Bonn Econ Discussion Papers

Discussion Paper 22/2009

Engel's Law Reconsidered

by

Manisha Chakrabarty and Werner Hildenbrand

September 2009

Bonn Graduate School of Economics
Department of Economics
University of Bonn
Kaiserstrasse 1
D-53113 Bonn

Financial support by the
Deutsche Forschungsgemeinschaft (DFG)
through the
Bonn Graduate School of Economics (BGSE)
is gratefully acknowledged.

Deutsche Post World Net is a sponsor of the BGSE.

Engel's Law Reconsidered

Manisha Chakrabarty
Indian Institute of Management Calcutta, India

Werner Hildenbrand
University of Bonn, Germany

18.09.2009

1. Introduction

"Of all empirical regularities observed in economic data, Engel's Law is probably the best established...",Houthakker (1957). This claim has been repeated frequently and was never seriously questioned. Given this unanimity of opinion, naturally, one expects that there is no ambiguity in the definition of Engel's Law. Yet, as we shall show, this is not the case.

Our original motivation for the present study was purely historical. We wanted to know exactly what Engel contributed in his two famous publications of 1857 and 1895, since there are concepts and claims attributed to Engel in the economic literature which certainly have nothing to do with Engel's thinking or writings.

Engel analysed in his publication "Die Productions- und Consumtionsverhältnisse des Königreichs Sachsen"(1857) income-expenditure data for Belgian working class households, which were collected by Ducpetiaux (1855). Engel summarized his insights, he speaks of a genuine step of induction ["*auf dem Wege ächter Induction*"] by the following statement that later has been called Engel's Law:

(i) "*je ärmer eine Familie ist, einen desto grösseren Antheil von der Gesamtausgabe muss zur Beschaffung der Nahrung aufgewendet werden"* ["the poorer a family, the greater the proportion of its total expenditure that must be devoted to the provision of food"][1].

How this statement should be interpreted? Clearly, it refers to income or total expenditure and budget shares for food (food share for short) for different households in a given population at a given period and not to changing (different) income of a given household. Food share is sometimes defined as consumption expenditures in current prices on food items divided by income but also by consumption expenditure on food divided by 'total expenditure' which is defined as expenditures on a well-specified large class of consumption goods and services.

[1] Translation by Stigler(1954), all other translations are by the authors.

Engel‘s statement (i), taken literally, claims a decreasing functional relationship between family income and food share. But this is not what Engel wanted to assert since he amended on p.30 in his publication (1857):
(ii) "*Freilich wird es auf Einzelne angewendet, nicht unter allen Umständen seine volle Richtigkeit behaupten, um so mehr aber in seiner Anwendung auf Bevölkerungsgruppen"* [Admittedly, applied to single households, it (statement (i)) will not be fully correct in all circumstances, yet it will be correct, if applied to groups of households**.]**

Therefore, it is clear that Engel's statement (i) concerns the bivariate distribution of income x_h and food share y_h across a population H of households h, and the statement (i) together with the amendment (ii) express a *negative stochastic association* or dependence of x and y, that is to say, large (small) values of x "tend" to be associated with small (large) values of y. Precise definitions of such concepts of negative stochastic association of a bivariate distribution were not available in the literature at the time when Engel formulated his Law. This explains Engel's somewhat unsatisfactory formulation of the Law by his statements (i) and (ii). It is easy, however, to fill this gap, since more than 150 years after Engel, one can find many well-defined concepts of negative stochastic association for bivariate distributions in the statistical literature, e.g., Kruskal (1958) or Lehmann (1966). In section 2.1 we present four different such concepts: negative Kendall's τ, negative quadrant dependence (Lehmann, 1966), decreasing population regression function and stochastically decreasing conditional food share distribution functions (Tukey, 1958). In the economic literature one generally defines Engel's Law by a decreasing regression. This is, as we shall show, the least useful and least informative property.

Engel analysed also other categories of consumption e.g., clothing, housing etc. However, *Engel's Law always refers to food share and to a comprehensive population* defined largely by a certain geographical area or nationality, e.g, the population of Belgium working class families or the population of private households in Saxony. Engel did not assume that all households in the population are facing the same prices, nor that the households are identical in certain household characteristics.

The goal of Engel in both articles (1857) and (1895) was to describe as carefully as possible an *observed empirical regularity*, he did not aim to explain deductively his Law by some postulates on individual household behaviour. This became the mainstream approach after Allen and Bowley (1935) [more details on this point at the end of the introduction and in section 2.2).

Once a property of stochastic association for the population distribution is well defined, one can turn to the more difficult problem of how to make inferences about this property from random samples of the population (see section 4). It is this step, that is to say, going form the specific (property of the sample, i.e, data) to the general (property of the population distribution) which Engel might have had in mind when he wrote on page 8 and page 28 in (1857) that he obtained his Law by a 'genuine step of induction'. This does not mean, of course, that we claim that Engel based his 'step of induction' on a statistical theory of random sampling (hence a probabilistic model) but rather he relied on

the idea of purposive sampling, i.e., the data is considered as "representative" for the population distribution. This explains why in the empirical literature on Engel's Law one generally does not distinguish explicitly between the population distribution and the sample, i.e., data. This distinction, however, is necessary for testing statistical hypotheses.

A property of a population distribution might be called a law if the class of populations is clearly specified for which serious attempts of falsification of the claimed property have not been successful.

In interpreting an observed stochastic association between two variables X and Y, as claimed by Engel for income and food share, one is constantly faced with the question or objection whether the association between X and Y, in fact, is really (intrinsically) due to an association of each with a third variable Z. For this reason, most economists – yet certainly not Engel! - speak of Engel's Law only if it refers to stratified subpopulations where "all" observable explanatory variables other than income (e.g., prices, household attributes and demographics) are held constant. For recent examples of insisting on this ceteris paribus clause see the articles in the New Palgrave (2^{nd} ed.) by Browning (2008) and Lewbel (2008). However, without an explicit theory of individual household behaviour, which specifies a complete set of explanatory variables, the above objection remains since without such a theory there is always the possibility that a relevant explanatory variable is missing. For this reason, Allen and Bowley (1935) argued that the analysis of family budgets should be linked to micro-economic theory that is to say, to a model of individual household behaviour and to the derived theoretical concept of an individual demand function.

We want to emphasize, however, that this link to micro-economics, which we shall discuss in section 2.2, is not in the spirit of Engel. Nowhere in Engel's thinking or writings, has occurred the concept of an individual demand function. Utility-based demand functions were developed by Jevons and Walras in 1870's. In 1857 Engel could not have known this concept. Also in his later contribution (1895) he did not use it. The micro-economic theory of a consumer and the notion of a utility-based demand function was mostly used as a logical tool to explore conceptually the properties of alternative market organization and economic policy. In empirical work utility theory played a less important role. Theorists in general were not engaged in empirical work. "The utility theorists as a class have always expressed the greatest enthusiasm for empirical work compatible with abstention from it" (Stigler, 1954).

The paper is organized as follows: section two describes four concepts of negative stochastic association and the relationship of Engel's Law with individual behaviour. In section three, the data from Ducpetiaux (1855) is presented with the re-interpretation of Engel's original work and the statistical analysis of the empirical literature on Engel's Law is presented. In section four two modern data sets which differ in place and time are used to provide empirical support for the concepts of association, which are presented in section 2.1.. Finally, in section five conclusion are drawn.

2.1 Concepts of stochastic association

In defining a notion of negative stochastic association of a general bivariate distribution which might serve as a candidate for defining Engel's Law, it is convenient to consider a pair (X,Y) of random variables whose joint distribution μ is the population distribution of income and food share across the population in question. Then, a realization of X and Y can be interpreted as income and food share of a randomly drawn household from the population.

1. Negative Kendall's τ (1938)
The pair (X,Y) of random variables (or its joint distribution μ) is *negatively associated in the sense of Kendall if*

(1) $$P\{(X_i - X_k)(Y_i - Y_k) > 0\} < P\{(X_i - X_k)(Y_i - Y_k) < 0\}$$

that is to say, if one chooses two households, say i and k, at random from the population under discussion, then observing *discordance*, i.e., $(x_i - x_K)(y_i - y_K) < 0$ is more likely than observing *concordance*, i.e., $(x_i - x_K)(y_i - y_K) > 0$.

More generally, one can define a measure of the degree of stochastic association for any bivariate distribution μ, called Kendall's τ (tau), by

$$\tau(\mu) := P\{(X_i - X_k)(Y_i - Y_k) > 0 \mid X_i \neq X_k \text{ and } Y_i \neq Y_k\} - P\{(X_i - X_k)(Y_i - Y_k) < 0 \mid X_i \neq X_k \text{ and } Y_i \neq Y_k\}.$$

It follows that $-1 \leq \tau(\mu) \leq +1$. If X and Y are independent, then $\tau(\mu) = 0$ (not the converse) and $\tau(\mu) = -1(+1)$ implies that there is a decreasing (increasing) functional relationship between X and Y. Kendall (1938) gave a very thorough discussion of $\tau(\mu)$ and its associated sampling theory. The basic notion goes back to Fechner (1897) and Lipps (1906).

2. Negative quadrant dependence (Lehmann (1966))

The pair (X,Y) of random variables (or its joint distribution μ) is *negatively associated in the sense of Lehmann* (or *negatively quadrant dependent*) if

(2) $$P\{Y \leq y \mid X \leq x\} \leq P\{Y \leq y\}, \text{ for all x, y}$$

That is to say, the knowledge of X being small (i.e., $X \leq x$) decreases the probability of Y being small (i.e., $Y \leq y$). Or in other words, if one draws at random a household first in the entire population and second in the subpopulation of all households with income less than x, then, the probability of observing a food share less than y in the first case is larger than in the second case.

There is a very useful characterization: *negative quadrant dependence is equivalent with* $\operatorname{cov}(\varphi(X), \psi(Y)) \leq 0$

for any non-decreasing functions φ *and* ψ provided covariance is defined. It follows (2) implies (1), yet not the converse, and any linear least square fit of $\psi(Y)$ on $\varphi(X)$ is non-increasing.
A strenthening of (2), that we call *monotone negative quadrant dependence*, is defined by

(2a) $P\{Y \le y \mid X \le x\}$ is **non-decreasing in x for every y,**
i.e. the conditional distribution function $F_Y(\cdot \mid X \le x_1)$ of Y stochastically dominates $F_Y(\cdot \mid X \le x_2)$ if $x_1 < x_2$, i.e., the graph of $F_Y(\cdot \mid X \le x_1)$ lies below the graph of $F_Y(\cdot \mid X \le x_2)$. This implies that $E(Y \mid X \le x)$ is non-increasing in x.

3. Decreasing regression

The pair (X,Y) of random variables (or its joint distribution μ) has a *decreasing regression* of Y on X if

(3) $E(Y \mid X = x_1) \ge E(Y \mid X = x_2)$
for any $x_1 < x_2$ in the range of X, i.e., the mean food share of all households with income x_1 is larger or equal to the mean food share of all households with income x_2.

This is the most popular definition of Engel's Law in the economic literature. However, property (3) alone is not interesting from a distributional point of view since a decreasing regression does neither imply (1), a negative Kendall's τ, nor (2), negative quadrant dependence. In other words, there are different distributions μ_1 and μ_2 with identical decreasing regression such that μ_1 satisfies property (1) or (2) yet μ_2 does not. The reason is that the conditional expectation $E(Y \mid X = x)$ does not give sufficient information on the conditional distribution of Y given $X = x$. This suggests to extend the monotonicity property of $E(Y \mid X = x)$ in (3) to the conditional distribution function $F_Y(\cdot \mid x)$ which is defined by $F_Y(y \mid x) = P\{Y \le y \mid X = x\}$, for all y.

