Modelling Income Distribution in Spain: A Robust Parametric Approach

Mercedes Prieto Alaiz and **Maria-Pia Victoria-Feser**[1]

April 1996

Published as a DARP Discussion Paper no 20, STICERD,
London School of Economics

[1]The second author is partially supported the ESRC, grant # R000 23 5725.

Abstract

This paper presents a robust estimation of two income distribution models using Spanish data for the period 1990-91 under three different concepts of income. The effect on the estimates of the Theil index due to the choice of the definition of income and of the estimation method is also analysed.
Key words: Income distribution, inequality, M-estimators, robust statistics, parametric estimation.
JEL classification: C13, D31.

1 Introduction

Fitting a parametric model to income data can be a valuable and informative tool of distributional analysis. Not only can one summarize the information contained in thousands of observations, but also useful information can be drawn directly from the estimated parameters. For example one could be interested in measuring income inequality, comparing different distributions or elaborating income redistribution policy: these concepts may sometimes be directly derived from parameters of a fitted distribution.

In view of this, it is important that model estimation is carried out appropriately. This means that one has to take into account the nature of the data in selecting an estimation procedure. In particular, the way 'income' is defined and also the quality of the data at hand play central roles.

This paper firstly discusses different definitions of income. In Section 2 the discussion on the choice of the income distribution is illustrated by an application to Spanish data. We use the *Encuesta de Presupuestos Familiares(EPF)* for the period 1990-1991. The EPF is a continuous survey whose effective sample consists of 21,155 households and primarily aimed at the elaboration of the weights for the Retail Index Price (Instituto Nacional de Estadistica 1992, p.7).

In distributional analysis a important controversy can arise when one arrives at the choice of the definition of income. This is of course and important problem, but it often undermines another as important problem which is the presence of gross errors or outliers in the data. Common sources of errors include recording errors or comma error (the income is inadvertently recorded 10 times its true value) and definition errors (the weekly income can be confused with the monthly income). Worse are the so-called legitimate extreme incomes because they lie far from the bulk of data and usually income models do not capture them satisfactorily. For examples see Cowell and Victoria-Feser (1996). Therefore in this paper we propose the use of robust methods to fit a model to the data.

In Section 3 we introduce key concepts of robust statistics, in particular the so called *influence function*. We present a robust estimator and give an algorithm to compute it. This estimator is defined generally for any parametric model, but in this paper we apply it to some income distribution models.

In Section 4, different types of functional form are proposed to fit Spanish

income data. The exercise of Salem and Mount (1974) for comparing the Lognormal and the Gamma parametrization is performed for three different kinds of income definition assigned to the Spanish households for the period 1990-1991. In particular, robust fits are compared for the two models.

2 Characterization of income distribution in Spain

Before describing the income distribution in Spain, we need to define what is understood by "income".

2.1 Income or expenditure

In this subsection, we discuss the different possible definitions of income in the particular case of Spanish data. We follow here the discussion in Atkinson (1983) (chapter 3) and the recommendations of the EPF for the period 1990-1991 (Instituto Nacional de Estadistica 1993). The economic unit and the equivalence scale need also to be defined.

Income is actually used to describe the economic status of economic units. However other variables such as expenditure and wealth are rival candidates. Wealth is usually not used because of the difficulty of wealth appraisal. The controversy lies between income and expenditure. This is particularly the case in Spain where the discrepancy between income and expenditure data is usually large. In fact, 56% of the interviewed households report higher expenditure than their corresponding income in the EPF 1990-91. Traditionally, this discrepancy has been attributed to the lack of reliability of income data. According to Alcaide and Alcaide (1983), there has been a systematic underestimation of income. These authors estimate a negative saving for the total Spanish households that represents 14.4% of the total disposable income according to the 80-81 EPF. In the same vein, Ruiz-Castillo (1987) maintains that more than 60% of the households sample report a higher expenditure than their income in the 80-81 survey. However, Sanz (1992) raises doubts about the reliability of expenditure data. Pazos (1994) points out that the difference between the sample period and the reference periods

of the expenditure[1] could influence the expenditure data. To see the latter point, Carrascal (1995) (p.140) gives a practical example: if a household reports the purchase of a coat, the cost of the coat will appear twelve times in its total expenditure. That means that the expenditure will be overestimated unless this household really buys a coat every month.

Although the choice between income or expenditure is not clear, we prefer to use income data. As Atkinson (1983) argues, we prefer to consider the economic position (income) rather than the level of welfare (expenditure) of the economic unit. Moreover, although the reliability of income data can be questioned, we have serious doubts about the reliability of expenditure data.

2.2 What income?

A difficulty with income is to choose the items that define it. To do that, one has to bear in mind that this variable has to reflect the economic position of the income unit. From the EPF (see Instituto Nacional de Estadistica 1992, p. 24-27 and Instituto Nacional de Estadistica 1993, p. 33), one can extract three different definitions:

1. Zubiri (1985) defines the income of an income unit in a given period as the increment in the potential spending power of the income unit in this period. He labels this definition as monetary income. The items included are: earning, self employment income, investment income and social security benefits net of income taxes and social contributions.

2. Atkinson (1983) points out that 'the income in a given period is the amount a person could have spent while maintaining the value of his wealth intact'. From this definition, the components that define the income are not only those of the monetary income but also, as Atkinson emphasizes, those items related to nonmonetary income: rents in kind (fringe benefits), production for home consumption and imputed rent.

3. Finally, another definition extracted from the EPF is that including monetary and non-monetary income plus the components of the extra-

[1]While the sample period is the week, the reference periods of the expenditure data vary according to the products. For example the reference period of the food is the week, of the clothes is the month, etc

ordinary income, namely lottery and gambling. This definition can be considered as to total income.

2.3 The income unit

There are three possible candidates for the income unit: the person, the nuclear family and the household. The income unit of the EPF is the household, that is the group of people living together at the same address with common housekeeping (see Instituto Nacional de Estadistica 1992, p.15). If one is interested in comparing the income of different households one should analyse income distribution of the household bearing in mind their size. A very simple way is to obtain the income per capita. However, as Coulter, Cowell, and Jenkins (1992) points out, the disadvantage of the income per capita is that it does not take into account that the marginal cost of an extra person may change as household size changes. Equivalence scale rates attempt to achieve both effect: the household size and the economies of scale generated by the size. By weighting the household income by a scale rate, an equivalised income is obtained. Atkinson (1983) distinguishes two main approaches in constructing equivalence scales. The first one is based on a priori consideration of the needs of children and adults. The second approach is based on the observed differences in consumption patterns for households of different size. The scales derived from both methods are subject to reservations and there is no general agreement about which equivalence to use. In fact, Coulter, Cowell, and Jenkins (1992) stress that 'there is no single *correct* equivalence scale for adjusting incomes'; for this reason 'a range of scale relativities is both justifiable and inevitable'.

The EPF introduces the OECD scale, called Oxford scale in the EPF (see Instituto Nacional de Estadistica 1992, p.34). The weights given to the members of the household are:

- 1.0 for the first adult in the household,
- 0.7 for the other adults and
- 0.5 for each child

2.4 The data

In this section we consider only the total income which can be rescaled as (Atkinson 1983):

1. *Total Income* which is actually the household total income,
2. *Equivalised Income* which is the household total income divided by the Oxford equivalence scale,
3. *Per Capita Income* which is actually the household total income divided by the number of members of the household.

We consider here the distribution by income units, i.e. without weighting the income by the number of household members.

