valid data observations but these observations deviate from the majority. It's more a question of data preparation than of econometrics if outliers should be included into research but all in all, effects which diminish after outlier reduction can hardly be considered robust. The quadratic form of the Engel curve for milk and milk products consists and these kind products are identified as luxury goods for low income households as well as child goods and adult goods.

2.6.1. Outlier Analysis

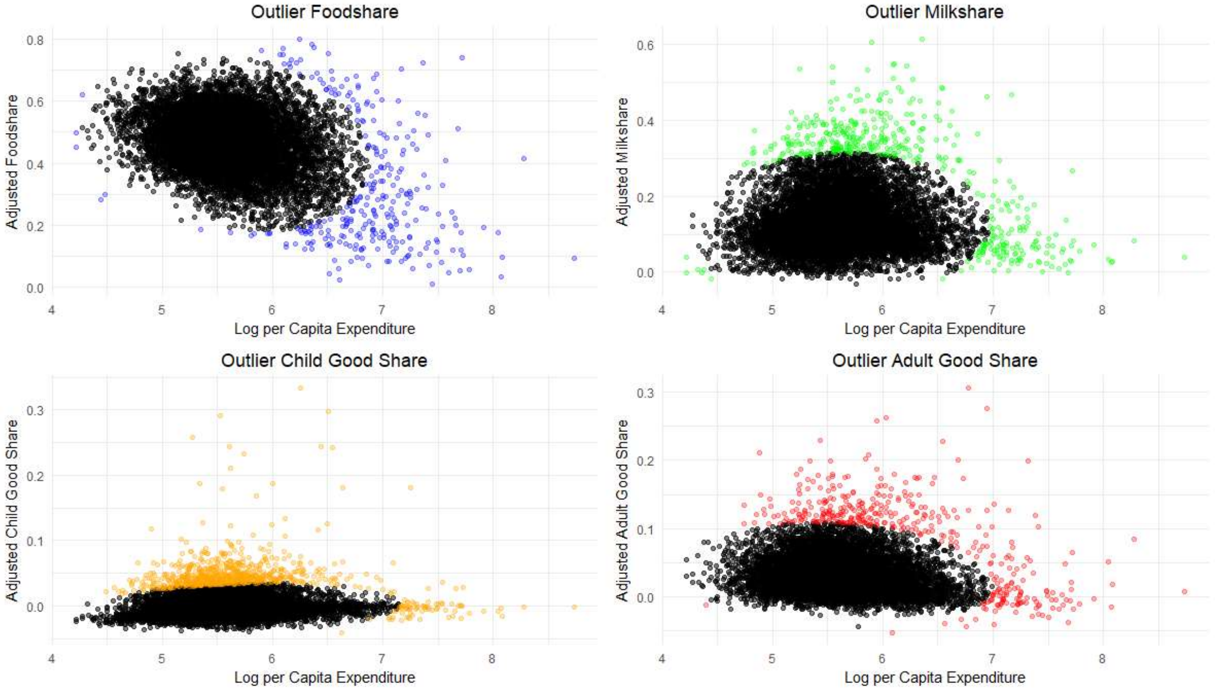

Figure 7: Outliers