4. Stochastically decreasing conditional distribution functions (Tukey (1958))

The pair (X,Y) of random variables (or its joint distribution μ) is *negatively associated in the sense of Tukey* (1958) if the conditional distribution function $F_Y(\cdot \mid x)$ of Y given $X = x$ is stochastically decreasing in x, i.e., for any $x_1 < x_2$ in the range of X,

(4) $P\{Y \le y \mid X = x_1\} \le P\{Y \le y \mid X = x_2\}$ for all y.
Property (4) implies (3), yet (4) is much more restrictive, it also implies (2), (2a) and (1) (see Lehmann (1966) Lemma 4 and Corollary of Lemma 3).

The decisive question now is which of these candidates (or possibly alternatives e.g. Gini's or Spearman's measure of concordance) should be chosen to define Engel's Law? The answer is obvious. The chosen notions of association must have satisfactory empirical support. We discuss this important point in section 4. One expects that properties (1) and (2) will pass the test, yet the monotonic versions (2a), (3) or (4) might not hold over the whole domain of the income distribution.

Remark: If one would have *a priori* knowledge on the functional form of the population distribution, then some of the above concepts might be equivalent. For example, in the case of a bivariate normal distribution each of the four concepts of negative association is equivalent with a non-positive correlation coefficient. Another example which is often considered in the statistical literature is the following case: the random variable y is defined by $Y = m(X) + \varepsilon$ where X and ε are independent random variables. Then, properties (3) and (4) are equivalent. If $m(X)$ is linear, then (X,Y) is either negatively or positively associated in the sense of Tukey (4), depending on the sign of cov(X,Y).

2.2 Engel's Law and Micro-Economics

In the introduction we explained why Allen and Bowley (1935) argued that the analysis of family budgets should be linked to micro-economic theory, that is to say, to a model of individual household behaviour. Naturally, Allen and Bowley (and all their followers) model individual behaviour by the hypothesis of preference (utility) maximization under budget constraints. From this hypothesis one can derive a micro-economic behavioural relation $s(.)$ of the form

$y^h = s(x^h, v_1^h, v_2^h, ...)$

where x^h denotes income of household h, the vector $v^h = (v_1^h, v_2^h, ...)$ consists of all parameters, other than income, which define the maximization problem and y^h denotes food share of household h. In the simplest case of an atemporal decision, one has $v^h = (p^h, \leq^h)$, where p^h denotes the price system which household h faces and $\leq^h$ is the preference relation of household h.

In this micro-economic setting, a population of households is described by the behavioural relation $s(.)$ and a joint distribution π of the explanatory variables x and y. Let (X,V) be a generic random variables with joint distribution π. The bivariate distribution μ of income and food share which is relevant for Engel's Law, is then given by the distribution of $(X, s(X,V)) \equiv (X,Y)$.

The following proposition answers the question which hypotheses on individual behaviour imply Engel's Law and conversely, what are the implications of Engel's Law on the postulated model of individual behaviour.

Proposition: Consider a micro-economically defined population of households $[s(.),(X,V)]$ *such that* X *and* V *are stochastically independent, i.e.income* x *and the vector* v *of all explanatory variables other than income are independently distributed across the population.*[2]

(a) If all households' budget share functions $s(.,v)$ *for food are non-increasing, then the conditional distribution function of food share given income is stochastically decreasing in income, i.e., the joint distribution* $(X, s(X,V))$ *of income and food share is negatively associated in the sense of Tukey. Conversely,*

(b) If the range of X *is an interval* $[a,b]$*, the range of* V *is finite and households' budget share functions* $s(.,v)$ *are continuous, then all budget share functions for food are non-increasing on (a,b) if the conditional distribution functions of food share are stochastically decreasing.*

In both conclusions one can not drop the independence assumption of (X,V).

Proof: Independence of X and V implies that the conditional distribution function $F_Y(.\mid x)$ *of* $Y = s(X,V)$ given $X = x$ is the distribution function of the random variable $s(x,V)$. Indeed,

$$P\{s(X,V) \le y \mid X = x\} = \frac{P\{s(x,V) \le y \text{ and } X = x\}}{P\{X = x\}} = P\{s(x,V) \le y\}.$$

(a) If $s(.,v)$ is non-increasing, one obtains $P\{s(x_1,V) \le y\} \le P\{s(x_2,V) \le y\}$ for $x_1 < x_2$ in the range of X and all y, which proves the claim.

(b) Since $s(.,v)$ is continuous on (a,b) for every $x_1 \in (a,b)$ there is $x_2 \in (a,b)$, $x_1 < x_2$ such that $x \in (x_1,x_2)$ and $s(x_1,v_i) < s(x_1,v_j)$ implies $s(x_2,v_i) < s(x_2,v_j)$. Note there are only finitely many v_k's. If there exists $x \in (x_1,x_2)$ such that for some v_k, $s(x_1,v_k) < s(x,v_k)$ then one obtains a contradiction to the assumption that $F_y(\cdot \mid x)$ is stochastically decreasing in x. (Note, that the weaker property of a decreasing regression would not lead to a contradiction.) Hence $s(x_1,v) \ge s(x,v)$ for all $x \in (x_1,x_2)$ and all v, which proves that $s(\cdot,v)$ is non-increasing in (a,b).

Remark: The proposition is a purely theoretical result without empirical content. This, unfortunately, is typical for many results in economic theory! Part (a) of the proposition can not be viewed as a micro-economic explanation (deductive derivation) of Engel's Law – even if one accepts as evident the assumption on households' behaviour - since it refers to a population which can not be identified.

---

[2] *This independence assumption does not imply that (X,Y) satisfies a standard assumption made in regression analysis, i.e.,* $Y = m(X) + \varepsilon$ where X and ε are stochastically independent. For this conclusion one needs that $s(\cdot,\cdot)$ is separable, i.e., $s(x,v) = s_1(x) + s_2(v)$.

Indeed, typically some of the explanatory variables in the vector V are unobservable (e.g., preference parameters or expectations). The standard practice to consider subpopulations by stratifying on "suitably chosen" observable household attributes (demographics) does not resolve the difficulty, since there is no guarantee that (X,V) becomes independent if conditioned on the *ad hoc or "suitably"* chosen households attributes. In any case, Engel did not formulate his Law for such hypothetical homogeneous subpopulation but rather for comprehensive populations which are defined by geographical area. For the same reason, part (b) of the proposition does not allow the conclusion that households' budget share functions for food can be assumed to be non-increasing since it is not clear at all whether the required strong negative association in the proposition holds for the subpopulation in question(see section 4).

3. Empirical support of Engel's Law in the literature

3.1. Engel (1857) and (1895): An early non-parametric statistician

"The first and most famous of all statistical analyses of budgets was made in 1857" [Stigler (1954), p. 209]. We can add, it is also the first non-parametric statistical analysis of budgets [Härdle (1990), p. xi].

Engel analysed income-expenditure survey data for Belgian working class families, which were collected by the Provincial Statistical Commission and were procured under the direction of Edouard Ducpetiaux (1855). The commission defined three socio-economic categories: (1) families dependent upon public assistance; (2) families just able to live without such assistance; and (3) families in comfortable circumstances. For each commune in the nine provinces one family of each category were chosen, resulting in 199 families. Moreover, most families consisted of a father, mother and four children whose ages were 16, 12, 6 and 2 respectively. This family composition was considered by the commission as 'typical' for Belgium resulting in a 'representative' data-set. This view was criticized by Engel (1895) p. 23. The published data in Ducpetiaux(1855) contained for all 199 families' information on income, expenditure on food items and many other consumption goods and services as well as for 153 families the belonging of the social-economic category.

Discussing the quality of the survey data Ducpetiaux observes [1855, p. 17] that for many families total expenditure exceeds income (for 19 families out of 199, even food expenditure exceeds income) since in some communes not actual expenditure on food but the need for food is reported. "Cet fait provient sans doute de ce que les recettes on été calculées d'après les salaires réelment gagnés, tandis que les dépenses on été indiquées d'après les besoin constatés des ménages, abstraction faite de ce que ceux-ci dépensent effectivement."

By modern standards, Ducpetiaux's data is not satisfactory, empirical results based on this data should be taken with caution! However, they were among the first. Interest in family budgets had its beginning in England at the close of the eighteenth century(see

D.W. Douglas, "Family Budgets", Encyclopaedia of the Social Sciences, London): Arthur Young(1771), David Davies(1795), Sir Frederick Eden(1797) or Le Play(1855). Engel analyzed in (1857) Ducpetiaux's data only in tabulated form (see tables 1 and 4-6 in Engel (1857))**.** We ignore whether Engel also looked at a graphical representation. If so, he would have obtained the following scatter plots 3.1a and 3.1b.

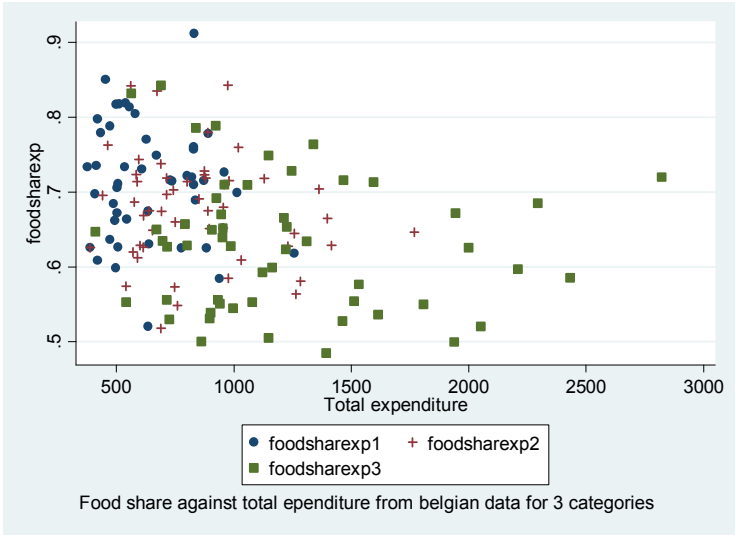


Figure 3.1a: Food share against total expenditure

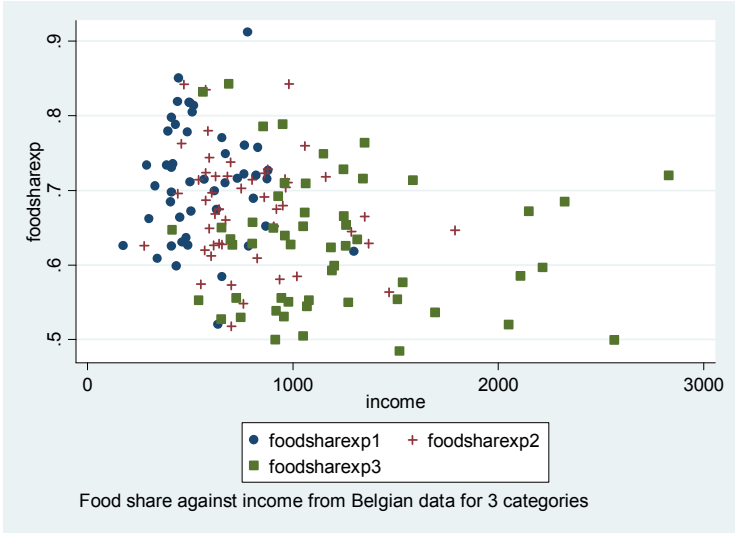


Fig 3.1b: Food share against income

* represents first socio-economic category, + represents 2nd category and ■ represents 3rd category.

