A New Composite Lognormal-Pareto Type II Regression Model to Analyze Household Budget Data via Particle Swarm Optimization

Hande Konşuk Ünlü (hkonsuk@hacettepe.edu.tr)
Hacettepe University: Hacettepe Universitesi https://orcid.org/0000-0003-3572-0254

Research Article

Keywords: Composite regression model, Particle Swarm Optimization, Complex Survey Design, Household Budget Survey, Lognormal Regression Model, Lomax Regression Model

Posted Date: July 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-672186/v1

Noname manuscript No.
(will be inserted by the editor)

A New Composite Lognormal-Pareto Type II Regression Model to Analyze Household Budget Data via Particle Swarm Optimization

Hande Konşuk Ünlü

Received: date / Accepted: date

Abstract When data exhibits heavy-tailed behavior, traditional regression approaches might be inadequate or inappropriate to model the data. In such data analyses, composite models, which are built by piecing together two or more weighted distribution at specified threshold(s) are alternative models. When data contains covariate information, composite regression models could be used. In the existing literature, there is not many work done on this topic. The only study is Gan and Valdez (2018)'s paper. In this study, a novel Lognormal-Pareto Type II composite regression model was proposed. Particle swarm optimization (PSO) was performed to obtain model parameters of the proposed model. The proposed model was applied to model monthly consumption expenditure and affecting factors using National Household Budget Survey, conducted by Turkish Statistical Institute, annually. Since the sampling design of Household Budget Survey is stratified two-stage cluster sampling, the parameters were estimated under weighted data by updating the proposed model and estimation method, PSO. Additionally, the proposed regression model performance was compared with Lognormal and Lomax regression models. The results showed that the proposed model provided better fit to data.

Keywords Composite regression model · Particle Swarm Optimization · Complex Survey Design · Household Budget Survey · Lognormal Regression Model · Lomax Regression Model

Hande Konşuk Ünlü
Institute of Public Health, Hacettepe University, Ankara, Turkey
Tel.: +90-312-3053141
Fax: +90-312-3093699
E-mail: hkonsuk@hacettepe.edu.tr

1 Introduction

The researchers have still been working on finding a model that fits well into data where a single component distribution usually does not provide a better fit due to the properties of the data such as heterogeneity, multimodality, heavy-tailed etc.. Finite mixture models have been widely used by researchers when modeling the data involving heterogeneous sub-populations with a single distribution is almost impossible. Whether a uni-modal, bi-modal or even multi-modal distribution is used depends on the degree of heterogeneity among all sub-populations [29]. Especially after the publication of the monograph by McLachlan & Basford (1998) [20], the researchers working in various fields have been started to interest the potential usefulness of mixture models for inference and clustering [22]. Finite mixture distributions are obtained by taking a weighted average of a finite number of the same type of distributions with different parameter values or completely different distributions [9]. The numerous works which address the special topic reviews for finite mixture models in the literature have been carried out by many authors. Some of the studies are: the mixture of two normal distributions by Cohen (1967) [6], the mixture of t-distributions by Peel & McLachlan (2000) [25], the mixture of skew normal and skew t-distributions by Lee & McLachlan (2013) [19], the mixture of generalized hyperbolic distributions by Browne & McNicholas (2015) [4]. For detail information, interested readers can also refer to the books written for example by Everitt and Hand (1981) [13], McLachlan and Peel (2000) [21] and Frühwirth-Schnatter (2006) [14].

Finite mixture models have been extended by introducing relevant risk covariates usually associated with the distribution parameters or even the mixing proportions to allow for the modeling of heterogeneous regression relationship which are called as finite mixture regression model [8]. Some of the studies where finite mixture regression model is used are: a two-component finite mixture Lognormal regression model by Tooze et al. (2002) [33], a two-component finite mixture normal regression model by Yau et al. (2003) [34], a two-component finite mixture generalized linear mixed effect regression model by Hall and Wang (2005) [16].

Composite models constructed by joining the two (or more) weighted distributions at a given threshold value(s) can be interpreted as a finite mixture models with mixing weights c and $(1-c)$ [10]. These models are especially used to model data with fat-tailed behavior. Some of the studies where composite models used are: composite Lognormal-Pareto model by Cooray and Ananda (2005) [7], composite Lognormal-Pareto models by Scollnik (2007) [28], composite exponential-Pareto models by Teodorescu and Vernic (2009) [32], composite Pareto models by Teodorescu and Vernic (2013) [31], composite Stoppa models by Calderin-Ojeda and Kwok (2016) [5].

The composite regression models formed by modeling the distribution parameters using regression as in finite mixture regression models can be used in order to capture both fat-tailed behavior and heterogeneity in the data. Although there are many studies on composite models in the literature, there

is only one study on composite regression models by Gan and Valdez (2018). Gan and Valdez (2018) [15] proposed the use of two composite regression models consisting of three components to model Singapore auto claims data which capture both fat-tail behavior of data and the policyholders heterogeneity.