2.6.2. Engel Curves and Regression Tables

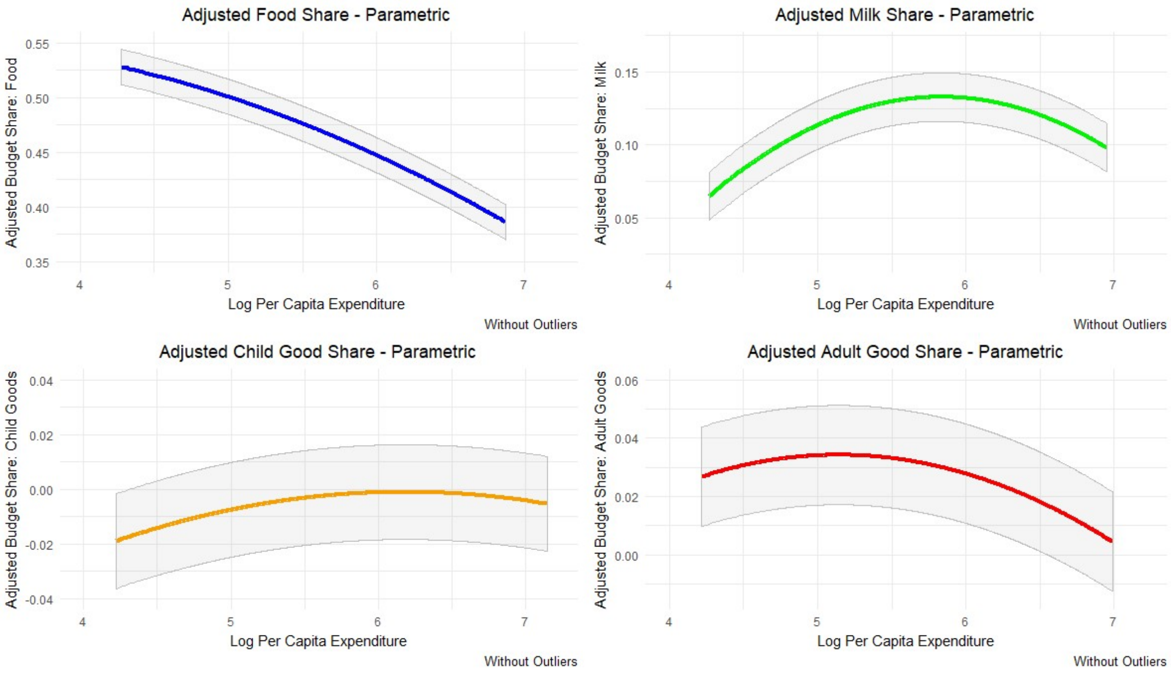

Figure 8: Parametric Engel Curves without Outliers

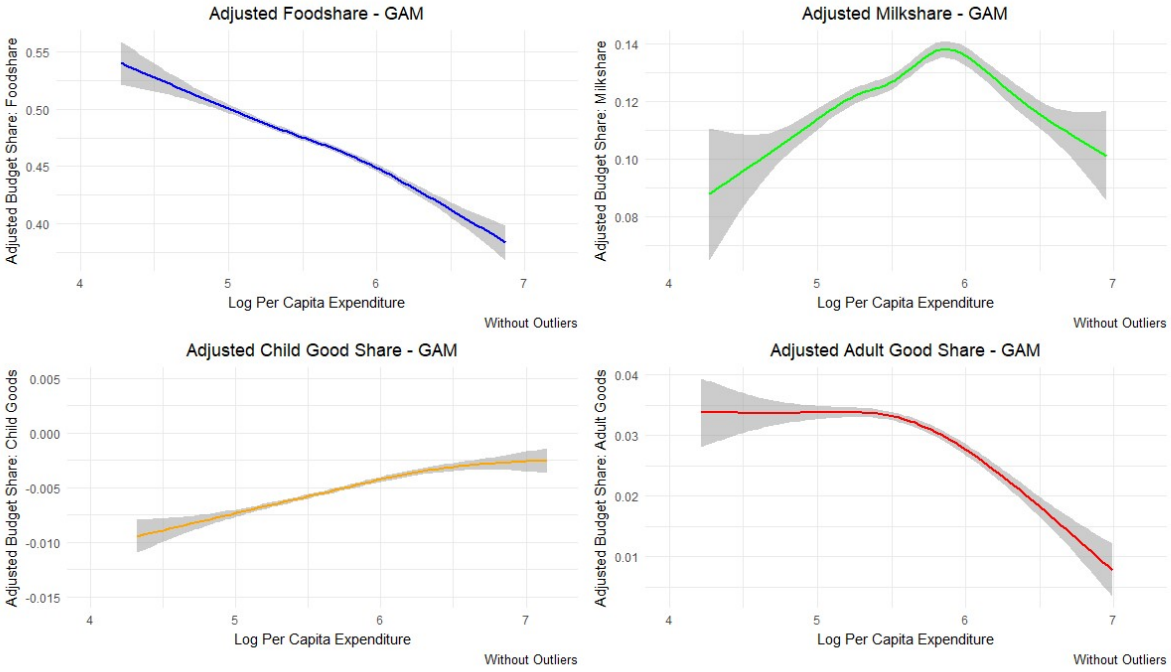

Figure 9: Nonparametric Engel Curves without Outliers

Table 3: Parametric Foodshare without Outliers

	Adjusted Foodshare
Log Expenditure	0.045
	t = 1.034
Log Expenditure Square	-0.009
	t = -2.324**
Observations	9,455
R^2	0.065
Adjusted R^2	0.064
Residual Std. Error	0.087 (df = 9452)
F Statistic	326.876*** (df = 2; 9452)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