Figures 1, 2 and 3 show the histograms of the three series[2]

The histograms for the three income variables are right-skewed. Both scaled incomes have very similar histogram. It seems that the simplest correction of the household size (that is weighting the total income by the number of household members) produces similar results to those obtained by weighting the total income by the Oxford scale, in terms of the shape of the empirical distribution.

3 Robust Methods for fitting a parametric model

3.1 The influence function

When fitting a parametric model to data, classical statistical theory assumes that the data at hand follow exactly the proposed parametric model. In practice however, we know that the sample might contain either gross errors or extreme observations that make the classical assumptions too strict. These

[2]In other to compare them, we have chosen the same scale for the X-axis. This supposes that the displayed range of the three series does not correspond to the original one, in that the highest incomes do not appear on the Figures. The highest income is 100 millions of pesetas for the total income, of around 45 millions for equivalised income and of 33 millions for the per capita income.

extreme observations, or indeed any slight model misspecification can ruin the analysis. Robust statistics on the other hand, deal with such situations by proposing estimators that are more stable in these conditions.

In robust theory, one assumes that the data are generated by a distribution in a neighbourhood of the parametric model. This neighbourhood can be described by the mixture distribution

$$F_\varepsilon := (1-\varepsilon)F_\theta + \varepsilon H \tag{1}$$

where F_θ is the parametric model, H is any distribution and $\varepsilon \in (0,1)$ is usually very small. Therefore, it is assumed that the data are generated from the parametric model F_θ with probability $(1-\varepsilon)$, which is large, and from the contamination distribution H with probability ε, which is small. The classical procedures are optimal under the assumption that the data are generated exactly by F_θ (i.e. $\varepsilon = 0$), but they are useless under slight violations of this assumption, i.e. under F_ε.

A procedure that gives reasonable results under (1) is called robust. In fact, "reasonable" means here that the estimator of the model's parameter θ is not influenced or is very little influenced by the presence of slight model deviations. One way of assessing this robustness property is by means of the *influence function (IF)*. The IF was first introduced by Hampel (1968) and Hampel (1974), and further developed in Hampel, Ronchetti, Rousseeuw, and Stahel (1986). It is widely used not only to study robustness properties of statistics such as estimators, but also to build robust estimators and robust test procedures (Hampel et al. 1986, Heritier and Ronchetti 1994, Victoria-Feser 1996). The IF measures the influence upon any statistic of an infinitesimal amount of contamination introduced in the model. Let T be a statistic that can be written as a functional of any distribution F, i.e. $T(F)$. We than suppose that F lies in a neighbourhood of the parametric model and study its effect on T. The worst effect is obtained when one considers the particular neighbourhood (see Hampel et al. 1986)

$$F_\varepsilon^{(z)} := (1-\varepsilon)F_\theta + \varepsilon H^{(z)} \tag{2}$$

where $H^{(z)}$ is the elementary distribution function which has unit point mass at z

$$H^{(z)}(x) = \begin{cases} 0 & \text{if } x < z \\ 1 & \text{if } x \geq z \end{cases} \tag{3}$$

The IF is then defined as

$$IF(z; F, F_\theta) := \frac{\partial}{\partial \varepsilon} T\left(F_\varepsilon^{(z)}\right)\Big|_{\varepsilon=0} \tag{4}$$

The IF can be thought of as a first order approximation of the bias on the statistic due to the introduction of the contamination. In other words, if one plots the maximum (absolute) bias of the statistic as a function of ε, then the maximum (absolute) value the IF can take is the slope of the tangent at $\varepsilon = 0$. This means that if the IF can be infinite, then so can be the bias of the statistic (Hampel et al. 1986).

3.2 Robust estimation

Classical parameter estimators like the maximum likelihood estimator (MLE) are typically not robust. It is easy to show it since their IF is proportional to the score function (Hampel et al. 1986). In particular, for all known parametric models used to describe income distributions, the MLE is not robust (Victoria-Feser 1993).

According to Huber (1977), a good statistical procedure should

1. achieve a reasonably good efficiency level for the assumed model,
2. be robust, that is, if there are small deviations from the model, the procedure should give similar results to those obtained when the model is exact (in terms of bias, asymptotic variance and the power for tests),

Several procedures have been proposed, among them we distinguish the following

1. those that try to broaden the scope of Non-Parametric Statistics
2. research and rejection of outlier by means of diagnostic tools,
3. robust theory.

We are interested in robust theory as developed by Huber (1964), Huber (1965), Hampel (1968) and Hampel et al. (1986). According to Hampel et al. (1986), there are two approaches in robust estimation. One of them

is Huber's minimax approach. In his important paper, Huber (1964) gives the foundation to the theory of robust estimation. He introduces a general class of estimators called M-estimator. They can be seen as a generalization of the maximum likelihood estimator since they are defined as the solution in θ of

$$\sum_{i=1}^{n} \psi(x_i; \theta) = 0 \tag{5}$$

where ψ is function satisfying very mild hypothesis (Huber 1964). For the MLE, $\psi(x;\theta)$ is simply the scores function $s(x;\theta)$. He also introduces the neighbourhood or gross-error model (1) and with these two ingredients Huber proposes as a robust estimator one that minimizes the maximum bias of the estimator due to the contamination.

The second approach, also called the *infinitesimal approach* is based on the IF and originally developed by Hampel (1968). Later, in Hampel et al. (1986) optimally IF-bounded estimators or optimal B-robust estimators (OBRE) are proposed. Actually, they are the best trade off between efficiency and robustness. Indeed, the most efficient but non robust estimator is the MLE. To make it robust, one necessarily sacrifices efficiency. The aim, is to find a robust estimator such that the efficiency loss is minimum. Hampel et al. (1986) propose finding the M-estimator with a bounded IF that has the minimum (asymptotic) variance. There are several optimal estimators depending on the way one chooses to bound the IF (Hampel et al. 1986). We propose here to use the standardized OBRE (which has proved to be numerically more stable) defined as the solution in θ of

$$\sum_{i=1}^{n} \psi_c^{A,a}(x_i; \theta) = 0 \tag{6}$$

where

$$\psi_c^{A,a}(x;\theta) = A(\theta)\left[s(x;\theta) - a(\theta)\right] W_c(x;\theta) \tag{7}$$

$$W_c(x;\theta) = min\left\{1; \frac{c}{\|A(\theta)\left[s(x;\theta) - a(\theta)\right]\|}\right\} \tag{8}$$

and A,a respectively a $dim(\theta) \times dim(\theta)$ matrix and a $dim(\theta)$-dimensional vector are determined by the equations

$$E\left[\psi_c^{A,a}(x;\theta)\psi_c^{A,a}(x;\theta)^T\right] = I \tag{9}$$

$$E\left[\psi_c^{A,a}(x;\theta)\right] = 0 \tag{10}$$

The OBRE is defined in terms of a weighted standardized scores function. It is robust because of the weights and since it depends on the scores function it keeps a level of efficiency close to the MLE. The matrix A and the vector a can be viewed as Lagrange multipliers for the constraints resulting from a bounded IF and Fisher consistency, i.e. (10). The constant c is the bound on the IF and can be interpreted as the regulator between robustness and efficiency: for a lower c one gains robustness but loses efficiency and vice versa. The most robust estimator can be obtained by choosing the lower bound $c = \sqrt{dim(\theta)}$. On the other hand, $c = \infty$ gives the MLE. Typically, c is chosen as to achieve a 95% efficiency at the model. In general this value depends on the model itself.