The composite regression models are quite new topic and many studies are needed in this area. In this study a novel composite regression model was proposed. The proposed composite regression model consists of two component-Lognormal and Pareto type II distributions. The proposed model was employed for analyzing monthly consumption expenditure and affecting factors using National Household Budget Survey, conducted by Turkish Statistical Institute (TurkStat), annually via proposed model. Monthly consumption expenditure data shows heavy-tailed behavior. Therefore traditional regression approach is not suitable to model this data. In the proposed model, particle swarm optimization was adapted to obtain the model parameters.

The rest of the paper is organized as follows. Section 2 involves the proposed Lognormal-Pareto Type II regression model along with a brief information about composite models. Particle swarm optimization (PSO) which is used to estimate parameters of the proposed model is introduced with its steps in Section 3. Section 4 presents the simulation study to assess the performance of PSO. The application of the proposed model to household budget survey data is given in Section 5. It is concluded with a brief discussion of the results of this study in Section 6.

2 Composite Models

Composite models consisting of two parts, a sub-threshold and an over-threshold distribution, can be more suitable to model data which has skewed distribution with a heavier right tail than the models including the Gamma, the Lognormal, the Weibull and the Pareto [3].

Assuming that data Y involves two sub-populations respectively with a light-tailed probability density function $f_1(y)$ and a heavy-tailed probability density function $f_2(y)$ then the random variable Y has the following probability density function:

$$f(y) = \begin{cases} cf_1(y), & y \leq \theta \\ cf_2(y), & y > \theta \end{cases} \tag{1}$$

where θ is threshold and $c = \dfrac{1}{\int_0^\theta f_1(y)\,dy + \int_\theta^\infty f_2(y)\,dx}$ is the normalizing constant.

A smooth probability density function can be obtained by imposing the continuity condition $f_1(\theta-) = f_2(\theta+)$ and differentiability conditions $f_1^{'}(\theta-) = f_2^{'}(\theta+)$ at threshold [7]. Here, as the mixing weight is fixed and known priorly.

Scollnik (2007) [28] expressed the model given in Eq. 1 as convex combination of two probability density functions:

$$f(y) = \begin{cases} c f_1^*(y) & , y \leq \theta \\ (1-c) f_2^*(y) & , y > \theta \end{cases} \tag{2}$$

where $0 \leq c \leq 1$ and $f_1^*(y)$ and $f_2^*(y)$ are adequate truncations of $f_1(y)$ and $f_2(y)$ given respectively:

$$\begin{aligned} f_1^*(y) &= \frac{f_1(y)}{\int_0^\theta f_1(y)\,dy} = \frac{f_1(y)}{F_1(\theta)} && , y \leq \theta \\ f_2^*(y) &= \frac{f_2(y)}{\int_\theta^\infty f_2(y)\,dy} = \frac{f_2(y)}{1 - F_2(\theta)} && , y > \theta \end{aligned} \tag{3}$$

The mixing weight c is a function of the threshold θ and the parameters of $f_1(y)$ and $f_2(y)$ varying in the interval $[0,1]$ obtained by imposing the continuity and differentiability conditions at threshold.

2.1 Composite Regresion Models

This section involves the probability density functions for composite Lognormal-Pareto Type II regression model along with its log-likelihood function.

2.1.1 Lognormal-Pareto Type II Composite Regression Model

Suppose the first and second components in Eq. 3 have the Lognormal distribution with location parameter $\mu > 0$ and shape parameter $\sigma > 0$ and Pareto type II distribution with scale parameter $\alpha > 0$, shape parameter $\lambda > 0$ and location parameter $\theta > 0$ respectively and the scale parameters of these distributions are modeled using regression. Then, the Composite Lognormal-Pareto Type II regression model is:

$$f(y_i) = \begin{cases} c\,\Phi\left(\dfrac{ln(\theta) - \mu(\boldsymbol{\beta}, \boldsymbol{X})}{\sigma}\right)^{-1} \dfrac{1}{\sigma y_i \sqrt{2\pi}} \exp\left(-\dfrac{1}{2}\left(\dfrac{ln(y_i) - \mu(\boldsymbol{\beta}, \boldsymbol{X})}{\sigma}\right)^2\right) & , y_i \in (0, \theta] \\ (1-c)\,\dfrac{\lambda}{\alpha(\boldsymbol{\beta}, \boldsymbol{X})}\left[1 + \dfrac{y_i - \theta}{\alpha(\boldsymbol{\beta}, \boldsymbol{X})}\right]^{-(\lambda+1)} & , y_i \in (\theta, \infty) \end{cases} \tag{4}$$

Here, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n), i = 1, 2, \ldots, n$ are the values of response variable. Let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p)$ denotes the $n \times p$ matrix of values of explanatory variables, with the first column assumed to be a 1 to accommodate the estimation of an intercept, $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T\right)$ denotes the component specific parameter vectors. The explanatory variables are incorporated through link function $g(\cdot)$.