Table 4: Parametric Milkshare without Outliers

	Adjusted Milkshare
Log Expenditure	0.324
	t = 10.737***
Log Expenditure Square	-0.028
	t = -10.436***
Observations	9,300
R^2	0.015
Adjusted R^2	0.015
Residual Std. Error	0.064 (df = 9297)
F Statistic	72.399*** (df = 2; 9297)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

Table 5: Parametric Adult Good Share without Outliers

	Adjusted Adult Good Share
Log Expenditure	0.089 t = 7.981***
Log Expenditure Square	-0.009 t = -8.803***
Observations	9,349
R^2	0.032
Adjusted R^2	0.031
Residual Std. Error	0.024 (df = 9346)
F Statistic	152.557*** (df = 2; 9346)
Notes:	***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Table 6: Parametric Child Good Share without Outliers

	Adjusted Child Good Share
Log Expenditure	0.058 t = 10.656***
Log Expenditure Square	-0.005 t = -9.839***
Observations	8,993
R^2	0.034
Adjusted R^2	0.033
Residual Std. Error	0.013 (df = 8990)
F Statistic	156.231*** (df = 2; 8990)
Notes:	***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

3. List of References

Bhalotra, S., & Attfield, C. (1998). Intrahousehold resource allocation in rural Pakistan: A semiparametric analysis. *Journal of Applied Econometrics*, *13*(5), 463–480. https://doi.org/10.1002/(SICI)1099-1255(1998090)13:5<463::AID-JAE510>3.0.CO;2-3

Härdle, W. (1987). Resistant Smoothing Using the Fast Fourier Transform. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *36*(1), 104.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, *56*(4), 931. https://doi.org/10.2307/1912705

4. Appendix

```
> sessionInfo()
R version 3.4.1 (2017-06-30)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows >= 8 x64 (build 9200)

Matrix products: default

locale:
[1] LC_COLLATE=German_Germany.1252
[2] LC_CTYPE=German_Germany.1252
[3] LC_MONETARY=German_Germany.1252
[4] LC_NUMERIC=C
[5] LC_TIME=German_Germany.1252

attached base packages:
[1] grid      stats     graphics  grDevices utils
[6] datasets  methods   base

other attached packages:
 [1] survey_3.32-1      survival_2.41-3
 [3] Matrix_1.2-10      stargazer_5.2
 [5] mvoutlier_2.0.8    sgeostat_1.0-27
 [7] KernSmooth_2.23-15 mgcv_1.8-17
 [9] nlme_3.1-131       ggplot2_2.2.1
[11] foreign_0.8-69     dplyr_0.7.1

loaded via a namespace (and not attached):
 [1] mclust_5.3            Rcpp_0.12.12
 [3] vcd_1.4-3             mvtnorm_1.0-6
 [5] lattice_0.20-35       class_7.3-14
 [7] zoo_1.8-0             assertthat_0.2.0
 [9] lmtest_0.9-35         VIM_4.7.0
[11] R6_2.2.2              plyr_1.8.4
[13] MatrixModels_0.4-1    stats4_3.4.1
[15] pcaPP_1.9-72          e1071_1.6-8
[17] rlang_0.1.1           lazyeval_0.2.0
[19] diptest_0.75-7        minqa_1.2.4
[21] data.table_1.10.4     SparseM_1.77
[23] kernlab_0.9-25        car_2.1-5
[25] nloptr_1.0.4          robCompositions_2.0.5
[27] labeling_0.3          splines_3.4.1
[29] lme4_1.1-13           sROC_0.1-2
[31] munsell_0.4.3         compiler_3.4.1
[33] pkgconfig_2.0.1       nnet_7.3-12
[35] tibble_1.3.3          reshape_0.8.6
[37] rrcov_1.4-3           laeken_0.4.6
[39] MASS_7.3-47           GGally_1.3.1
[41] gtable_0.2.0          magrittr_1.5
[43] scales_0.4.1          flexmix_2.3-14
[45] bindrcpp_0.2          sp_1.2-5
[47] robustbase_0.92-7     pls_2.6-0
[49] boot_1.3-19           RColorBrewer_1.1-2
[51] tools_3.4.1           fpc_2.1-10
[53] glue_1.1.1            trimcluster_0.1-2
[55] DEoptimR_1.0-8        cvTools_0.3.2
[57] parallel_3.4.1        pbkrtest_0.4-7
[59] colorspace_1.3-2      cluster_2.0.6
[61] prabclus_2.2-6        bindr_0.1
[63] quantreg_5.33         modeltools_0.2-21
```