These scatter plots of Ducpetiaux' data suggest some degree of 'negative association' between income and food share. In fact, Kendall's τ in figures 1a and 1b are -0.19 and -

0.21 respectively. Yet, as we explained in the introduction, Kendall's τ or other measures of association, as discussed in section 2.1, were not known in 1857.

To give empirical support to his statements (i) and (ii) (see Introduction), Engel computed "category food share", i.e., the ratio of mean food expenditure and mean income (or mean total expenditure) across each of the three categories, and obtained the following table

Table 3.1a: Reproduced from tables 4 and 6 of Engel (1857)

category	Category 1	Category 2	Category 3
mean income	565.0	796.7	1197.8
total expenditure	648.7	845.5	1214.5
food share out of income	81.4	71.5	63.3
food share out of total expenditure	70.9	67.4	62.4

Thus, as Engel claimed in his statement (ii) (see Introduction), *the smaller the mean income of the category, the larger its food share*. Note, however, that the three categories are socio-economic groups. These are not income classes since their income-ranges overlap considerably, as can be seen from figures 3.1a and 3.1b. Engel was aware that his findings of table 1 are not a satisfactory support for his law. Indeed, he wrote in his publication in 1895, p. 36: "In meiner Abhandlung aus dem Jahre 1857 habe ich nachgewiesen, zu welchem Schlusse diese Ergebnisse berechtigen. Die Berechtigung wird unleugbar grösser, wenn man das subjective Ermessen, ob man es mit einer dürftigen, auskommenden oder sparfähigen Familie zu thun habe, ganz bei Seite lässt und die Klassifiktation der Budgets lediglich sowohl nach der Höhe der Jahreseinnahmen als auch nach der Höhe der Jahresausgaben jeder einzelnen Familie vornimmt." [In my study of the year 1857 I have shown to which conclusion these results lead. The justification becomes undeniably stronger if one puts aside all together the subjective judgment of whether one deals with a family classified as on relief, poor but independent or comfortable, and instead classifies the family exclusively according to income per year as well as the level of expenditure per year.]

Engel defined first eighteen income classes, but later reduced it to five, in order to have more observations in each class. Then he computed, analogously as for the three socio-economic categories in (1857), for every income class the 'food share', and obtained again the result: *the lower the income class the higher its food share.* This is described in the following table 3.1b.

Table 3.1b: **Reproduced from Table 2 of Engel (1895)***

income class	$x \leq 600$ I_1	$600 < x \leq 900$ I_2	$900 < x \leq 1200$ I_3	$1200 < x \leq 2000$ I_4	$x > 2000$ I_5
mean total expenditure	501.63	762.09	1010.44	1460.99	2306.41
mean food expenditure	356.07	516.66	665.92	904.95	1444.27
income class 'food share'	70.98	67.79	65.90	61.94	62.62
mean food share	70.89	67.68	65.90	62.35	62.08
Numbers of observations	42	70	46	35	6

*Number of observations are reproduced from table 2 of Engel (1895) based on total expenditure according to Ducpetiaux report. Mean values are calculated by taking total expenditure as reported in Ducpetiaux, not as reported in Engel (1857).

Comment: To match the number of observations within each class, as reported in Engel (1895), one has to take total expenditure according to Ducpetiaux, not as reported in Engel (1895). There are some discrepancies between Engel's reported total expenditure (1857) and Ducpetiaux reported total expenditure [52 out of 199]. Engel's reported values [Summe der Ausgaben pro familie, table 2 page 38 1895 article] are wrong even when one takes total expenditure data as reported by himself in 1857.

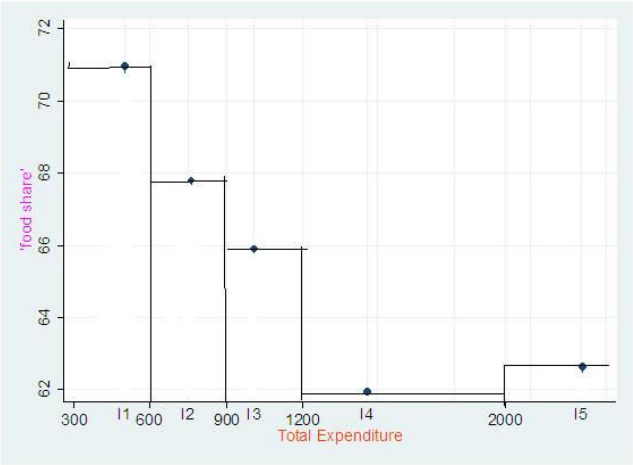


Figure 3.1c: Graphical Representation of income class food shares of Table 2.

We now claim that the step function in Figure 3.1c which we call the Engel smoother was viewed by Engel as a non-parametric estimator for the unknown population regression function. Note, that the step function in figure 3.1c is not a regressogram since the mean of households' food shares across an income class (which defines the regressogram) is different from the category 'food share', calculated as the ratio of mean food expenditure and mean income. The difference, however, is small for narrowly defined income classes, since

$$\Delta_I = \left| \frac{1}{\#I} \sum_{h \in I} \frac{c_h}{x_h} - \frac{\sum_I c_h}{\sum_I x_h} \right| \le \left| \mathrm{cov}_I\left(\frac{c_h}{x_h}, \frac{x_h}{x_I}\right) \right| \le \left(\mathrm{var}\frac{c_h}{x_h}\,\mathrm{var}\frac{x_h}{x_I}\right)^{1/2} \text{ where } x_I = \frac{1}{\#I}\sum_I x_h .$$

The following figure shows Engel's smoother of figure 3.1c and the corresponding regressogram where $\Delta_1 = 0.09, \Delta_2 = 0.11, \Delta_3 = 0.00, \Delta_4 = 0.41, \Delta_5 = 0.54$. Note, the sample regressogram is decreasing while Engel's smoother is not decreasing in the last step.

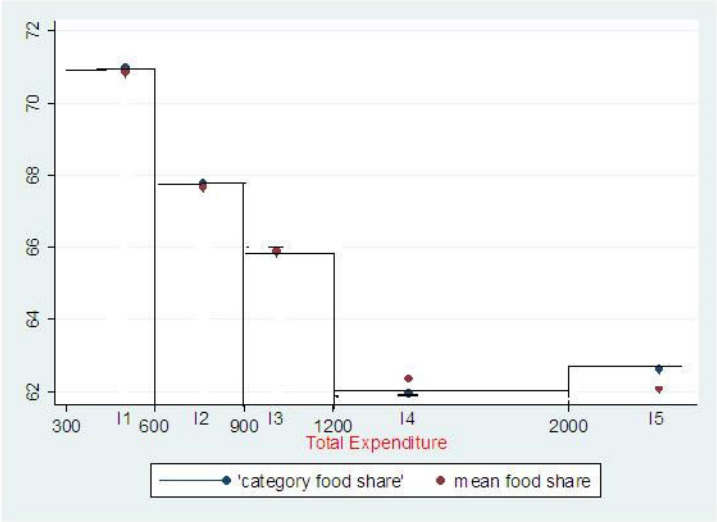


Figure 3.1d: Comparison of Regressogram and Engel's smoother

Thus asymptotically (for more and more narrowly defined income classes), there is no difference between Engel's smoother of figure 3.1c and the corresponding regressogram. Collomb (1977) and Lecoutre (1983) have shown that for suitably chosen income classes the regressogram of a random sample is a non-parametric estimator for the population regression function with good statistical properties. Since a regressogram is a special Kernel estimate [Härdle (1990), p. 67], we have linked Engel's statistical analysis to modern non-parametric Kernel estimation.

In summary, Engel's statistical analysis in support of his law is non-parametric. For the given data set of Ducpetiaux he computed his version of a regressogram, the Engel smoother, which he considered as an estimator for the population regression. Of cause, Engel could not show that this estimator has good statistical properties. Note, the Ducpetiaux' data set is not a random sample from the population distribution. Engel probably did not feel the need to distinguish between the observed sample property and the claimed population property, since the data was typically viewed as 'representative'. We emphasize that Engel never assumed a parametric functional form of the population regression. This is the reason why Engel did not use any curve fitting method that were known at his time, in particular, the method of linear least square- which Engel should

have known, since this method was published 50 years earlier [Legendre (1805) and Gauss (1809)]. The non-parametric approach to study Engel's Law was given up in the first part of the 20' century in favor of parametric (linear) regression analysis, most likely for computational reasons. After high speed computers became available non-parametric methods – more than 100 years after Engel – were used again for analysing large cross section data (see section 3.3).

Remark:
In contradiction to our claim in the summery, one can find in the literature (e.g. Houthakker(1957)) the view that Table 8 in Engel(1857) is evidence for a parametric analysis, since the scatter plot of food expenditure c^h against income x^h in a double logarithmic scale seems to lie on a straight line.

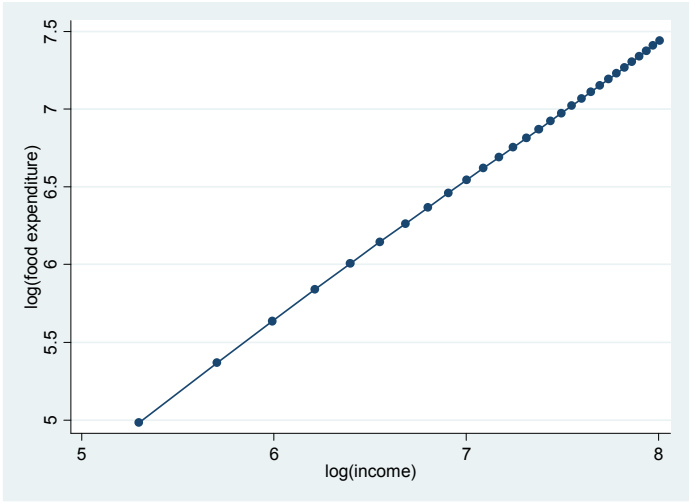


Figure 3.1e

Engel wanted to illustrate his Law by his Table 8. However he did not explain how, if at all, he obtained Table 8 from Ducpetiaux's data-set [see Perthel(1975) for an attempt to solve this puzzle]. There are at least two arguments against the above view. First the scatter plot $(\log x^h, \log c^n)$ in Figure 3.1e does not really lie on a straight line – even if one allows for random errors. This can easily be seen by looking at a scatter plot of $\log y^h = \log(\frac{c^h}{x^h})$ against $\log x^h$, which shows a clear nonlinear shape.

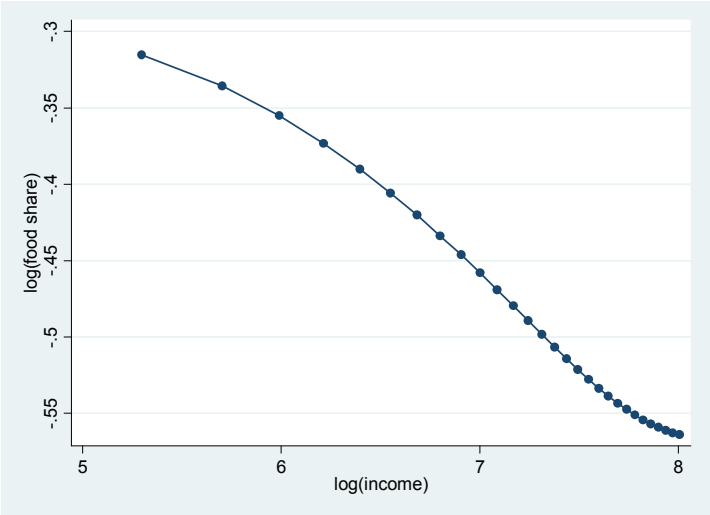


Figure 3.1f

Note that the scatter plot $(\log x^h, \log y^h)$ lies on a straight line if and only if $(\log x^h, \log c^n)$ does so.