In this setting, the scale parameters μ and α are related to the linear predictors through identity and exponential link functions as $\mu(\boldsymbol{\beta}, \boldsymbol{X}) = \boldsymbol{\beta}_1^T \boldsymbol{X}[\mathbb{1}\{y_i \leq \theta\}]$ and $\alpha(\boldsymbol{\beta}, \boldsymbol{X}) = \exp\left(\boldsymbol{\beta}_2^T \boldsymbol{X}[\mathbb{1}\{y_i > \theta\}]\right)$ respectively.

Log-likelihood function for the proposed model given in Eq. 4 is as follows:

$$\begin{aligned}\log L = \sum_{i=1}^{n} I_{(0,\theta]}(y_i) \Bigg[log(c) - \frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2}\left(\frac{\log(y_i) - \mu(\boldsymbol{\beta}, \boldsymbol{X})}{\sigma}\right)^2 \\ -\Phi\left(\frac{\log(\theta) - \mu(\boldsymbol{\beta}, \boldsymbol{X})}{\sigma}\right)\Bigg] + I_{(\theta,\infty)}(y_i)\{\log(1-c) + \log(\lambda) \\ -\log(\alpha(\boldsymbol{\beta}, \boldsymbol{X})) - (\lambda+1)\log\left(1 + \frac{y_i - \theta}{\alpha(\boldsymbol{\beta}, \boldsymbol{X})}\right)\Bigg\} \end{aligned} \tag{5}$$

Let $E_1(Y|Y \leq \theta)$ and $E_2(Y|Y > \theta)$ denote the contribution of Lognormal distribution and Pareto type II distribution to the expected value of response variable, respectively. The expectation of response variable with probability density function could be obtained as $E(Y) = E_1(Y|Y \leq \theta) + E_2(Y|Y > \theta)$.

$$E_1(Y|Y \leq \theta) = c\exp(\mu(\boldsymbol{\beta}, \boldsymbol{X}))\frac{\Phi\left(\frac{ln(\theta) - \mu(\boldsymbol{\beta}, \boldsymbol{X}) - \sigma^2}{\sigma}\right)}{\Phi\left(\frac{ln(\theta) - \mu(\boldsymbol{\beta}, \boldsymbol{X})}{\sigma}\right)}\exp\left(\frac{\sigma^2}{2}\right) \tag{6a}$$

$$E_2(Y|Y > \theta) = (1-c)\left(\theta + \frac{\alpha(\boldsymbol{\beta}, \boldsymbol{X})}{\lambda - 1}\right) \tag{6b}$$

As seen from Eq. 6a, the contribution of Lognormal distribution to the expected value is proportional to $\exp(\mu(\boldsymbol{\beta}, \boldsymbol{X}))$ and other terms in the Eq. 6a do not affect the proportionality. Similarly, in Eq 6b, the contribution of Pareto type II distribution to the expected value is proportional to $\alpha(\boldsymbol{\beta}, \boldsymbol{X})$. Consequently, the effect of explanatory variables on the response variable Y could be interpreted proportionally.

3 Particle Swarm Optimization

Heuristic methods such as Genetic Algorithm, Differential Evolution Algorithm and PSO inspired by the events in nature are powerful optimization methods used to solve a complex system. PSO was developed by Eberhart and Kennedy in 1995 [18] inspired by the social behavior of bird flocking or a school of fish. It has been observed that the actions of the animals moving in herd, often randomly, such as food and safety, enable them to reach their goals more easily. When a school of fish, flocks of birds and other social animals were examined, it was seen that these animals interacted in search of food, and when one found a food, the others turned their position to the location of the food and updated their speed accordingly without breaking

away from the herd. Therefore, PSO was developed by using social interaction between birds.

In PSO, individuals are called "particles" (i.e. each bird in swarm) in the *swarm*. The change of particle position in the search space is based on the social and psychological tendency of individuals to imitate the success of other particles. Therefore, the changes of particles in the group are affected by the experience or knowledge of their neighbors. Hence, the search behavior of particles is affected by the search behaviors of other particles in the group. The result of modeling this social behavior is that the search process causes particles to randomly return to previously successful regions in the search space [23].

The PSO starts with randomly generated a set of solutions called as "swarm". In the swarm, each solution defined as "particle". Particles (birds) are flown through the multidimensional search space. Each particle has its own position and speed information that guides its flight and a fitness value obtained by the fitness function (log-likelihood function - in this study) to be optimized. Each particle adjusts its position according to its own and its neighbors' previous experiences. PSO is mainly based on approximating the position of particles in the swarm to the best positioned particle of the swarm. This approximation approach develops randomly, and most of the time, particles in the swarm positioned better by their new movements than their previous position and this process continues iteratively until the fitness function is optimized. At each iteration, each particle is updated according to the two "best" values. One of them is personal best (pbest), the best fitness obtained by particle so far, and the other is the global best (gbest), the global solution obtained so far by whole swarm [27],[30].