3.3 Computation of OBRE

To compute the OBRE requires solving (6) under the conditions (9) and (10). We propose here an algorithm based on the Newton-Raphson method. The main idea is to compute the matrix A and the vector a for a given θ by solving (9) and (10). This is followed by a Newton-Raphson step for (6) given these two matrices, and these steps are iterated until convergence.

More precisely, the algorithm can be defined by the following four steps:

Step 1: Fix a precision threshold η, an initial value for the parameter θ and initial values $a = 0$ and $A = J^{\frac{1}{2}}(\theta)^{-T}$ where

$$J(\theta) = \int s(x,\theta)s(x,\theta)^T dF_\theta(x)$$

is the Fisher information matrix.

Step 2: Solve the following equations with respect to a and A:

$$A^T A = M_2^{-1}$$

and

$$a = \frac{\int s(x,\theta)W_c(x,\theta)dF_\theta(x)}{\int W_c(x,\theta)dF_\theta(x)} \ ,$$

where

$$M_k = \int [s(x,\theta) - a][s(x,\theta) - a]^T W_c(x,\theta)^k dF_\theta(x) \ \ , \ k = 1, 2.$$

The current values of θ, a and A are used as starting values to solve the given equations.

Step 3: Compute M_1 and $\Delta\theta = M_1^{-1}\{\frac{1}{n}\sum_{i=1}^{n}[s(x_i,\theta) - a]W_c(x_i,\theta)\}$.

Step 4: If $|\Delta\theta| > \eta$ then $\theta \to \theta + \Delta\theta$ and return to Step 2, else stop.

The algorithm is convergent provided the starting point is near to the solution. In the first step, we can take for instance the MLE as initial value for the parameter. However, it can be argued that a more robust starting point like a trimmed moment estimator or a moment estimate based on the median and MAD[3] would be preferable. An alternative is to use the MLE as the starting point but then compute an OBRE with a high value of the bound c and then use the estimate as starting point for another more robust (lower value of c) estimator.

The choice of the initial values for the matrices A and a in the second step is due to the fact that these values solve the equations for $c = \infty$ (corresponding to the MLE). Notice that integration can be avoided in Step 2 by replacing F_θ by its empirical distribution function. This means replacing the integrals with averages over the sample.

4 Modelling Spanish income

4.1 Functional forms for income distribution

Since the early work of Pareto (1896), a large number of models have been proposed to describe the distribution of incomes. The most frequently used in applied work are the Pareto, the Lognormal (first proposed by Gibrat 1931), the Gamma (Salem and Mount 1974), and recently the Dagum (Dagum 1977, Dagum 1980) and Singh-Maddala (Singh and Maddala (1976)) models. The generalised Beta distribution of the second kind is also used in practice (McDonald 1984, Slottje 1989), partly because it includes as special cases most of the models for income distribution.

All these functions constitute a set of possible statistical models for describing the income distribution. However, the problem of the choice of the

[3]MAD denotes median absolute deviation and is often used as a robust estimator of the standard error in a normal model.

functional form still remains. One can choose the model because of its theoretical foundation, or because it has worked well in previous analyses. One may also argue that while a two-parameter model may be too simple to reflect the impact of economic fluctuations, a three- or four-parameter model may be appropriate. The question is then: is the cost of an additional parameter worthwhile, especially knowing that the variability of the estimators can be very large? An "ideal" income distribution model is unlikely to exist and the development of new functional forms for income distributions could continue indefinitely, because no model is perfectly adapted to every set of data. An alternative approach to the search for a better model could be to improve the estimation techniques and diagnostic checking. So we propose here the use of robust methods to estimate the parameters of the chosen model. The choice of the model itself is the subject of another paper (Victoria-Feser 1996).

Since robust statistics allow the chosen model not to be the exact representation of the data, a two parameter model should be sufficient to describe the distribution of incomes in Spain. This is not a general statement of course, since every sample should be studied separately, but we believe that in most cases two parameters should be enough. We consider here the most used two parameters models, namely the Gamma and the Lognormal distribution.

The advantage of using these two distributions is not only that they involve only two parameters but also these parameters have a clear economic interpretation. For the Lognormal distribution, as Aitchison and Brown (1957) point out, μ is the logarithm of the geometric mean income and σ^2 is the variance of the logarithm of income and it is related to inequality measurement in the sense that the larger σ^2, the larger the inequality measure. On the other hand, for the Gamma distribution, α is the parameter directly related to inequality measurement in such way that if the value of α increases the population in the left tail decreases (Haro Garcí 1993, p. 21) and thus the inequality decreases. The second parameters (μ and λ) are scale parameters.

The MLE of the parameters for these two models are not robust. Indeed, their densities and score functions are given by

- Gamma

$$f_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \ , \ 0 < x < \infty$$

$$s(x;\alpha,\lambda) = \begin{bmatrix} \log(\lambda) - \tilde{\Gamma}(\alpha) + \log(x) \\ \frac{\alpha}{\lambda} - x \end{bmatrix} ,$$

where $\alpha, \lambda > 0$, $\Gamma(\alpha) = \int_0^\infty e^{-u} u^{\alpha-1} du$ and $\tilde{\Gamma}(\alpha) = \frac{\partial}{\partial \alpha} \log \Gamma(\alpha)$.

- Lognormal

$$f_{\mu,\sigma^2}(x) = \frac{1}{x\sqrt{\sigma^2 2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\left(\log(x) - \mu\right)^2\right\} \ , \ x < \infty$$

$$s(x; \mu, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}(\log(x) - \mu) \\ -\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(\log(x) - \mu)^2 \end{bmatrix}$$

It is easy to see that the MLE for these models is not robust. That means that a single point can drive the MLE to an arbitrary extreme.

4.2 Application to Spanish data

In this subsection, we fit the Gamma and Lognormal distribution to the Spanish data (total income), by estimating the MLE and the OBRE. We also redo the old experiment of Salem and Mount (1974) of comparing the two distributions. In their paper, they conclude that the Gamma fits better the income distribution (using the MLE) in the United States for the years 1960 to 1969. We carry out the same model comparison using the three concepts of income assigned to the household in Spain, 1990-1991, with the MLE and the OBRE.

In order to give another picture of the income distribution, we also compute an inequality index. We choose the Theil (1967) index given by

$$I_{Theil} = E\left[\frac{x}{\mu} \log\left(\frac{x}{\mu}\right)\right]$$

where $\mu = E[x]$. It is computed by taking the expectation at the model with the estimated parameters.

The estimates of Theil index are given by

$$I_{Theil}(\hat{\alpha}) = \frac{1}{\hat{\alpha}} + \tilde{\Gamma}(\hat{\alpha}) - \log(\hat{\alpha})$$

for the Gamma distribution, and

$$\frac{1}{2}\hat{\sigma}^2$$

for the Lognormal distribution, where $\hat{\alpha}$ and $\hat{\sigma}^2$ denote estimators of α and σ^2 respectively. For any parametric model F_θ, it can be shown that the IF of the Theil index is proportional to the IF of the estimators of the parameters (see Cowell and Victoria-Feser 1996). Therefore, an unbounded IF for the estimators of the parameters implies an unbounded IF of the Theil index.

4.2.1 Gamma modelling

Victoria-Feser and Ronchetti (1994) stress the effect of a small amount of contamination in the MLE of the Gamma distribution and on the value of Theil index through a simulation study . In fact, the bias of the MLE in the Gamma model becomes substantial with only 1% of contamination. Is it also the case with the Spanish data?