Second if Engel would have had in mind a linear relation between log income and log food share why then he wrote on p. 30 "Das Gesetz, mit welchem man es hier zu thun hat, ist kein einfaches" [The Law, with which one has to deal here, is not a simple one.] and further more "… so entsprechen die Ausgaben folgender 8. Tabelle ziemlich nahe den Bedingungen des Gesetzes, obschon diese selbst noch nicht auf einem präzisen mathematischen Ausdruck gebracht werden konnten." [the data in the 8th Table represent rather well the conditions of the law, even so the law itself could not yet be expressed in a precise mathematical expression.]

3.2 Econometric Studies on Engel's Law: Parametric regression analysis

The early econometric literature on Engel's Law is bases on the linear least square regression model e.g, Allen and Bowley (1935), Working (1943), Prais and Houthakker (1955), Houthakker(1957) and Leser (1963). This literature can be summarized as follows: one selects a specification of a relationship between income x and food share y which is of the form

$$\psi(y) = a + b\varphi(x), x \in [\alpha, \beta]$$

where $\psi(y) = \tilde{y}$ and $\varphi(x) = \tilde{x}$ are known increasing transformations and a, b are unknown parameters. The above quoted literature differs in the choice of ψ and φ, e.g., in Allen and Bowley $\varphi(x) = -1/x,\ \psi(y) = y$, since they assume that food expenditure is linear in income; in Working $\varphi(x) = \log x,\ \psi(y) = y$; in Leser $\varphi(x) = x,\ \psi(y) = y$; and in Houthakker $\varphi(x) = \log x,\ \psi(y) = \log y$. The problem with this literature is the ad hoc choice of the functional form of the relationship between income and food share. There is no theory which justifies a particular choice. Computational simplicity alone is not sufficient.

Given a distribution μ of income and food share, one considers the **linear least square smoother** (fit) of the transformed distribution $\tilde{\mu}$ defined as the straight line $\tilde{y} = a^* + b^* \tilde{x}$ which minimizes the expression $E(\psi(Y) - a - b\varphi(X))^2 = \int (\tilde{y} - a - b\tilde{x}) d\tilde{\mu}$. It is well-known that

$$b^*(\mu, \psi, \varphi) = \frac{\mathrm{cov}(\varphi(X), \psi(Y))}{\mathrm{var}(\varphi(X))} \text{ and } a^*(\mu, \psi, \varphi) = E(\psi(Y)) - b^*(\mu, \psi, \varphi) E(\varphi(X)).$$

Obviously, there is no a priori reason why the conditional expectation $E(\tilde{Y} \mid \tilde{X} = x)$ should be linear in x on the whole range of the income distribution. If it happens to be the case[3] (possibly after truncation on an interval $[\alpha, \beta]$), then the regression $E(\tilde{Y} \mid \tilde{X} = x)$ is equal to the linear least square smoother. Invariably one obtains, for the data sets of income and food share which are analysed in the literature, and for the above specifications of (ψ, φ) - as well as for many other choices - that the linear least square smoother is decreasing. This is equivalent with $\mathrm{cov}_{\mu_s}(\varphi(x), \psi(y)) < 0$, where the covariance is taken with respect to the sample distribution μ_s.

This very robust empirical regularity, or its population distribution analogue, is often regarded in the literature as Engel's Law. Of course, for a clear definition, one ought to specify the class of transformations (ψ, φ), for which $\mathrm{cov}(\varphi(X), \psi(Y))$ is supposed to be negative. The larger this class of transformations, the stronger would be the law. Recall, a distribution μ is negatively quadrant dependent (definition 2, section 2.1) if and only if $\mathrm{cov}_{\mu}(\varphi(X), \psi(Y)) \leq 0$ for all increasing φ and ψ.

Certainly, however, the above empirical regularity alone is not sufficient to derive a satisfactory statistical support (not to say a test) for any of the four candidates of Engel's Law which we discussed in section 2.1. For this, one needs suitable a priori knowledge on the structure of the population distribution. But where should this knowledge come from?

Interestingly, a standard assumption of independence between the random variables X and ε in linear regression analysis such as $\psi(Y) = \alpha + \beta\varphi(X) + \varepsilon$ where α and β are unknown parameters, alone implies almost what one wants to show. Indeed, this assumption implies that $(\varphi(X), \psi(Y))$ and hence also (X, Y) is either negatively or positively associated in the sense of Tukey (Def. (4) section 2.1) as $\beta \leq 0$ or $\beta \geq 0$. Therefore the data are only needed to decide which case is prevailing and for this it is sufficient to know the sign of $\mathrm{cov}(X, Y)$.

---

[3] For the FES data it has been shown that it is very unlikely that any of the above transformations(φ , ψ) lead to a linear regression on the whole support of the income distribution. See Härdle and Jerison(1991) for support of this claim.

3.3 Non-parametric Regression Analysis

As explained in section 2.1. Engel seems to be the first – certainly in the economic literature – who proposed a **non-parametric estimator** for the regression function of a general bivariate distribution. The statistical properties of Engel's estimator (or alternative ones such as regressogram- or kernel-estimator) were developed much later in the mathematical statistical literature, starting in the 1960's. Today this is a well-developed field; for standard references see Härdle(1990), Simonoff(1996) or Li & Racine (2007) . The first applications of these non-parametric methods to Family Budgets and Engel's Law appeared in the economic literature in the 80's and early 90's of the last century: see K. Hildenbrand and W. Hildenbrand(1986), Härdle and Jerison(1988) and (1991), Bierens and Pott-Buter(1990), Lewbel(1991), Blundell et al(1993), Engel & Kneip (1996).

Non-parametric methods allow one to compute a uniform confidence region for the estimated regression. This region can be used to test hypotheses about the form of the underlying regression function. The large empirical literature for different populations varying in time and space supports well the hypothesis of a decreasing regression of food share on income or total expenditure if one neglects the near boundary region of the income distribution.

4. New empirical Support of Engel's Law: Four Measures of Stochastic Association

In this section, two large data sets on consumption expenditure are considered; one from UK: Family Expenditure Survey (FES) and one from India: Consumer Expenditure Survey of National Sample Survey Organization (NSSO), which not only vary in terms of place of origin and time, but also in terms of its size. To save space we represent here only the results for FES in 1994 and for NSSO in 2005. FES data consists of 6657 observations after omitting extreme observations and for this data set not only total expenditure, but also income data are reported. [4] The consumer expenditure data for the rural population in India consists of 63028 observations. For the Indian data only total expenditure is reported.

4.1: Kendall's τ

The sample estimates of Kendall's τ with confidence interval for these two data sets are reported in table 4.1 a. Similar values are abtained for other years. No structural assumption of the population distribution is required here except that both X *and* Y are assumed to be distributed continuously. Given a random sample of size n of observations (X_i, Y_i), we may estimate and test the population values of Kendall's τ by the corresponding sample statistics $\hat{\tau}$ using the relative frequencies for each pair of

[4] In the FES data set two definitions of income and total expenditure are used; one is including (before) and other is excluding (after) housing costs within which gross rent, water-electricity charges, council water charges, mortgage etc are included. We use both total expenditure and income after housing cost because of the compatibility with the Indian data, where total expenditure in computed without considering housing cost.

observations. The details of the calculation of sample statistics is given in Newson's (2002) which is based on Hoeffding's U-statistics (1948). The confidence limits are calculated by jackknifing the U-statistics (Arvesen (1969)).

Table 4.1a: Estimates of Kendall's τ

Measures	τ	95% confidence interval	
Estimates of Kendall's τ for food share and tex_ahc* (FES)	-0.38	-0.39	-0.37
Estimates of Kendall's τ for food share and inc_ahc** (FES)	-0.40	-0.41	-0.39
Estimates of Kendall's τ for food share and total expenditure (NSSO)	-0.63	-0.633	-0.626

Data source: Family Expenditure Survey (FES) 1994 & National Sample Survey data of India (NSSO) 61st round, 2005.

- *Food share is calculated by dividing expenditure on food only by total expenditure after subtracting housing cost(tex_ahc).
- ** Food share is calculated by dividing expenditure on food only by income after subtracting housing cost(inc_ahc).

The consideration of subpopulations has important implications on individual behaviour , which is described in section 2.2. A large number of subpopulations of FES and NSSO have been analysed. In all cases estimates of Kendall's τ and the entire confidence interval are negative. For a small selection of subpopulations, the relevant statistics are presented in the following table 4.1b.[5] The description of the subgroups is given in the following table.

Description of subgroups

Subgroups from FES	Groups
Group 1: 2 adults &Employed (864)	Group 1:Hindu, SC-ST, AL & 2 Adults (468) ^
Group 2: 2 adults &Unoccupied (983)	Group 2: Hindu, OBC & 2 Adults+1 Baby [0-5 years age] (565)
Group 3: 2 adults& Self-employed (190)*	Group 3: Hindu, Upper Caste & 2 Adults+1 baby+1 adult child [6-15 years age] (279)
	Group 4: Muslim & 2 Adults (416)

*Employment status describes the status of household head.
^ Castesd are indicated by: SC-ST: Scheduled caste and Scheduled Tribe, OBC: Other backward caste. : AL: agricultural labour : indicates household main occupation.

[5] The relevant statistics, presented here for UK data, are for food share out of total expenditure. Similar results are observed for food share and income after housing cost.

4.1b: Kendall's τ for subgroups from UK (FES) and Indian (NSSO) data

Subgroups from FES data	Kendall's τ	95% confidence interval
Group 1 (864)	-0.40	[-0.44 -0.36]
Group 2 (983)	-0.50	[-0.53 -0.46]
Group 3 (190)	-0.40	[-0.49 -0.31]
Subgroups from NSSO data	Kendall's τ	95% confidence interval
Group 1 (468)	-0.23	[-0.29 -0.17]
Group 2 (565)	-0.31	[-0.36 -0.25]
Group 3 (279)	-0.32	[-0.40 -0.25]
Group 4 (416)	-0.33	[-0.38 -0.27]

4.2 Quadrant dependence

Lehmann's (1966) quadrant dependence condition, as described in definition 2 in section 2.1, can be formulated in terms of stochastic dominance of cumulative distribution functions (CDF's); the marginal distribution function of food share $F_Y(y)$ and the conditional distribution function $F_Y(y \mid X \le x)$ for any income level x. Hence the null and alternative hypotheses are formulated as:

$$H_0 : F_Y(y \mid X \le x) \le F_Y(y) \text{ for all } y \text{ and } x$$
$$H_1 : F_Y(y \mid X \le x) > F_Y(y) \text{ for some } y.$$

In this paper we follow the test proposed by Barrett and Donald (2003). They assume continuity of the two CDF and allow for random samples of different sizes n and m, from two distributions $F_Y(y \mid X \le x)$ and $F_Y(y)$ respectively. The test statistic for testing the hypothesis is $\hat{S}_1 = \sqrt{\frac{n \times m}{n+m}} \sup_y (\hat{F}_Y(y \mid X \le x) - \hat{F}_Y(y))$ where $\hat{F}_y$ denotes the empirical distribution function. The P-values for stochastic dominance have closed-form distribution $\exp(-2(\hat{S}_1)^2)$.

For presentation purpose, we consider only three conditional CDF of food share given three particular values of total expenditure; first at a 25th quantile (x_1), second at median (x_2) and third at 75th quantile (x_3) of total expenditure. For each particular values of x, the test is proceeded in two-steps: first by examining if the conditional CDF of food share for a particular quartile of total expenditure level $F_Y(. \mid X \le x)$ stochastically

dominates the marginal CDF of food share for the whole sample; and then in the second step one tests if $F_Y(.)$ stochastically dominates $F_Y(.|X \le x)$. If we fail to reject the first step but can reject the second step, we conclude that $F_Y(y|X \le x)$ stochastically dominates $F_Y(y)$. If we reject or fail to reject both steps of the test, we conclude that there is no stochastic dominance relation between the two distribution functions.