The algorithm basically consists of the following steps;

1. The initial swarm is created with randomly generated starting locations and velocities for prespecified swarm size ($\mathbb{K}$).
2. Fitness values are calculated for each particle.
3. Inertia weight (w), c_1 and c_2 are set.
4. The personal best (p_{best}) is found for each particle.
5. The global best (g_{best}) is found among all the particles in the swarm.
6. The velocity (Eq. 7a) and the location (Eq. 7b) are updated at each iteration using the formulas given below:

$$\begin{aligned} v_{kj}\left(t+1\right) = {} & wv_{kj}\left(t\right) + c_1 r_{1,kj}\left(t\right)\left[p_{best,kj}\left(t\right) - x_{kj}\left(t\right)\right] \\ & + c_2 r_{2,kj}\left[g_{best,j}\left(t\right) - x_{kj}\left(t\right)\right] \end{aligned} \tag{7a}$$

$$X_{kj}\left(t+1\right) = X_{kj}\left(t\right) + v_{kj}\left(t+1\right) \tag{7b}$$

where $X_k\left(t\right) = \left(x_{k1}, x_{k2}, \ldots, x_{kj}, \ldots, x_{kd}\right), k = 1, 2, \ldots, \mathbb{K}$ shows the location of k^{th} particle and $v_{kj} = \left(v_{k1}, v_{k2}, \ldots, v_{kj}, \ldots, v_{kd}\right), k = 1, 2, \ldots, \mathbb{K}$ represents the velocity of k^{th} particle at the t^{th} iteration in the d-dimensional search

space. Here, c_1 and c_2 are accelerating factors used to scale cognitive and social components respectively, w is the inertia weight and $r_{1,kj}$ and $r_{2,kj}$ are random numbers generated from Uniform distribution, $U[0,1]$. $p_{best,k}$ *(personal best)*, represents the best location of the particle i experienced so far and g_{best} *(global best)*, represents the best location among all the particles in the swarm experienced so far [1],[2],[12],[24].

7. The steps 1-6 are repeated until stopping criterion is satisfied.

Velocity regulation is very important in the PSO algorithm. In the Eq. 7a, the coefficient w, inertia weight is used to limit the velocity of the particles. When $w > 1$, the velocities of particles increase with time to the maximum speed and the swarm diverges. On the other hand, small inertia values facilitates the local search but weaken the global search capability of PSO [17], [30]. Inertia weight was taken as constant in the early studies in the literature. Later in the paper by Eberhart and Shi(2000) [11], linear decreasing value for the inertia weight (w) was proposed.

In this study, the PSO is used to estimate the parameters of the models given in Eq. 4. PSO is an efficient algorithm in terms of being simple to use, not needing any score functions or their derivatives. Its computational simplicity and lower elapsed time compared to traditional optimization algorithms makes PSO so appealing especially for complex models. By taking into consideration that the score functions of the likelihood function are extremely complex and it is difficult to obtain solutions using traditional methods, it was decided to use PSO method to obtain the model parameter estimates.

4 Simulation Study

To determine the value of inertia weight is vital in the PSO algorithm. Therefore, the simulation setting was carried out as two-stage process. In the first stage, the simulation study was designed in order to determine optimal hyper parameters of PSO algorithm. In the second stage, another simulation study was designed to obtain model parameters given in Eq 4 using PSO method with its hyper parameters which were chosen in stage 1.

In both stages, the number of explanatory variables was taken as 4, component specific parameter vectors were taken as $\beta_1 = (0.69, -0.29, 0.41, 0.92)$, $\beta_2 = (0.41, 0.83, 0.56, -0.51)$ and the value of threshold was specified as 10. Explanatory variable matrix was generated using normal distribution with mean 2 and standard deviation 0.5. The values of dependent variables being lower and higher than the threshold θ were generated respectively from Lognormal distribution with location parameter $\mu(\boldsymbol{\beta}, \boldsymbol{X}) = \boldsymbol{\beta}_1^T \boldsymbol{X_1}$ and shape parameter $\sigma = 0.1$ and from Pareto type II distribution with scale parameter $\alpha(\boldsymbol{\beta}, \boldsymbol{X}) = \exp\left(\boldsymbol{\beta_2}^T \boldsymbol{X_2}\right)$ shape parameter $\lambda = 2$ and location parameter $\theta = 10$. Here, the mixing weight was taken as 0.50.

In the first stage of the simulation study, grid search was implemented in order to optimize inertia weight parameter of PSO algorithm. The inertia weight values were taken fixed as $(0.1, 0.3, 0.5, 0.7, 0.9, 1.1)$. In addition, inertia weight value was taken dynamically decreasing linearly from 1.4 to 0.4. The other hyper parameters of PSO algorithm used in first stage were given below:

- Sample size was taken as 300.
- Swarm size was taken as 20.
- The coefficients c_1 and c_2 were taken equal to 1.49.