	Total income		Equivalised income		Per capita income	
	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\lambda}$
MLE	2.8355	0.1314e-5	3.3819	0.3816e-5	3.0161	0.4265e-5
$c = 5$	2.9917	0.1412e-5	3.7809	0.4384e-5	3.386	0.4945e-5
$c = 2$	3.0006	0.1453e-5	4.0453	0.4810e-5	3.8133	0.5911e-5

Table 1: MLE and OBRE (Gamma) for Spanish data 1990-1991

Table 1 shows the MLE and OBRE of the Gamma distribution for the different types of income in Spain. From the table we can see that the difference between the MLE and the OBRE reduces when the value of the constant c increases. This confirm the fact that when we relax the constraint of bounded IF the OBRE coincides with the MLE. The difference between the MLE and the OBRE is considerably large for the equivalised and per capita income but not so large for the total income.

In Table 2 are presented the corresponding values for the Theil index. Here as well we can see that the more robust the estimator is, the less inequality is found. This is because one extreme income can substantially influence estimated inequality (see Cowell and Victoria-Feser 1996) and ro-

bust estimators downweight extreme incomes. The difference between the Theil indexes computed form the MLE and the OBRE ($c = 2$) being quite large (especially for the equivalised and per capita income), this suggests that the data contain some extreme returns.

	Total income	Eqivalised income	Per capita income
MLE	0.1661	0.1406	0.1567
$c = 5$	0.1579	0.1265	0.1405
$c = 2$	0.1575	0.1185	0.1254

Table 2: Theil's index for the Gamma model

From the plots comparing the two estimated distributions (MLE and OBRE ($c = 2$)) with the histograms of the empirical distribution (Figures 4, 5 and 6), we can see that for all these series the OBRE appears to fit the model better throughout the whole range, but is especially good in the middle of the distribution. It is worth noticing that the OBRE tends to underestimate the proportion of high incomes, whereas the MLE tends to overestimate it.

4.2.2 Lognormal modelling

Table 3 show the MLE and the OBRE of the Lognormal distribution for the different types of income in Spain. The difference between the MLE and the OBRE of the location parameter for different values of c is very small. However, the estimates of the variance of the logarithms differ considerably, especially for the equivalised and per capita income. The direction of the change is such that for more robust estimator, we get less inequality (i.e. lower σ^2). This is confirmed in Table 4 where the estimates of Theil index for the Lognormal distribution are presented. As with the Gamma distribution, the equivalised income tends to show less inequality.

	Total income		Equivalised income		Per capita income	
	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\mu}$	$\hat{\sigma}^2$
MLE	14.3977	0.3811	13.5397	0.2980	13.2941	0.3367
$c = 5$	14.3997	0.3724	13.5422	0.2791	13.2964	0.3163
$c = 2$	14.4287	0.3681	13.5301	0.2568	13.2889	0.2876

Table 3: MLE and OBRE (Lognormal) for Spanish data 1990-1991

	Total income	Equivalised income	Per capita income
MLE	0.1906	0.149	0.1835
$c = 5$	0.1862	0.1396	0.1582
$c = 2$	0.1841	0.1284	0.1438

Table 4: Theil's index for the Lognormal model

Figures 7, 8 and 9 show the MLE and OBRE ($c = 2$) fit for the three distributions. The two estimators lead to similar fits but the OBRE fits the middle of the range better.

Finally, Salem and Mount (1974) noted that the MLE of Lognormal model exaggerates the skewness for the USA income for the period 1960 and 1969. In Spain, this is true when we use the total income variable. However, with the OBRE, this is not the case.

4.2.3 Comparison of the fits

If we compare the estimated densities corresponding to the Lognormal (OBRE with $c = 2$) and the Gamma (OBRE with $c = 2$) distributions, we can see that the Lognormal model provides a better fit for the equivalised and per capita income and the Gamma model provides a better fit of the total income

(see figures 10, 11 and 12). None of the three estimated densities tend to exaggerate the skewness (this result contrasts with that obtained by Salem and Mount 1974).

5 Conclusion

Several aspects should be taken into account when income distribution is studied. We emphasizes the role of the definition of income assigned to the household and the method of estimation of parametric models. From the 'Encuesta de Presupuestos Familiares', three definitions of income can be used and two equivalence scales can be used, leading to 9 'income' variables. In this paper we have concentrated (for the moment) on total household income, scaled with the Oxford scale and with the number of people in the household.

We study the three series by fitting two commonly used parametric models to the data. Because income data in general and Spanish data in particular are not entirely reliable, we argue the use of robust methods of estimation. We then propose to compare fitted models using the classical MLE and using a robust estimator, namely the OBRE. This exercise leads to the conclusion that the data present a few extreme values which (slightly) distort the picture of the shape of the distributions when the Gamma and the Lognormal distributions are used as income distribution models and the MLE is used as estimator of the parameters. The OBRE then performs better than MLE in the two models and for the three income variables.

Salem and Mount (1974) experiment of comparing the Lognormal and the Gamma distribution is carried out for the three income variables on Spanish data using OBRE and MLE. The Gamma distribution (OBRE) fits the total income data better than the Lognormal model (OBRE), but the Lognormal (OBRE) fits the equivalised and per capita income better than the Gamma (OBRE) distribution. There is no evidence that the Gamma (OBRE) and the Lognormal (OBRE) tend to exaggerate the skewness for the total, equivalised and per capita income.

References

Aitchison, J. and J. C. Brown (1957). *The Log-Normal Distribution*. Cambridge, Massachussets: Cambridge University Press.

Alcaide, A. and J. Alcaide (1983). Distribución personal de la renta Española en 1980. *Hacienda Pública Española 85*, 485–509.

Atkinson, A. B. (1983). *The Economics of Inequality*. Oxford: Clarendom Press.

Carrascal, U. (1995). *Escales de Equivalencia de Consumo: Aplicación al caso Español*. Ph. D. thesis.

Coulter, F., F. Cowell, and S. Jenkins (1992). Differences in needs and assessement of income distributions. *Bulletin of Economic Research 44*, 77–124.

Cowell, F. A. and M.-P. Victoria-Feser (1996). Robustness properties of inequality measures. *Econometrica 64*, 77–101.

Dagum, C. (1977). A new model of personal income distribution: Specification and estimation. *Economie Appliquée 30*, 413–436.

Dagum, C. (1980). Generating systems and properties of income distribution models. *Metron 38*(3-4), 3–26.

Gibrat, R. (1931). *Les Inegalités Economiques*. Paris: Sirey.

Hampel, F. R. (1968). *Contribution to the Theory of Robust Estimation*. Ph. D. thesis, University of California, Berkeley.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association 69*, 383–393.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley.

Haro Garcí, J. (1993). Modelos de distribución de la renta biparamétricos. *Cuadernos de Ciencías Económicas y Empresariales 11*.

Heritier, S. and E. Ronchetti (1994). Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association 89*(427), 897–904.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics 35*, 73–101.

Huber, P. J. (1965). A robust version of the probability ratio test. *Annals of Mathematical Statistics 36*, 1753–1758.

Huber, P. J. (1977). Robust covariances. In S. S. Gupta and D. S. Moore (Eds.), *Statistical Decision Theory and Related Topics*, Volume 2, pp. 1753–1758. New York: Academic Press.

Instituto Nacional de Estadistica (1992). Encuesta de presupuestos familiares 1990-1991. Metodologa, Madrid.