The following tables 4.2a and 4.2c present the p-values for UK (FES) and Indian (NSSO) data for the whole data and for subgroups respectively.

Table 4.2a: Negative quadrant dependence (P-Values)

(FES)	$X \le 25th\ \ qtl$	$X \le 50th\ \ qtl$	$X \le 75th\ \ qtl$
$F_Y(.\|X \le x)$ SD $F_Y(.)$ *	0.99	0.99	0.99
$F_Y(.)$ SD $F_Y(.\|X \le x)$	0.00	0.00	0.00
$F_Y(.\|X \le x)$ SD $F_Y(.)$ **	1.00	1.00	0.99
$F_Y(.)$ SD $F_Y(.\|X \le x)$	0.00	0.00	0.00

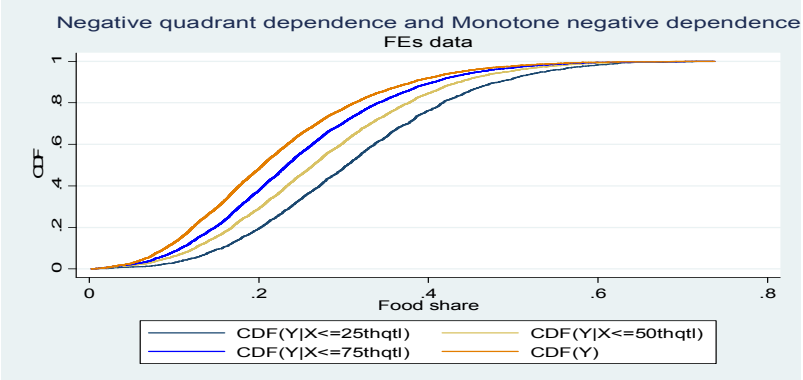
(NSSO)	$X \le 25th\ \ qtl$	$X \le 50th\ \ qtl$	$X \le 75th\ \ qtl$
$F_Y(.\|X \le x)$ SD $F_Y(.)$	0.99	1.00	1.00
$F_Y(.)$ SD $F_Y(.\|X \le x)$	0.00	0.00	0.00

- * & **: As described in the footnote of table 4.1a

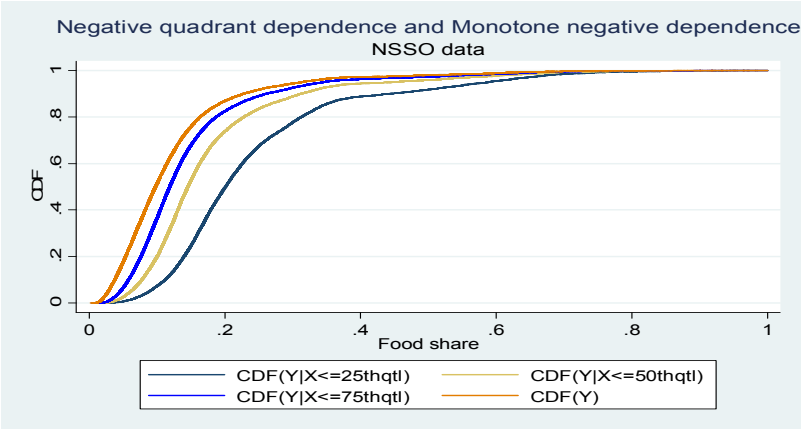
We also plot the CDFs for both FES and NSSO only for three different values of total expenditure, shown in figures 4.2a and 4.2b[6]. These plots suggest that even monotone negative dependence (definition 2a in section 2.2) prevails.

[6] Similar pattern in the plot is observed for FES data corresponding to income level, which are available from the authors on request.

4.2a: Negative quadrant dependence for FES whole sample



4.2b: Negative quadrant dependence for NSSO whole sample



Both results, the p-values and the graphs, indicate the fact that for UK as well as for Indian data, the probability of smaller food share is considerably less for a lower range of total expenditure (X), as compared to the whole range of total expenditure. We also test for *monotone negative quadrant dependence* (definition 2a in section 2.2) using the conditional CDF's for three different values of total expenditure, namely 25th quantile (CDF25th) , 50th (CDF50th) and 75th quantile (CDF75th). Therefore, using the stochastic dominance test, as described above, we test the null hypothesis $H_0 : F_Y(y \mid X \le x_1) \le F_Y(y \mid X \le x_2)$ *for all* y *and* $x_1 < x_2$. The p-values are described in table 4.2b for the whole sample. The diagrams and p-values for the whole sample support well the property of monotone quadrant dependence, which is stronger than Lehmann's negative quadrant dependence.

Table 4.2b: Monotone Negative Quadrant dependence (P- values) for the whole sample

FES	P-values
$F_Y(y \mid X \le x_1) \le F_Y(y \mid X \le x_2)$ *	1.00
$F_Y(y \mid X \le x_2) \le F_Y(y \mid X \le x_1)$ *	0.00
$F_Y(y \mid X \le x_2) \le F_Y(y \mid X \le x_3)$ *	0.99
$F_Y(y \mid X \le x_3) \le F_Y(y \mid X \le x_2)$ *	0.00
$F_Y(y \mid X \le x_1) \le F_Y(y \mid X \le x_2)$ **	0.99
$F_Y(y \mid X \le x_2) \le F_Y(y \mid X \le x_1)$ **	0.00
$F_Y(y \mid X \le x_2) \le F_Y(y \mid X \le x_3)$ **	0.99
$F_Y(y \mid X \le x_3) \le F_Y(y \mid X \le x_2)$ **	0.00
NSSO	
$F_Y(y \mid X \le x_1) \le F_Y(y \mid X \le x_2)$	0.99
$F_Y(y \mid X \le x_2) \le F_Y(y \mid X \le x_1)$	0.00
$F_Y(y \mid X \le x_2) \le F_Y(y \mid X \le x_3)$	1.00
$F_Y(y \mid X \le x_3) \le F_Y(y \mid X \le x_2)$	0.00

* & **: As described in the footnote of table 4.1a

Next, we consider the subgroups from each set of data. We only present the p-values for all subgroups, not the graphs in order to make the presentation less cumbersome[7].

4.2c Negative quadrant dependence for subgroups from UK (FES) and Indian (NSSO) data (P-values)

Subgroups from FES data	$X = 25th \quad qtl$	$X = 50th \quad qtl$	$X = 75th \quad qtl$
Group 1 (864)	$F_Y(. \mid X \le x)$ SD $F_Y(.)$ * =0.99	$F_Y(. \mid X \le x)$ SD $F_Y(.)$ * =0.99	$F_Y(. \mid X \le x)$ SD $F_Y(.)$ * =1.00
	$F_Y(.)$ SD $F_Y(. \mid X \le x)$ =0.00	$F_Y(.)$ SD $F_Y(. \mid X \le x)$ =0.00	$F_Y(.)$ SD $F_Y(. \mid X \le x)$ =0.00
Group 2 (983)	$F_Y(. \mid X \le x)$ SD $F_Y(.)$ * =0.99	$F_Y(. \mid X \le x)$ SD $F_Y(.)$ * =1.00	$F_Y(. \mid X \le x)$ SD $F_Y(.)$ * =1.00
	$F_Y(.)$ SD $F_Y(. \mid X \le x)$ =0.00	$F_Y(.)$ SD $F_Y(. \mid X \le x)$ =0.00	$F_Y(.)$ SD $F_Y(. \mid X \le x)$ =0.00
Group 3 (190)	$F_Y(. \mid X \le x)$ SD $F_Y(.)$ * =0.99	$F_Y(. \mid X \le x)$ SD $F_Y(.)$ * =1.00	$F_Y(. \mid X \le x)$ SD $F_Y(.)$ * =1.00
	$F_Y(.)$ SD $F_Y(. \mid X \le x)$ =0.00	$F_Y(.)$ SD $F_Y(. \mid X \le x)$ =0.002	$F_Y(.)$ SD $F_Y(. \mid X \le x)$ =0.08
Subgroups from	$X = 25th \quad qtl$	$X = 50th \quad qtl$	$X = 75th \quad qtl$

[7] The corresponding figures for CDFs are available from the authors on request. Also the p-values for FES subgroups are reported for a given value of total expenditure only, not for income.

NSSO data			
Group 1 (468)	$F_Y(.\mid X \le x)$ SD $F_Y(.)^*$ =0.93	$F_Y(.\mid X \le x)$ SD $F_Y(.)^*$ =0.99	$F_Y(.\mid X \le x)$ SD $F_Y(.)^*$ =1.00
	$F_Y(.)$ SD $F_Y(.\mid X \le x)$ =0.00	$F_Y(.)$ SD $F_Y(.\mid X \le x)$ =0.01	$F_Y(.)$ SD $F_Y(.\mid X \le x)$ =0.13
Group 2 (565)	$F_Y(.\mid X \le x)$ SD $F_Y(.)^*$=0.99	$F_Y(.\mid X \le x)$ SD $F_Y(.)^*$ =0.99	$F_Y(.\mid X \le x)$ SD $F_Y(.)^*$ =1.00
	$F_Y(.)$ SD $F_Y(.\mid X \le x)$ =0.00	$F_Y(.)$ SD $F_Y(.\mid X \le x)$ =0.00	$F_Y(.)$ SD $F_Y(.\mid X \le x)$ =0.00
Group 3 (279)	$F_Y(.\mid X \le x)$ SD $F_Y(.)^*$ =1.00	$F_Y(.\mid X \le x)$ SD $F_Y(.)^*$ =0.99	$F_Y(.\mid X \le x)$ SD $F_Y(.)^*$=0.99
	$F_Y(.)$ SD $F_Y(.\mid X \le x)$ =0.00	$F_Y(.)$ SD $F_Y(.\mid X \le x)$ =0.01	$F_Y(.)$ SD $F_Y(.\mid X \le x)$ =0.05
Group 4 (416)	$F_Y(.\mid X \le x)$ SD $F_Y(.)^*$ =0.99	$F_Y(.\mid X \le x)$ SD $F_Y(.)^*$ =0.99	$F_Y(.\mid X \le x)$ SD $F_Y(.)^*$ =0.99
	$F_Y(.)$ SD $F_Y(.\mid X \le x)$ =0.00	$F_Y(.)$ SD $F_Y(.\mid X \le x)$ =0.00	$F_Y(.)$ SD $F_Y(.\mid X \le x)$ =0.04

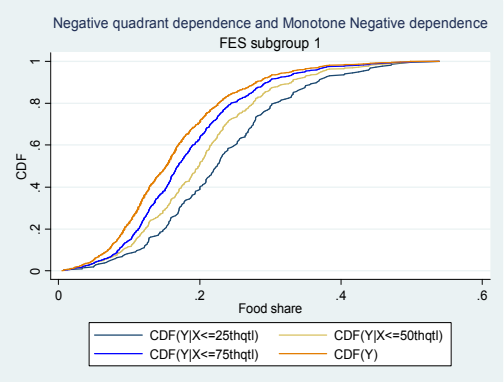
Table 4.2d: Monotone Negative Quadrant dependence (P- values) for the subgroups

Subgroups from FES data	$F_Y(y \mid X \le x_1) \le F_Y(y \mid X \le x_2)$	$F_Y(y \mid X \le x_2) \le F_Y(y \mid X \le x_1)$	$F_Y(y \mid X \le x_2) \le F_Y(y \mid X \le x_3)$	$F_Y(y \mid X \le x_3) \le F_Y(y \mid X \le x_2)$
Group 1 (864)	p value=0.99	p value=0.002	p value=0.99	p value=0.00
Group 2 (983)	p value=0.99	p value=0.00	p value=0.99	p value=0.00
Group 3 (190)	p value=0.99	p value=0.01	p value=0.99	p value=0.02
Subgroups from NSSO data	$F_Y(y \mid X \le x_1) \le F_Y(y \mid X \le x_2)$	$F_Y(y \mid X \le x_2) \le F_Y(y \mid X \le x_1)$	$F_Y(y \mid X \le x_2) \le F_Y(y \mid X \le x_3)$	$F_Y(y \mid X \le x_3) \le F_Y(y \mid X \le x_2)$
Group 1 (468)#	p value=0.93	p value=0.17	p value=0.99	p value=0.27
Group 2 (565)#	p value=1.00	p value=0.11	p value=0.99	p value=0.29
Group 3 (279)#	p value=0.99	p value=0.15	p value=0.99	p value=0.47
Group 4 (416)#	p value=0.99	p value=0.12	p value=0.99	p value=0.14

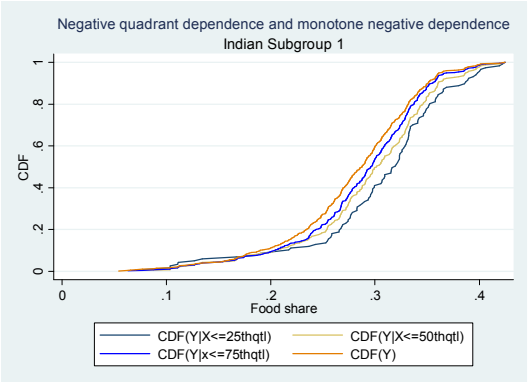
#: Monotone negative quadrant dependence is not satisfied.