Each scenario was repeated 500 times. The maximum number of iterations of PSO algorithm was taken as 20 and the simulation was stopped when PSO algorithm reached to the maximum iteration number. For each scenario, Akaike Information Criteria (AIC) and total elapsed time were calculated. The scenarios for different inertia weight values were compared according to AIC-values.

PSO algorithm given detailed in Section 3 and all data generation process were carried out using `R` ver.4.0.0 [26] with self-written code. The results of first stage of the simulation study were given in Table 1.

[Table 1 about here]

As seen from the Table 1, the total elapsed times for different scenarios were found similar. When the AIC values were compared, it was concluded that the dynamic inertia weight decreasing linearly from 1.4 to 0.4 had the lowest AIC value. Therefore, the PSO algorithm with dynamic inertia weight value was used to estimate the model parameters in Eq. 4 for second stage.

In the second stage of the simulation study, the scenarios given below were run in order to obtain parameter estimates of the model in Eq 4 according to maximum likelihood method.

- Sample size was taken as 300, 500 and 1000.
- Swarm size was taken as 20 and 40.
- The coefficients c_1 and c_2 were taken equal to 1.49.
- As a result of first stage of the simulation, inertia weight values were dynamically taken from 1.4 to 0.4 in linearly decreasing way.

Each scenario was repeated 1500 times. The maximum number of iterations of PSO algorithm was taken as 20 and the simulation was stopped when PSO algorithm reached to the maximum iteration number. For each scenario, Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), mean of total absolute bias and total elapsed time were calculated. The results of the proposed model obtained for different sample size and swarm size were compared. Using the simulated data, the parameters of the Lognormal regression and Lomax regression models were also estimated. The AIC and BIC values for these models were compared with the proposel model, Lognormal-Pareto Type II composite regression model. The results were given in Table 2, Table 3

and Table 4.

[Table 2 about here]

[Table 3 about here]

[Table 4 about here]

The parameter estimates and their MSE values for sample sizes 300, 500 and 1000 were given respectively in Table 2, Table 3 and Table 4 for swarm size 20 and 40. The AIC and BIC values were close to each other for each sample sizes. The results indicated that the AIC, BIC and total absolute bias values were decreasing as swarm size was increasing for each sample size. Furthermore, total elapsed time was increasing as the swarm size was increasing for each sample size. It was seen that swarm size 40 in each sample size had the lowest AIC and BIC values. When the AIC and BIC values obtained for different swarm size of each sample size were compared, it was seen that swarm size 40 had smallest AIC and BIC values. As the sample size was increasing, MSE of parameter estimates were decreasing. Therefore, it could be concluded that the parameter estimates were consistent.

The parameters of the Lognormal regression and Lomax regression models were also estimated using the simulated data. The probability density function with their mean structures of these models were given below:

Lognormal Regression Model:

$$f(y_i) = \frac{1}{\sigma y_i \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{ln(y_i) - \mu(\boldsymbol{\gamma}, \boldsymbol{X})}{\sigma}\right)^2\right), \qquad y_i > 0 \tag{8}$$
$$\mu(\boldsymbol{\gamma}, \boldsymbol{X}) = \boldsymbol{\gamma}^T \boldsymbol{X} = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3$$

where $\sigma > 0$ is shape parameter, $\mu > 0$ is scale parameter and $\boldsymbol{\gamma}^T = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ is parameter vector.

Lomax Regression Model:

$$f(y_i) = \left(1 + \frac{y_i}{\alpha(\boldsymbol{\beta}, \boldsymbol{X})}\right)^{-\lambda} \qquad , y_i > 0 \tag{9}$$
$$\alpha(\boldsymbol{\gamma}, \boldsymbol{X}) = \exp\left(\boldsymbol{\gamma}^T \boldsymbol{X}\right) = exp(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3)$$

where $\lambda > 0$ is shape parameter, $\alpha > 0$ is scale parameter and $\boldsymbol{\gamma}^T = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ is parameter vector.

The BIC values of Lognormal regression model for sample sizes 300, 500 and 1000 were 2418.2109, 4037.6981 and 8001.3424 respectively. The BIC values

of Lomax regression model for sample sizes 300, 500 and 1000 were 4045.0708, 6744.2876 and 13234.6622 respectively. Comparing the results of Lognormal and Lomax regression models with those of the proposed Lognormal-Pareto Type II regression model, it was seen that the difference between the BIC values between Lognormal regression model and the proposed model, $BIC_{\Delta} = BIC_{LogN} - BIC_{Prop}$ and the difference between Lomax regression model and the proposed model, $BIC_{\Delta} = BIC_{Lomax} - BIC_{Prop}$ for all sample sizes were greater than 10 ($BIC_{\Delta} >> 10$). Hence, it was concluded that the proposed model provided a better fit to the data than Lognormal and Lomax regression models.