Instituto Nacional de Estadistica (1993). Epf 90-91: Fichero para usarios. Technical report, Madrid.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica 52*, 647–664.

Pareto, V. (1896). Ecrits sur la courbe de la répartition de la richesse. In *Oeuvres complètes de Vilfredo Pareto*. Giovanni Busino. Librairie Droz, Genève, 1965.

Pazos, M. (1994). Variabilidad semanal de los gastos en la EPF. *Estadística Española 36*, 431–440.

Ruiz-Castillo, J. (1987). La medición de la pobreza y de la desigualdad en España. Estudios Económicos 42, Banco de España, Madrid.

Salem, A. B. Z. and T. D. Mount (1974). A convenient descriptive model of income distribution: The Gamma density. *Econometrica 42*, 1115–1127.

Sanz, B. (1992). La encuesta de presupuestos familiares 1990-1991. *Situación 2*(3), 151–166.

Singh, S. K. and G. S. Maddala (1976). A function for the size distribution of income. *Econometrica 44*, 963–970.

Slottje, D. J. (1989). *The Structure of Earnings and the Measurement of Income Inequality in theUS*. Amsterdam: North-Holland.

Theil, H. (1967). *Economics and Information Theory*. Amsterdam: North-Holland.

Victoria-Feser, M.-P. (1993). *Robust Methods for Personal Income Distribution Models*. Ph. D. thesis, University of Geneva, Switzerland. Thesis no 384.

Victoria-Feser, M.-P. (1996). Robust model choice test for non-nested hypothesis. *Journal of the Royal Statistical Society, Series B*. Under revision.

Victoria-Feser, M.-P. and R. Ronchetti (1994). Robust methods for personal income distribution models. *The Canadian Journal of Statistics 22*, 247–258.

Zubiri, I. (1985). Income inequality as a predictor of welfare inequality. SEED 40, Universidad de País Vasco. Instituto de Economia Pública.