We also present plots only for 1st subgroup from FES and 1st subgroup from NSSO to illustrate our results described in tables 4.2c and 4.2d.

4.2e: Subgroup 1 from FES data



4.2f Subgroup1 from NSSO data



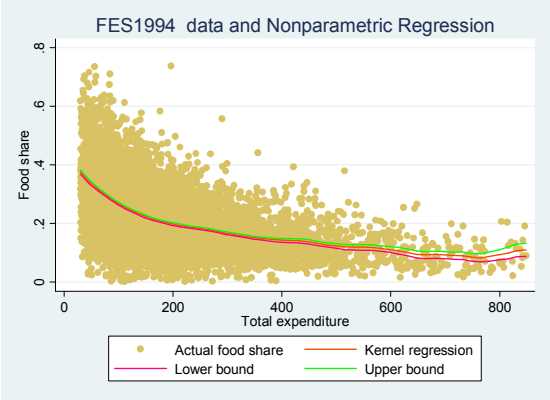
4.3 Decreasing regression:

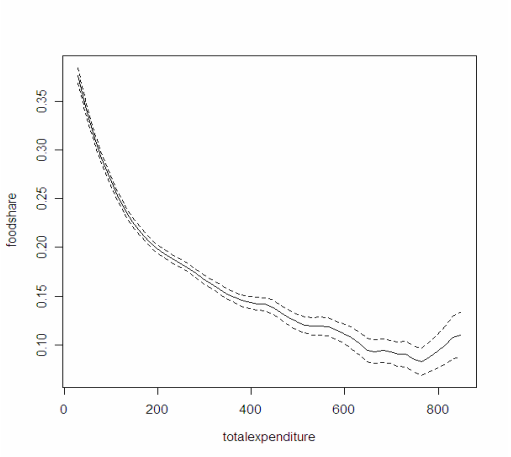
Estimation of regression function using nonparametric Kernel smoothing technique is standard today (see section 3.3 for references). Only for completeness we give the estimates with confidence region for FES and NSSO data sets. In the nonparametric regression model $Y_i = m(X_i) + \varepsilon_i$, where sample observations (X_i, Y_i) are i.i.d and $m(.)$ is a smooth function, one can estimate $m(.)$ nonparametrically using Kernel method. We use the local linear estimator of $m(.)$ (Stone (1977)) which is obtained by minimizing $\min_{\{a,b\}} \sum_{i=1}^{n} (Y_i - a - (X_i - x)'b)^2 K(\frac{X_i - x}{h})$. The smoothing parameter h is called the bandwidth parameter, and K is the Kernel function. We choose the optimum bandwidth using least-square cross-validation technique and the second order the Epanechnikov Kernel is used for estimation.

The nonparametrically estimated regression curve is shown in the following diagrams with the confidence bands constructed with asymptotic normality, first for the whole sample of FES considering both income and total expenditure as X variable, and for NSS data and for few subgroups from these data sets[8].

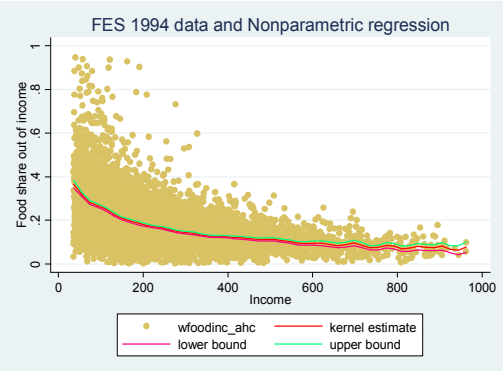
[8] The nonparametric regression curves for all the subgroups are available from the authors on request.

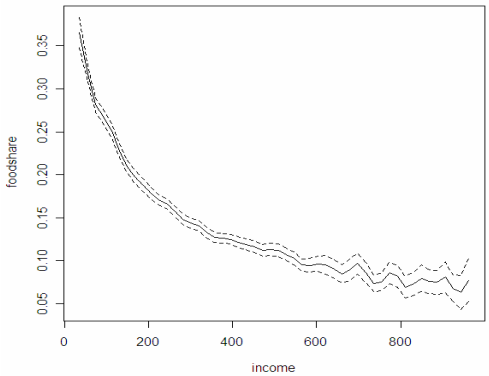
Figure 4.3a: Nonparametric Regression with confidence band from FES data





Food share is calculated by dividing expenditure on food only by total expenditure after housing cost.



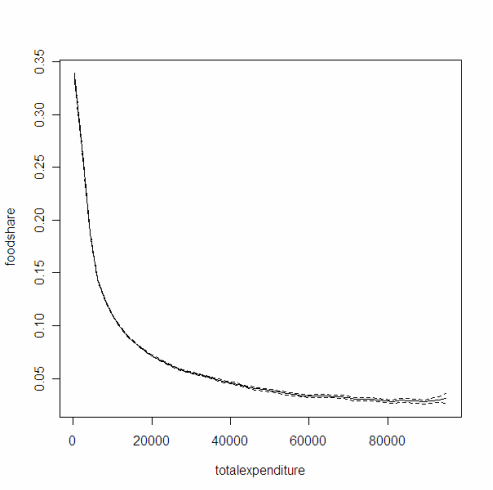


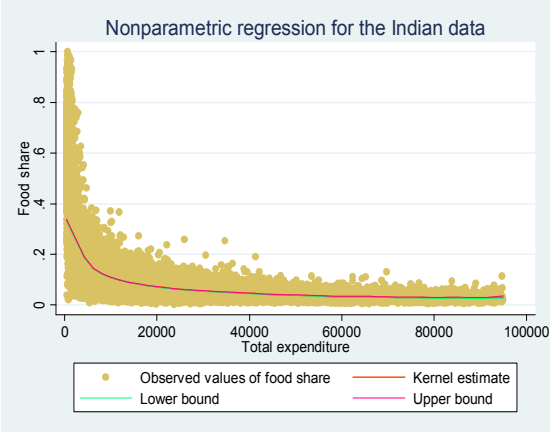
- Food share is calculated by dividing expenditure on food only by income after housing cost.

The property of decreasing regression is well supported if restricted on the main domain of income distribution. The nonparametrically estimated regression curves are also shown for few subgroups from FES and NSS in the following figures 4.3c and 4.3d respectively[9].

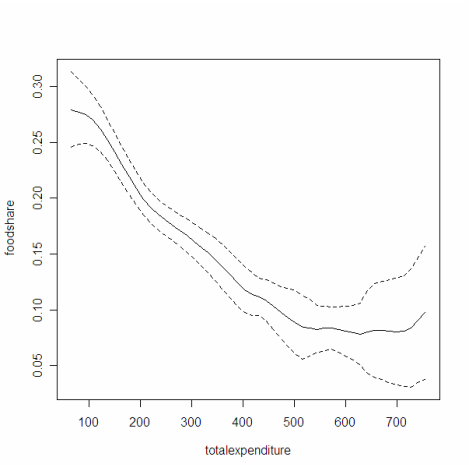
[9] For subgroups we have used adaptive nearest neighbour bandwidth for cross-validation purpose and use 499 bootstrap to compute the bandwidths.

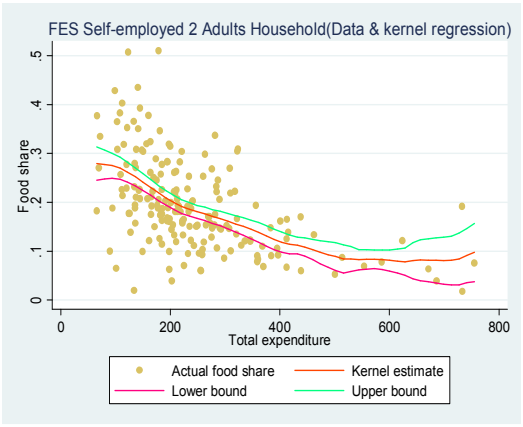
4.3b: Nonparametric regression for the Indian whole sample