5 Household Budget Data

The proposed Lognormal-Pareto Type II regression model was deployed to analyze monthly consumption expenditure and related factors using National Household Budget Survey in 2018 which was conducted by Turkish Statistical Institute (TurkStat). The household budget survey was first conducted in 1964. Since 2004, it has been carried out annually to reveal consumption patterns and income levels of individuals and households by socio-economic groups, rural & urban areas and regions of Turkey. In this study the survey which was conducted between January 1^{st} and December 31^{st}, 2018 on 1296 households was used. The sampling design of the Household Budget Survey was stratified two-stage cluster sampling method. Therefore, parameter estimates were obtained using sampling weights. The micro data of 2018 Household Budget Survey consists of three data sets which are Household, Individual and Consumption Expenditure. These three data sets were linked to each other with one-to-one matching on the "Household ID", which was common in all data sets in order to select variables to be used in the analysis. After mapping process, the data was filtered to the individuals who live alone in household since it is more practical to evaluate the relationship between monthly consumption expenditure and related factors for individual.

The explanatory variables used in the analysis including continuous variables and categorical variables with reference levels were given below:

- Age
- Gender: male (reference), female
- Education level: not completed any school (reference), primary education, secondary education, bachelor's degree and higher
- Marital status: single (reference), married
- Whether to have private health insurance or not: no (reference), yes
- Employment status: not working (reference), working
- Total annual individual income
- Ownership status of any real-estate: no (reference), yes
- Smoking habit: no (reference), yes

- Habit of playing games of chance such as lottary, numbers pool, tickets for horse race etc.): no (reference), yes
- Habit of eating out: no (reference), yes
- Habit of buying daily newspaper: no (reference), yes
- Having cable TV: no (reference), yes
- Habit of shopping via internet: no (reference), yes
- Whether to use credit card or not: no (reference), yes
- Savings habit: no (reference), yes

The explanatory variables were thought to be correlated with the dependent variable which was *monthly consumption expenditure*. The categorical variables were coded using leave-one-out method relative to the reference level.

As mentioned before, the parameters were estimated under weighted data by taking the sampling method into consideration. Besides, parameter estimates under unweighted data were also obtained and compared with the weighted results. Similar to the simulation study, parameter estimates were also obtained under weighted and unweighted data using Lognormal and Lomax regression models.

In order to provide ease of operation in the analysis, estimates were obtained by dividing the monthly consumption expenditure by 100. The histograms for weighted and unweighted of the monthly consumption expenditure were given in Figure 1.

[Figure 1 about here]

Summary statistics calculated under weighted and unweighted data for the variables to be used in the regression model were given in Table 5.

[Table 5 about here]

When the results in Table 5 were examined, it was seen that the summary statistics obtained for weighted and unweighted data were generally similar with some differences. As the data was obtained from the stratified two stage cluster sampling, summary statistics obtained for weighted data were interpreted. The average age of the participants in the study was 55.97 with standard deviation 18.47, 42.7% of them were male, 39.0% of them had primary education and 94.1% of them were single. More than half of the participants was working (57.3%), and about two-thirds of the participants had own property. 6.2% of them had the habit of buying daily newspapers. Only 10% of them had private insurance and one third of them had smoking habit. In addition, 11.7% had a cable TV subscription and 5.1% was playing games of chance. The average monthly consumption expenditure of the participants was 28.33 with standard deviation 23.35 and varied between 1.09 and 247.98. The average annual income of individuals was 2832.84 Turkish Lira (TL) with standard deviation 2334.91 and varies between 109 and 24798.19. 60% of the

participants had savings habit, 39% of them had a credit card and 77.5% of them had the habit of online shopping.

The parameter estimates for the Lognormal-Pareto Type II composite regression model, which were obtained for weighted and unweighted data using variables considered as related to the average monthly consumption expenditure, were given in Table 6. It was observed that the histograms of the monthly consumption expenditure for weighted and unweighted data were very similar which could be seen in Figure 1. Similarly, the summary statistics obtained under weighted and unweighted data were also found similar. However, when the parameter estimates given in Table 6 were examined, it was observed that the results of weighted and unweighted data were very different from each other. In particular, the estimate of threshold value (θ) for weighted data was around 9, while the estimate for the unweighted data was around 40. However, since the sampling method was stratified two-stage cluster sampling, parameters should be estimated under weighted data. Therefore, by examining models generated under weighted data, it was seen that the model with swarm size 40 had lowest AIC and BIC values and gave better results. The reason to obtain parameter estimates for weighted and unweighted data separately was to show that parameter estimates for weighted data were quite different from the parameter estimates for unweighted data, although both distributions were found quite similar in Figure 1.