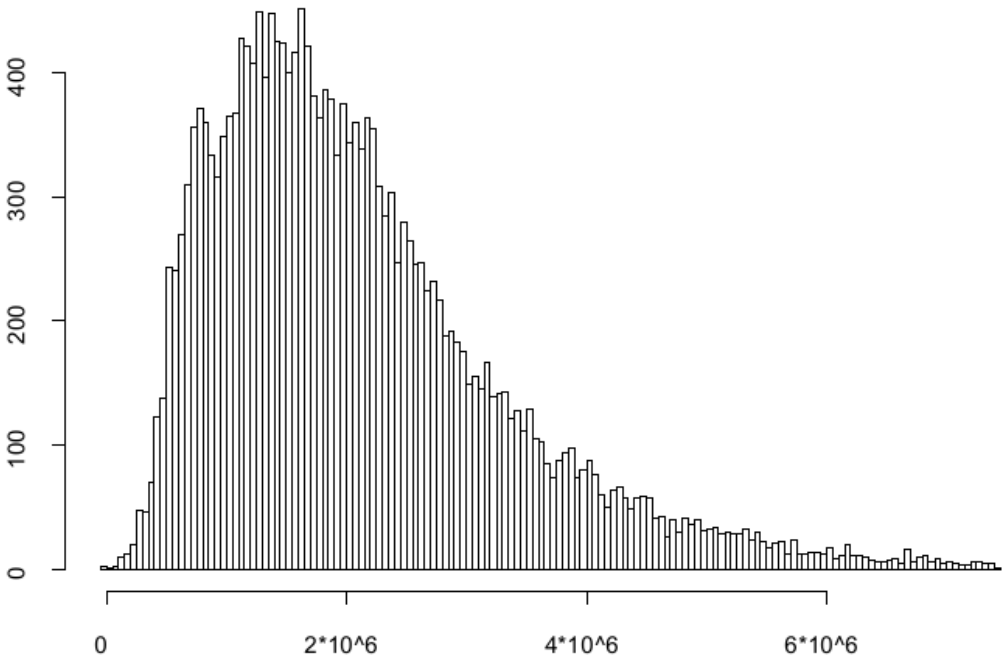

Figure 1: Total income in Spain, 1990-1991

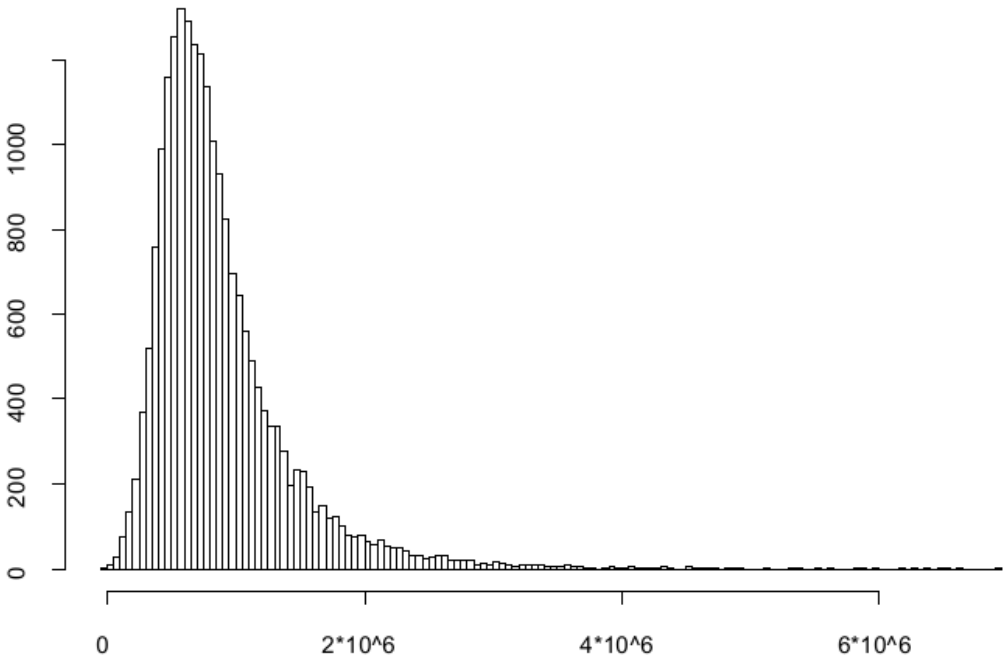

Figure 2: Equivalised income in Spain, 1990-1991

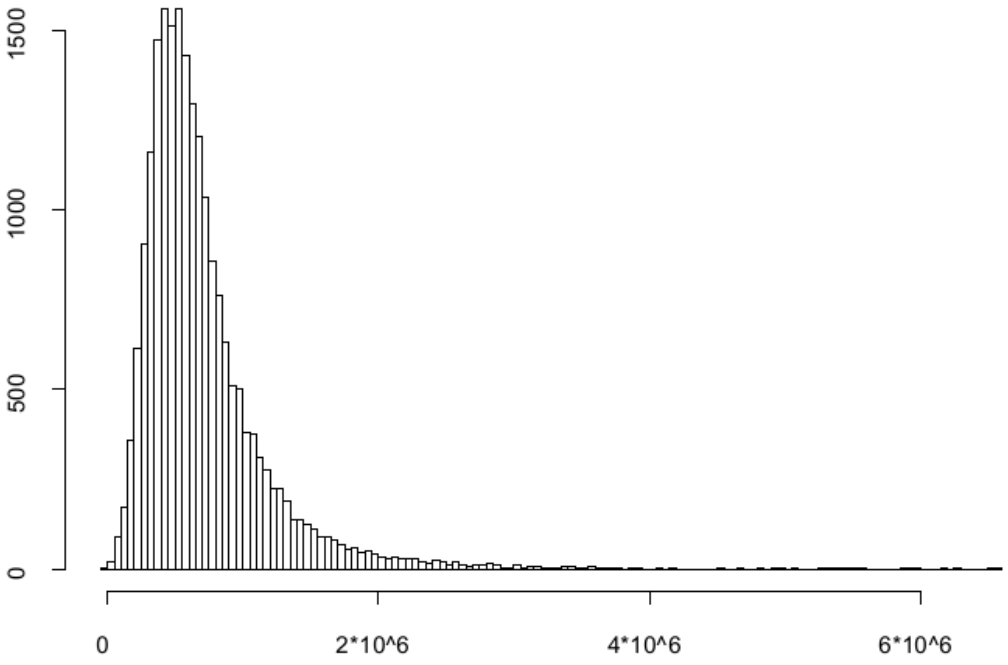

Figure 3: Per capita income in Spain, 1990-1991

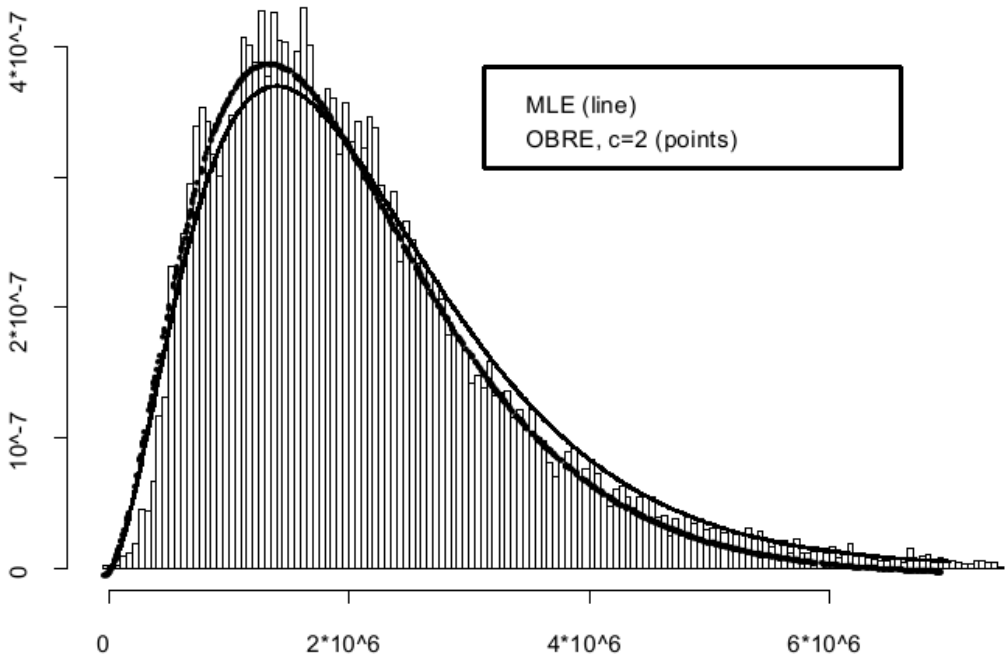

Figure 4: Gamma parametrization of Spanish data (total income)

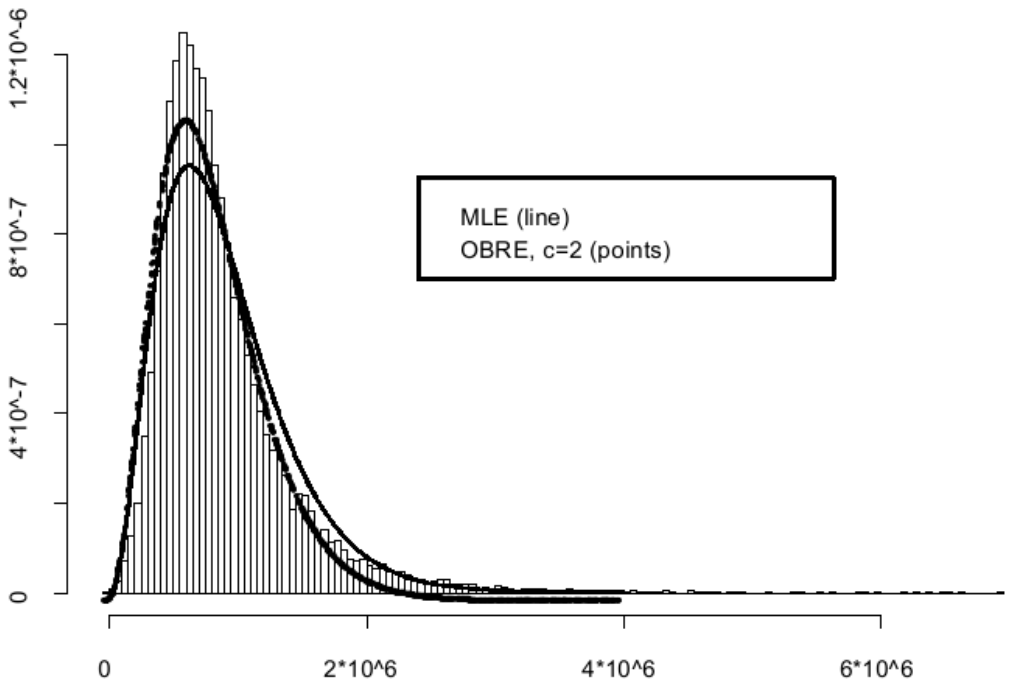

Figure 5: Gamma parametrization of Spanish data (equivalised income)

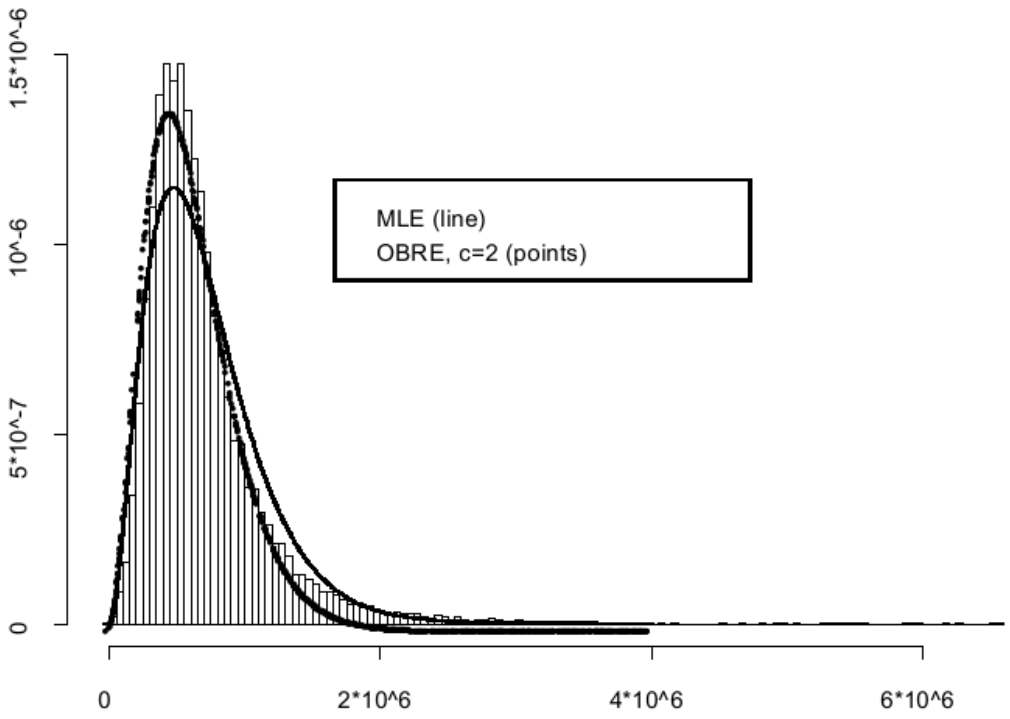

Figure 6: Gamma parametrization of Spanish data (per capita income)