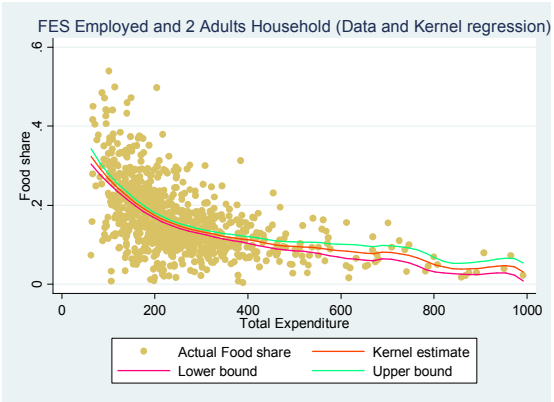


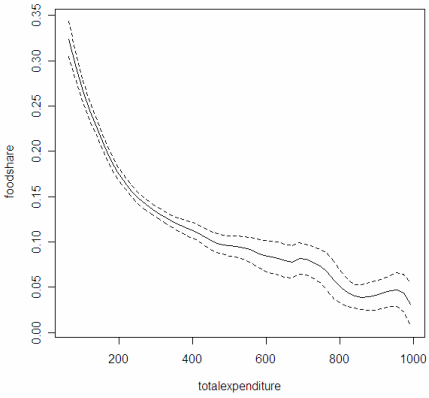


4.3c: Nonparametric regression for the subgroups of FES data

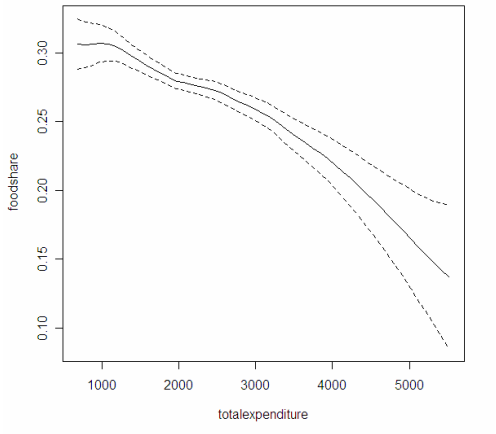


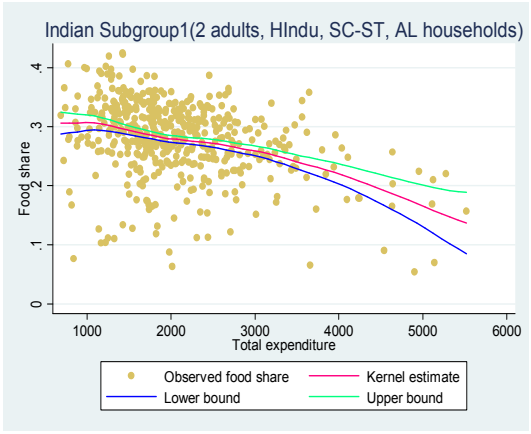


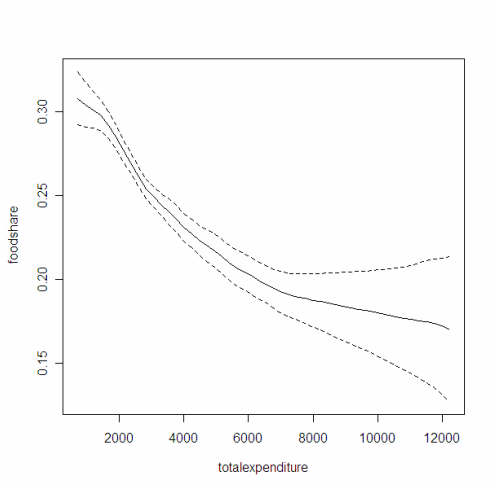


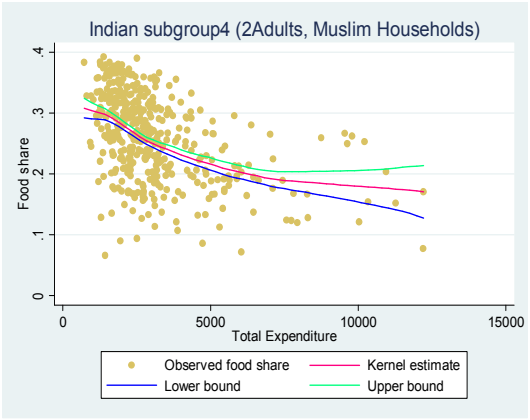


4.3d: Nonparametric regression from the subgroups of NSS data









4.4 Stochastically decreasing conditional distribution functions

Finally we consider stochastically decreasing conditional distribution functions (Tukey (1958)) as described in definition 4 of section 2.1. i.e., stochastic dominance of the conditional distribution function, conditioned on $X = x$, denoted by $F_Y(.\,|\,x)$.

Therefore, the hypothesis for Tukey's condition can be formulated as follows:
$H_0 : F_Y(y \,|\, x_1) \le F_Y(y \,|\, x_2)$ *for all y and for any* $x_1 < x_2$
$H_1 : F_Y(y \,|\, x_1) > F_Y(y \,|\, x_2)$ *for some y*

We consider two conditional distribution functions of food share given two small consecutive intervals of total expenditure/ Income level; one at $5^{th}\, qtl < x \le 10^{th}\, qtl$ (x_1) and other at $10^{th}\, qtl < x \le 15^{th}\, qtl$ (x_2) of total expenditure / income. The test is proceeded in two-steps as before following Barrett and Donald (2003): first by examining if

$F_Y(.|X=x_1)$ stochastically dominates $F_Y(.|X=x_2)$; and then in the second step tests if $F_Y(.|X=x_2)$ stochastically dominates $F_Y(.|X=x_1)$. If we fail to reject the first step but can reject the second step, we conclude that the $F_Y(y|x_1)$ stochastically dominates the $F_Y(y|x_2)$, thus satisfies condition 4 of negative stochastic association. If we reject or fail to reject both steps of the test, we conclude that there is no stochastic dominance relation.

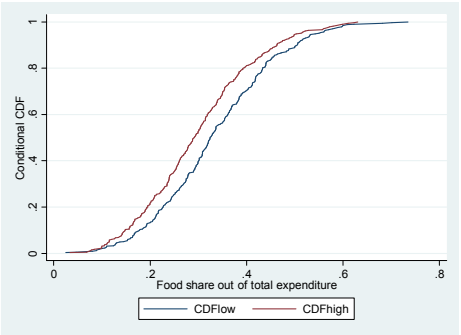
The test results for stochastically decreasing CDF are described in table 4.4a for the whole sample of FES and NSSO data and are illustrated in figures 4.4a and 4.4b. The strongest condition of negative stochastic association is well supported in the FES as well as in the NSSO data.[10] This is not so evident for consecutive intervals around income considered (see the p-values in table 4.4a and the right-most plot in figure 4.4a).

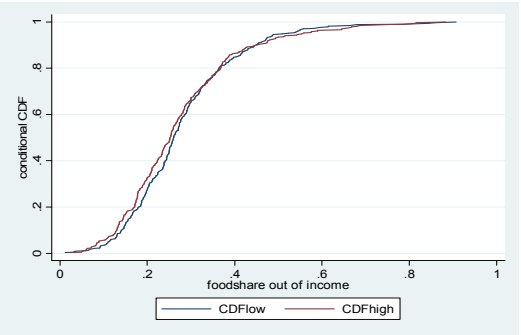
Table 4.4a: Test of stochastic dominance given total expenditure(tex_ahc)/income (inc_ahc) for quantiles (5-10th) and (10-15th).

From the FES data	P values*	P values **
$F_Y(y\|X=x_1)SDF_Y(y\|X=x_2)$	0.99	0.83 #
$F_Y(y\|X=x_2)SDF_Y(y\|X=x_1)$	0.00	0.12#
From the NSS data		
$F_Y(y\|X=x_1)SDF_Y(y\|X=x_2)$	1.00	
$F_Y(y\|X=x_2)SDF_Y(y\|X=x_1)$	0.00	

* & ** are described before in the footnote of table 4.1a. #: Weakly satisfied. In case of Indian data only total expenditure is reported, as stated earlier.

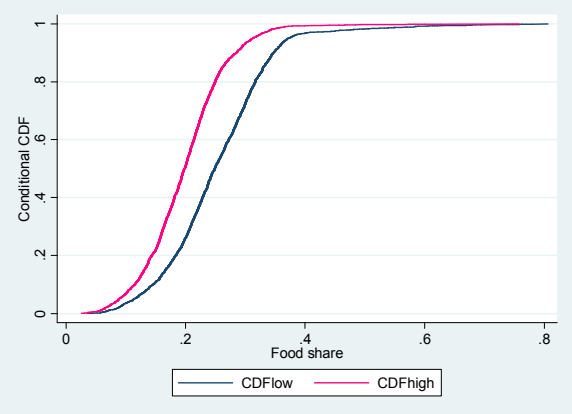
4.4a: Plot of Empirical CDFs of food share given two total expenditure/income levels of FES data





[10] The Tukey's condition for income levels are satisfied only weakly. Due to shortage of space the figures and tests of stochastic dominance for other intervals of expenditure and income levels are not reported here. Those are available from authors on request. Yet, we would like to mention here that similar feature is observed for whole range of income distributions.

Fig 4.4b: Plot of Empirical CDFs given two total expenditure levels of Indian data



The p-values for stochastic dominance test for the subgroups are reported in table 4.4b. The values indicate absence of clear stochastic dominance in several subgroups.

4.4b: Test of stochastic dominance given total expenditure for subgroups from FES and NSS data*

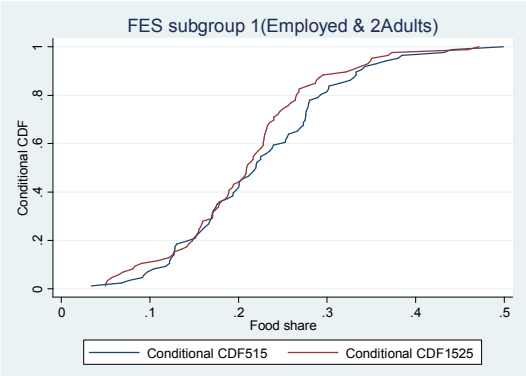
Groups from the FES data**	**Description**	**P values**
Group 1 (864)$	$F_Y(y \mid X = x_1) SDF_Y(y \mid X = x_2)$ $where..5thqtl \le x_1 < 15thqtl, 15thqtl \le x_2 < 25thqtl$	0.90
	$F_Y(y \mid X = x_2) SDF_Y(y \mid X = x_1)$ $where..5thqtl \le x_1 < 15thqtl, 15thqtl \le x_2 < 25thqtl$	0.10
Group 2 (983)	$F_Y(y \mid X = x_1) SDF_Y(y \mid X = x_2)$ $where..5thqtl \le x_1 < 15thqtl, 15thqtl \le x_2 < 25thqtl$	0.96
	$F_Y(y \mid X = x_2) SDF_Y(y \mid X = x_1)$ $where..5thqtl \le x_1 < 15thqtl, 15thqtl \le x_2 < 25thqtl$	0.04
Group 3 (190)@	$F_Y(y \mid X = x_1) SDF_Y(y \mid X = x_2)$ $where..5thqtl \le x_1 < 25thqtl, 25thqtl \le x_2 < 50thqtl$	0.97
	$F_Y(y \mid X = x_2) SDF_Y(y \mid X = x_1)$ $where..5thqtl \le x_1 < 25thqtl, 25thqtl \le x_2 < 50thqtl$	0.01
Subgroups from NSSO data	**Description**	**P values**
Group 1 (468)#	$F_Y(y \mid X = x_1) SDF_Y(y \mid X = x_2)$ $where..5thqtl \le x_1 < 15thqtl, 15thqtl \le x_2 < 25thqtl$	0.59
	$F_Y(y \mid X = x_2) SDF_Y(y \mid X = x_1)$ $where..5thqtl \le x_1 < 15thqtl, 15thqtl \le x_2 < 25thqtl$	0.48

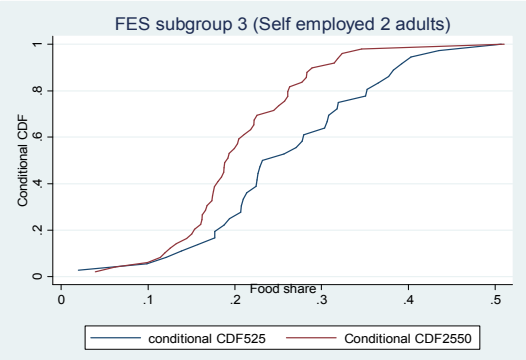
Group 2 (565) #	$F_Y(y \mid X = x_1) SDF_Y(y \mid X = x_2)$ $where..5thqtl \le x_1 < 15thqtl, 15thqtl \le x_2 < 25thqtl$	0.74
	$F_Y(y \mid X = x_2) SDF_Y(y \mid X = x_1)$ $where..5thqtl \le x_1 < 15thqtl, 15thqtl \le x_2 < 25thqtl$	0.48
Group 3# (279)@	$F_Y(y \mid X = x_1) SDF_Y(y \mid X = x_2)$ $where..5thqtl \le x_1 < 25thqtl, 25thqtl \le x_2 < 50thqtl$	0.15
	$F_Y(y \mid X = x_2) SDF_Y(y \mid X = x_1)$ $where..5thqtl \le x_1 < 25thqtl, 25thqtl \le x_2 < 50thqtl$	0.94
Group 4# (416)	$F_Y(y \mid X = x_1) SDF_Y(y \mid X = x_2)$ $where..5thqtl \le x_1 < 15thqtl, 15thqtl \le x_2 < 25thqtl$	0.91
	$F_Y(y \mid X = x_2) SDF_Y(y \mid X = x_1)$ $where..5thqtl \le x_1 < 15thqtl, 15thqtl \le x_2 < 25thqtl$	0.45

- *Due to few observations within each subgroup we have considered here bigger intervals as compared to the whole sample to have sufficient observations.
- @ For the 3rd subgroups from both the FES data & the NSSO data, we have considered much bigger interval due to very small observations in these subgroups.
- ** For the FES data the tests are reported only for given values of total expenditure, not for income.
- # indicates non-satisfaction of stochastic dominance.
- $ Indicates weak satisfaction of stochastic dominance

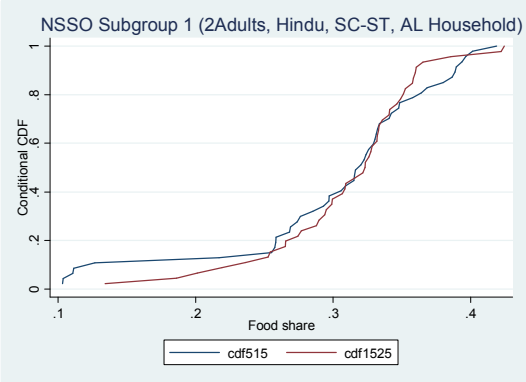
The following diagrams present CDFs for few sub-groups to support the test results described in table 4.4b.