Similar to the simulation study, parameter estimates were also obtained by using Lognormal and Lomax regression models. Using Lognormal regression model, AIC_w and BIC_w values for weighted data were found as 80190799.46 and 80191062.23, respectively. Using Lomax regression model, AIC_w and BIC_w values for the weighted data were found as 54178799.27 and 54179062.04, respectively. There was a significant difference between the BIC values of Lognormal-Pareto Type II composite regression model, Lognormal and Lomax regression models. This result confirmed that the most compatible model with the data was the proposed Lognormal-Pareto Type II composite regression model.

The parameter estimates under weighted data were presented in Table 6. The threshold value θ for the Lognormal-Pareto Type II composite regression model with swarm size 40 was estimated as 9.8795. Households, whose monthly consumption expenditure was equal to or less than 9.8795 TL were called "low expenditure class", those above 9.8795 TL were called "high expenditure class". This classification were used in following comments.

[Table 6 about here]

When parameter estimates in the weighted data for swarm size 40 in Table 6 were examined, it was seen that the age variable had an increasing effect to the amount of average monthly consumption expenditure in low, but it had a decreasing effect to the amount of high expenditure classes.

For low expenditure class, the average monthly consumption expenditure of women compared to men was exp (−0.0134) times greater. For high expenditure class, women spent exp (0.0129) times more than men.

For low expenditure classes, the average monthly consumption expenditure of people who had the level of primary or secondary education compared to those who did not completed any school was exp (0.1024) and exp (0.055) times higher, respectively, while people who had bachelor's degree was exp (−0.0534) times greater. For high expenditure class, the average monthly consumption expenditure was increasing as the level of education increased.

For both expenditure classes, the average monthly consumption expenditure was lower for married individuals compared to single ones.

The average monthly consumption expenditure of the participants in low expenditure class, who had the habit of dining out was exp (0.1321) times greater than the individuals who did not have this habit. For high expenditure class, those who had the habit of dining out spent exp (0.2792) times more than those who did not have the habit of dining out.

Those who had savings habit for both expenditure classes, had a higher amount of expenditure than those who did not have savings habit.

For both expenditure classes, the average monthly consumption expenditure was higher for individuals who were working compared to those who were not working.

The annual income variable had an increasing effect to the amount of average monthly consumption expenditure for both expenditure classes.

The average monthly consumption expenditure of the participants in low and high expenditure classes, who had their own property was exp (0.0999) and exp (0.3266) times greater, respectively, than the individuals who did not have any property.

For both expenditure classes, the average monthly consumption expenditure was higher for individuals who had habit of online shopping compared to those who did not have this habit.

Those who had smoking habit for both expenditure classes, had a higher amount of expenditure than those who did not smoke.

The average monthly consumption expenditure of the participants in low expenditure class, who had habit of buying daily newspaper was exp (0.0339) times higher than the individuals who did not have habit of buying newspaper. Similarly, in high expenditure classes, having habit of buying daily newspaper had an increasing effect {exp (0.1732)} on average monthly consumption expenditure.

For both expenditure classes, the average monthly consumption expenditure was higher for individuals who had Cable TV subscription compared to those who did not.

Those who had habit of playing games of chance in low expenditure class had a lower amount of expenditure than those who did not. On the other hand, those who had habit of playing games of chance in high expenditure class had exp (0.2336) times more than those who did not have this habit.

The average monthly consumption expenditure of the participants in low expenditure class, who had credit card was exp (0.0412) higher than the individuals who did not have credit card. Similarly, for spending amounts in high expenditure classes, having credit card had an increasing effect {exp (0.1267)} on average monthly consumption expenditure.

Those who had private health insurance in low expenditure class had a lower amount of expenditure than those who did not, while those who had private health insurance in high expenditure class had exp (0.2429) times more than those who did not have private health insurance.

Summary statistics calculated for low and high expenditure classes in Table 7.

[Table 7 about here]

6 Conclusion

The main purpose of this study was to propose a novel composite regression model for heavy-tailed data. The proposed model consisted of two components which were Lognormal and Pareto type II distributions. In the literature, the only study on composite regression models was Gan and Valdez (2018)'s paper. They [15] proposed Gamma-Pareto and Pareto-Type I Gumbel spliced regression models to estimate auto insurance claims. The proposed model in this study is different from the models proposed by Gan and Valdez. Consequently, this study provide an important contribution to the existing literature.

PSO algorithm is capable of dealing with a large number of parameters. PSO algorithm is a powerful tool for solving complex systems. Moreover, algorithm's elapsed time is quite short. In this study, PSO algorithm was used to obtain parameter estimation of the proposed model. PSO algorithm was implemented in two stages. The first stage was designed to find optimal parameter value for inertia weight which is one of the hyper parameter of PSO algorithm. According to the first stage result of the simulation study, the dynamic inertia weight decreasing linearly from 1.4 to 0.4 had the lowest AIC value compared to the other fixed inertia values. As a result of the first stage, the dynamically decreasing inertia weight was used for the second stage of the simulation setting. In the second stage of the simulation setting was performed to obtain parameter estimates of the proposed model using PSO algorithm. As expected, MSE values of parameter estimates were decreasing as the sample size was increasing for each swarm size. In addition, the proposed composite Lognormal-Pareto type II regression model was compared with those of Lognormal regression and Lomax regression models according to AIC and BIC values. The results of the simulation study showed strong support in favor of the proposed model.