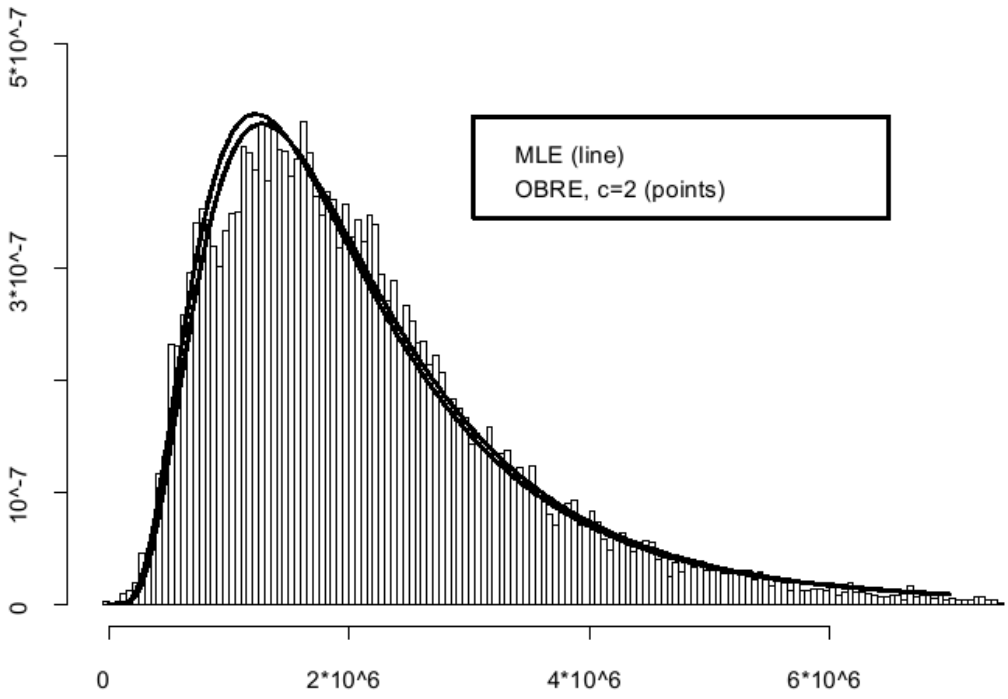

Figure 7: Lognormal parametrization of Spanish data (total income)

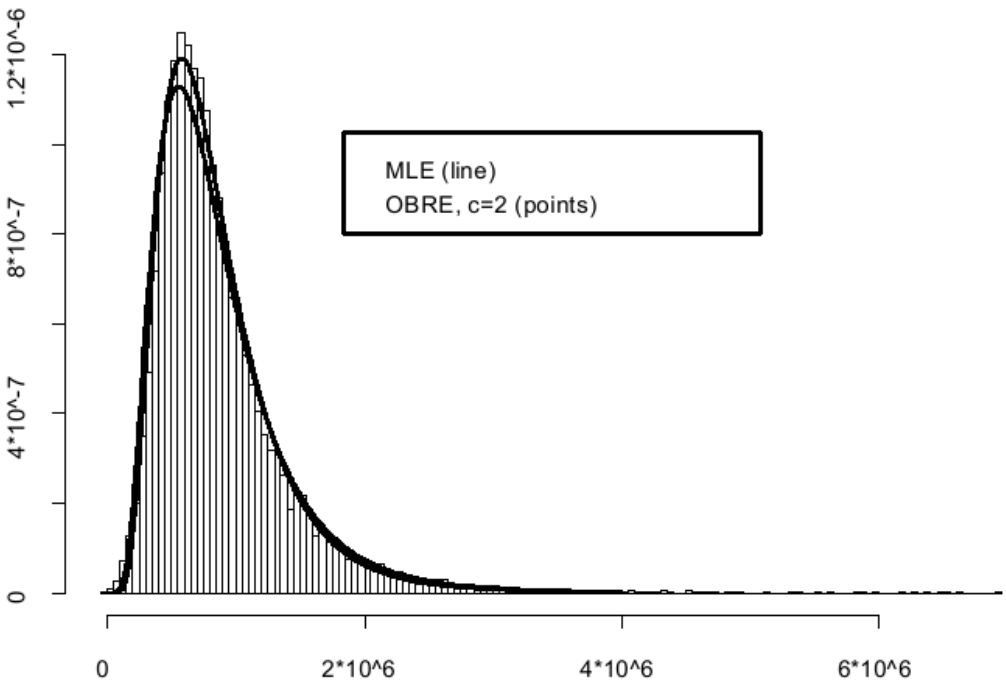

Figure 8: Lognormal parametrization of Spanish data (equivalised income)

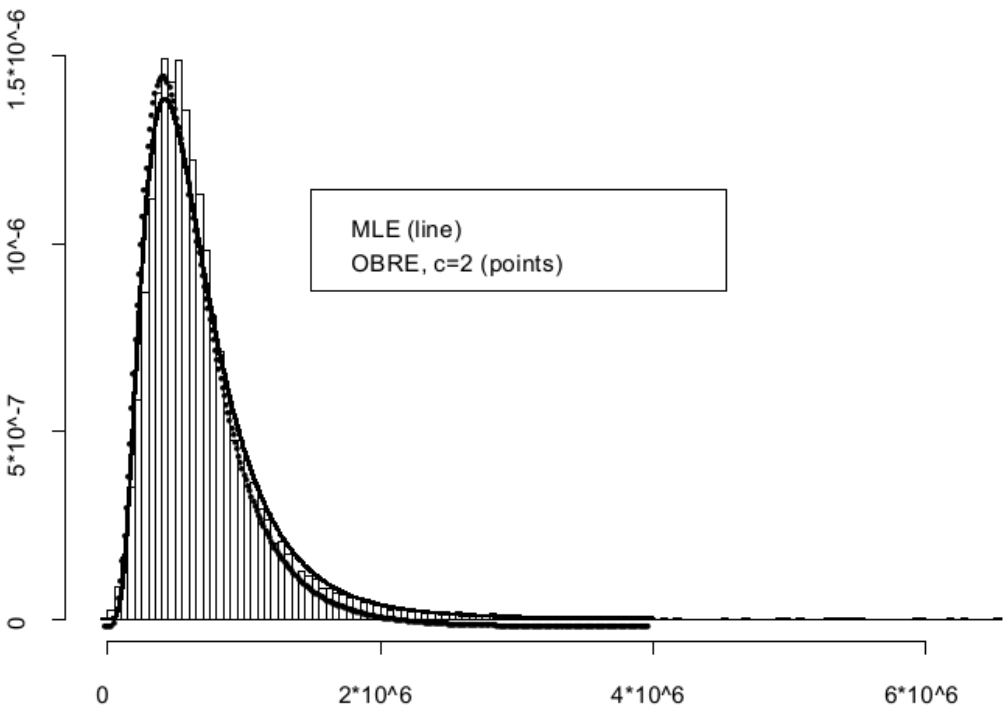

Figure 9: Lognormal parametrization of Spanish data (per capita income)