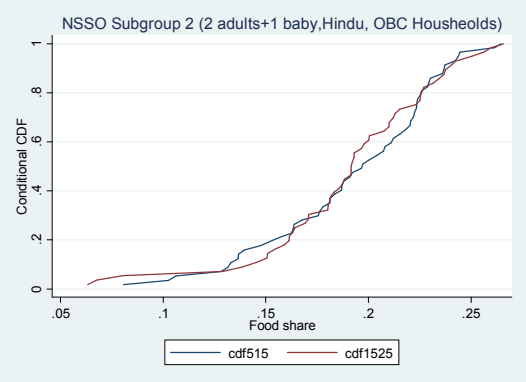
4.4c: Subgroups from FES data





4.4d: Subgroups from NSSO data





5: Conclusion

Engel's verbal formulation (see statement (i) and (ii) in the introduction) of his law expresses a 'negative stochastic association' of the bivariate distribution of income (total expenditure) and food share.

Among the many different definitions of 'negative stochastic association' which can be found in the statistical literature, we have chosen four: negative Kendall's τ, negative quadrant dependence, stochastically decreasing conditional food share distribution functions and decreasing regressions (see section 2.1). Only the last property is used in the economic literature in order to define Engel's law. Yet a decreasing regression does not imply useful information of its underlying bivariate distribution, in particular, it does not imply a negative Kendall's τ nor negative quadrant dependence. However, one expects these properties to be satisfied if one reads the two articles by Engel (1857) and (1895).

Further more if one wants to link Engel's law with individual behaviour, then as we have shown in section 2.2, the property of a decreasing regression function is not sufficient, stronger properties are needed. This motivates the empirical study of section 4. We have shown that a negative Kendall's τ and negative quadrant dependence has good empirical support for the whole as well as for subpopulations of FES and NSSO. We have also shown that monotone negative quadrant dependence and even stochastically decreasing conditional food share functions have satisfactory empirical support in the case of total expenditure for the whole population of FES and NSSO. We expect that these empirical findings for the whole (unstratified) population will also hold for other data sets. This, of course, has to be shown. If the answer is positive, then the property of 'stochastically decreasing conditional food share distribution functions' is the proper definition of Engel's law.

For subpopulations, obtained by stratification, the situation is less clear. Naturally, income is not the only explanatory variable for food share. If one stratifies the population

with respect to a certain observable explanatory variable, for example, family size, then one eliminates the influence of this variable on the stochastic association between food share and income. This might increase or decrease the 'degree' of stochastic association. For example, in the case where for given income, food share and family size is positively associated and income and family size is negatively associated then one might expect (one can easily give examples) that controlling for family size decreases the 'degree' of negative association of income and food share, e.g., one still obtains a negative Kendall's τ or negative quadrant dependence, yet not stochastic decreasing conditional distribution functions. Engel's Law for subpopulations requires further empirical research analysis.

References

Allen R.G.D. and A.L. Bowley (1935), " Family Expenditure", London, 1935.

Arvesen JN (1969). "Jackknifing U-statistics." Annals of Mathematical Statistics, 40, 2076–2100.

Barrett, G. and S. Donald (2003), "Consistent Tests for Stochastic Dominance," Econometrica, Vol 71, No 1, pp 71-104.

Bierens, H. and H. A. Pott-Buter (1990), "Specification of Household Expenditure Functions and Equivalence Scales by Nonparametric Regression," Econometric Reviews, 9, 123-210.

Browning. M (2008), "Engel's Law"; The New Palgrave Dictionary of Economics, Second Edition, Edited by Steven N. Durlauf and Lawrence E. Blume

Blundell, R., M. Browning and I. Crawford (2003), Nonparametric Engel curves and revealed preference," Econometrica, 71, 205-240.

Collomb, G. (1977)., "Quelques propriétés de la méthode du noyau pour l'estimation non-paramétrique de la regression en un point fixé.", *C. R. Acad. Sc. Paris* **285**: 289-92.

Ducpetiaux (1855) , "Budgets Economiques des Classes Ouvriers en Belgique", Brussels.

Davies, D (1795), "The Case of Labourers in Husbandry",London.

Engel, E (1857), *Die Productions- und Consumtionsverhaltnisse des Kdnig- reichs Sachsen" in Zeitschrift des Statistischen Buireaus des K. Sach- sischen, Ministerium des Innern,* No. 8 u. 9, pp1-54. It was reprinted as an appendix to "*Die Lebenskosten Belgischer Arbeiter Familien frfther und jetzt,"* Bulletin de l'institut international de statistique, tome IX, premiere livraison, Rome 1895.

Engel, E. (1895), "Die Lebenskosten Belgischer Arbeiter-Familien Fruher and jetzt," International Statistical Institute Bulletin, vol. 9, pp. 1-74.

Engel, J and A. Kneip (1996),“Recent Approaches to Estimating Engel Curves“. Journal of Economics, Vol-62, No-2, 187-212.

Fechner, Gustav, T (1897), *Kollektivemasslehre,* Wilhelm Engelmann, Leipzig.

Gauss (1809),” Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium Frid”;. Perthes et I H Besser, Hamburg, *Theory of the Motion of the Heavenly Bodies* ... (Trans. from original edition published in Latin in Hamburg, 1809). Reprinted, New York, 1963.

Hildenbrand, K and W. Hildenbrand (1986), “On the Mean Income Effect: A Data Analysis of the U.K. Family Expenditure Survey", in Hildenbrand and Mas-Colell, edited *Contributions to Mathematical Economics* in Honor of Gerard Debreu. North-Holland, Amsterdam.

Härdle, W (1990), “ Applied Nonparametric Regression”, CUP, Cambridge.

Härdle, W. and M. Jerison (1988), "Cross Section Engel Curves Over Time," discussion paper A-178, SFB 303, University of Bonn (**1988**).

Härdle, W. and M. Jerison (1991), "Cross Section Engel Curves Over Time," Recherches Economiques de Louvain, 57, 391-431.

Hoeffding, W (1948), “A non-parametric test of independence,” Annals of Mathematical Statistics, Vol 19, pp 546-557.

Houthakker, H. S (1957), “An International Comparison of Household Expenditure Patterns, Commemorating the Centenary of Engel's Law,”

Kendall, M, G (1938), “ A New Measure of Rank Correlation,” Biometrika, vol 30, pp 81-93.

Lehmann, E. L. (1966), “Some Concepts of dependence,” Annals of Mathematical Statistics, Vol. 37, No. 5 , pp. 1137-1153.

Legendre, A.M (1805) *Nouvelles methodes pour la determination des orbites des comktes.*Paris. An abridged translation into English of the Appendix on least squares was published in D. E. SMITH, *A Source Book of Mathematics,* repr., in 2vols., New York, NY, 1959.

Le Play(1855), "Les ouvriers européens," Paris.

Leser, C. E. V. (1963), "Forms of Engel Functions," Econometrica, 31, 694-703.

Lewbel, A. (1991), "The Rank of Demand Systems: Theory and Nonparametric Estimation," Econometrica, 59, 711-730.

Lewbel, A (2008), "Engel Curves', Entry for The new Palgrave Dictionary of Economics, 2nd edition, Edited by Steven N. Durlauf and Lawrence E. Blume

Li, Q. and Racine, J. S (2007), "Nonparametric Econometrics: Theory and Practice, Princeton University Press.

Lipps, G. F. (1906), Die Psychischen Massmethoden, Braunschweig.

Newson, R (2002), "Parameters Behind Nonparametric Statistics: kendall's tau, Somer's D and Median Differences," The Stata Journal, Vol 2, No 1, pp 45-64.

Perthel, D (1975), "Engel's Law Revisited"; International Statistical Review , Vol. 43, No. 2, pp. 211-218.

Prais, S.J and H. S Houthakker (1955), " The analysis of Family Budgets"; Cambridge.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, New York: Springer.

Stigler, G, J (1954), " The Early History of Empirical Studies of Consumer Behaviour", Journal of Political Economy, Vol-62, No-2, 95-113.

Stone, C (1977)," Consistent nonparametric regression", *Ann. Statist.* **5** (1977), pp. 595–645.

Tukey, J. W. (1958). A problem of Berkson, and minimum variance orderly estimators. Ann. Math. Statist. 29 588-592.

Working, H. (1943), "Statistical Laws of Family Expenditures," Journal of the American Statistical Association, 38, 43-56.

Young, A (1771), "Farmer's Letters**",** London, especially Letter V.

Sir Frederick Eden(1797)**,** "The State of the Poor", London.

Appendix:

Description of Belgian data, used by Ernst Engel which was collected by the Provincial Statistical Commission and were processed under the direction of Edouard Ducpetiaux, Commission Centrale de Statistique in 1855.

This survey data includes information of 199 families across nine Belgian provinces. To compare relative standard of living among these areas the commission chose three categories of families in each location which are as follows: (i) workers sustained by public assistance, (ii) poor workers just able to live without such assistance and (iii) well-to-do workers living in comfortable circumstances. Also commission considered only families of a single type consisting of a father, mother and four children whose ages were sixteen, twelve, six and two [**Ducpetiaux words]**. Among these 199 families, Engel considered in his original study of 1857 only 153 families due to nonavailability of information on categories of 46 families. But in his 1895 paper where he considered explicitly the income classes he considered 199 families altogether. Although the final report in 1855 published annual budgets, information was actually collected on a weekly basis.

Table A I: Descriptive statistics: mean values

Expenditure/Income	153 Families			199 families
	Category i (48)	Category ii (51)	Category iii (54)*	
Food expenditure	459.85	569.55	757.98	613.53
Total expenditure	648.68	845.45	1214.44	933.24
(Min)	(377.06)	(387.32)	(411.00)	
(Max)	(1256.32)	(1768.82)	(2822.54)	
Income	564.97	796.54	1198.33	879.92
(Min)	(175.00)	(275.00)	(411.00)	
(Max)	(1298.00)	(1790.00)	(2830.00)	
Food share	0.71	0.68	0.63	0.67

Source: Edouard Ducpetiaux,:*Budgets Economiques des Classes Ouvriers en Belgique* (Brussels, 1855). All data are in Belgium Fr.

The terms in parentheses indicate minimum and maximum values of total expenditure and income respectively for each category. These values clearly reflect the overlapping of values of income across these categories. * Although this category reflects families in comfortable position, only 33 families out of 54 have income higher than total expenditure.