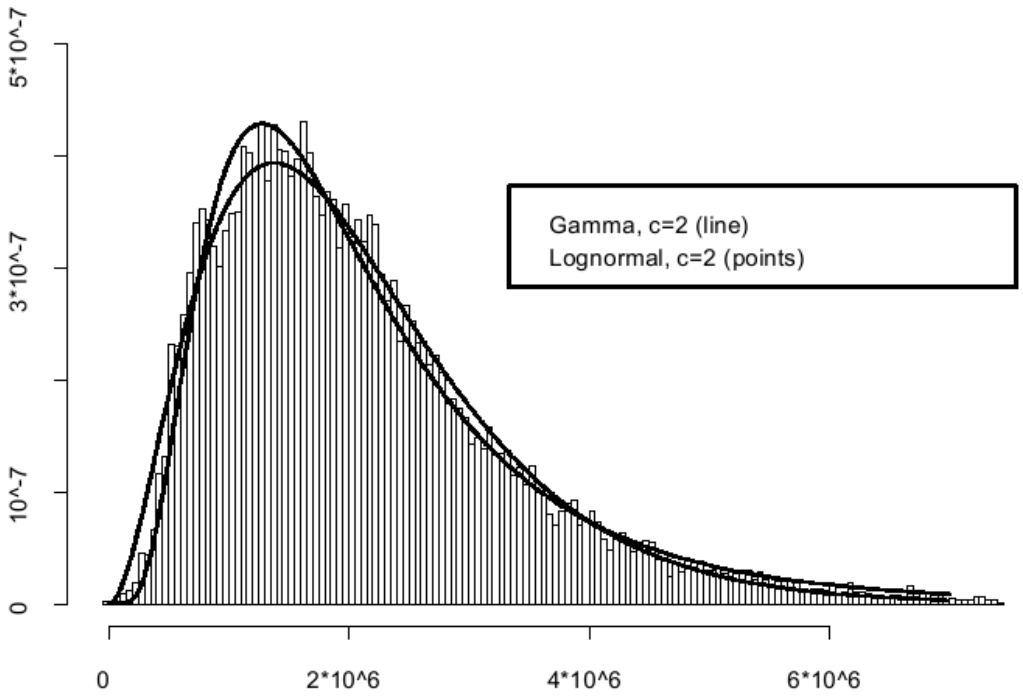

Figure 10: Robust parametrization of Spanish data (total income)

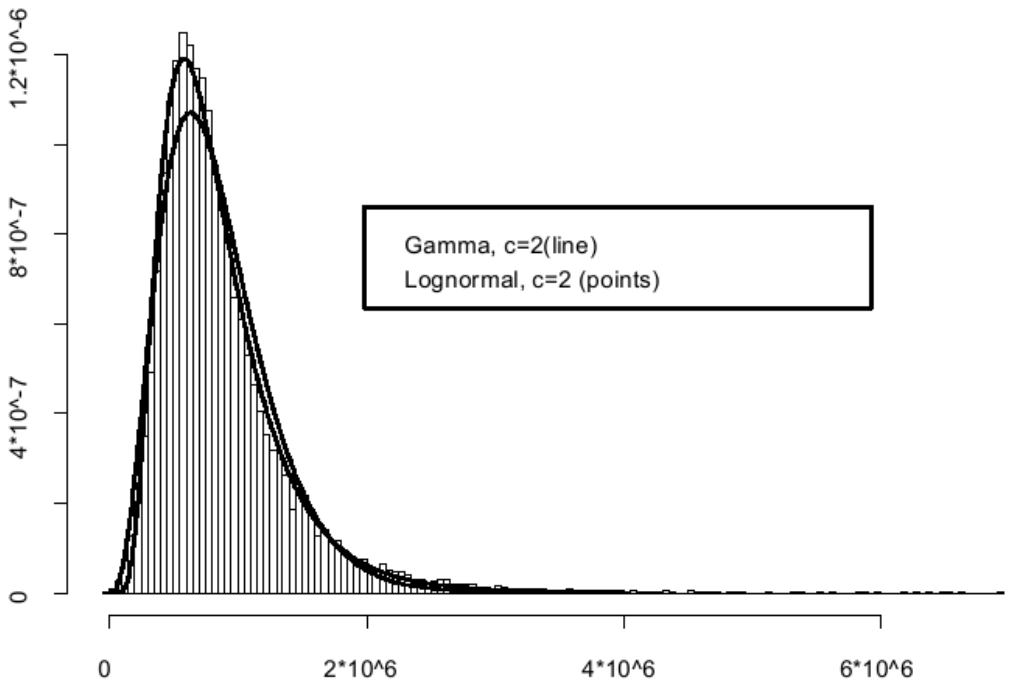

Figure 11: Robust parametrization of Spanish data (equivalised income)

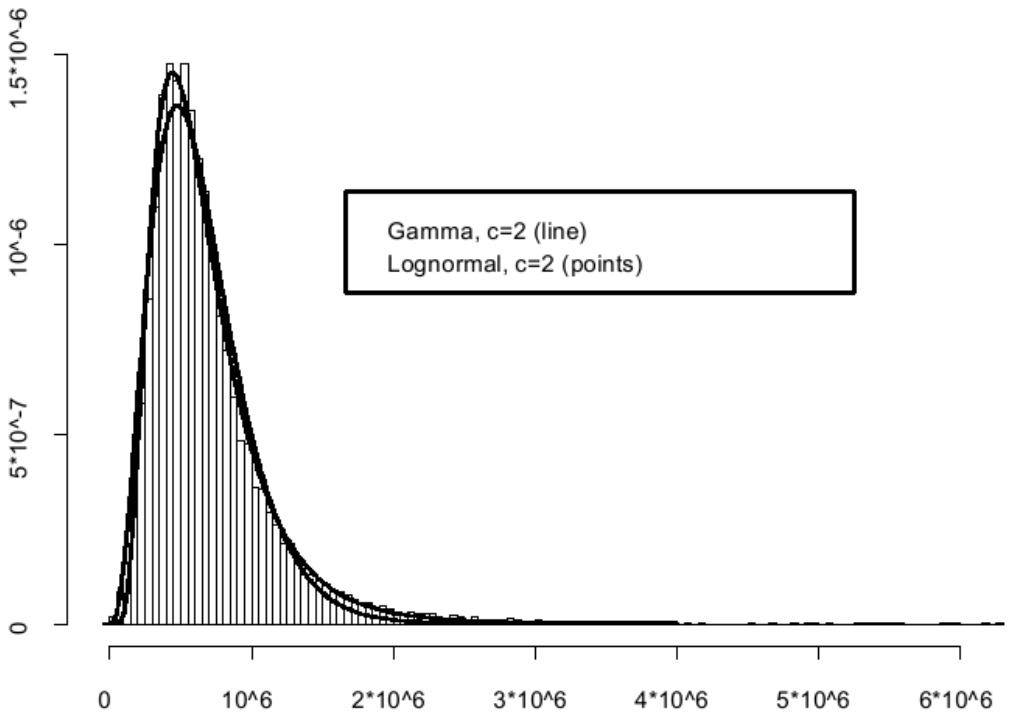

Figure 12: Robust parametization of Spanish data (per capita income)