After that, monthly consumption expenditure and affecting factors using National Household Budget Study conducted by Turkish Statistical Institute

(TurkStat), annually were modeled via the proposed composite Lognormal-Pareto type II regression model. The explanatory variables in the model were age, gender, level of education, marital and employment status, whether to have private health insurance or not, habits of smoking, buying daily newspaper, playing games of chance, eating out, online shopping and savings, own any property, having Cable TV subscription and credit card. To set up a regression model for providing better fit to the data was challenging. As seen in Figure 1, monthly consumption expenditure had heavy-tailed behavior. Therefore, standard regression models were deficient to capture the relationship between monthly consumption expenditure and its related variables. The results of Household Budget Study data showed that the proposed regression model indicated more favorable fit than Lognormal and Lomax regression models.

Additionally, the sampling design of Household Budget Study was stratified two-stage clustering sampling. Accordingly, it was another challenge to obtain parameter estimates under weighted data. To handle with this challenge, model equation and PSO algorithm were updated by taking into account sampling weights. Finally, parameters were estimated for weighted data and the effect of explanatory variables on average monthly consumption expenditure was explained in detail.

Acknowledgements I would like to thank the Turkish Statistical Institute (TurkStat) in Turkey for their valuable contributions in providing the data for this research. I would also like to thank Dr. Ayten Yiğiter for valuable comments to help improve the presentation of this paper.

Declarations

Funding : Not applicable
Conflicts of interest/Competing interests : Not applicable
Availability of data and material : The data was provided Turkish Statistical Institute (TurkStat) with restricted use. I am not allowed to share with third parties.
Code availability : Not applicable

References

1. Acitas, S., Aladag, C.H., Senoglu, B.: A new approach for estimating the parameters of weibull distribution via particle swarm optimization: an application to the strengths of glass fibre data. Reliability Engineering & System Safety **183**, 116–127 (2019)
2. Amoshahy, M.J., Shamsi, M., Sedaaghi, M.H.: A novel flexible inertia weight particle swarm optimization algorithm. PloS one **11**(8), e0161558 (2016)
3. Bølviken, E.: Computation and modelling in insurance and finance. Cambridge University Press (2014)
4. Browne, R.P., McNicholas, P.D.: A mixture of generalized hyperbolic distributions. Canadian Journal of Statistics **43**(2), 176–198 (2015)

5. Calderín-Ojeda, E., Kwok, C.F.: Modeling claims data with composite stoppa models. Scandinavian Actuarial Journal **2016**(9), 817–836 (2016)
6. Cohen, A.C.: Estimation in mixtures of two normal distributions. Technometrics **9**(1), 15–28 (1967)
7. Cooray, K., Ananda, M.M.: Modeling actuarial data with a composite lognormal-pareto model. Scandinavian Actuarial Journal **2005**(5), 321–334 (2005)
8. Dietz, E., Böhning, D.: Statistical inference based on a general model of unobserved heterogeneity. Statistical modelling pp. 75–82 (1995)
9. van Dijk, B.: Essays on finite mixture models. Ph.D. thesis, Erasmus University Rotterdam , Tinbergen Institute (2009)
10. Dominicy, Y., Sinner, C.: Distributions and composite models for size-type data. Advances in statistical methodologies and their application to real problems p. 159 (2017)
11. Eberhart, R.C., Shi, Y.: Comparing inertia weights and constriction factors in particle swarm optimization. In: Proceedings of the 2000 congress on evolutionary computation. CEC00 (Cat. No. 00TH8512), vol. 1, pp. 84–88. IEEE (2000)
12. Engelbrecht, A.P.: Particle swarm optimization: Global best or local best? In: 2013 BRICS congress on computational intelligence and 11th Brazilian congress on computational intelligence, pp. 124–135. IEEE (2013)
13. Everitt, B.S., Hand, D.J.: Finite Mixture Distributions. Springer Netherlands (1981)
14. Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer Science Business Media, LLC, New York (2006)
15. Gan, G., Valdez, E.A.: Fat-tailed regression modeling with spliced distributions. North American Actuarial Journal **22**(4), 554–573 (2018)
16. Hall, D.B., Wang, L.: Two-component mixtures of generalized linear mixed effects models for cluster correlated data. Statistical Modelling **5**(1), 21–37 (2005)
17. Karaboğa, D.: Yapay Zeka Optimizasyon Algoritmalari. Nobel Akademi Yayıncılık (2017)
18. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of ICNN'95-International Conference on Neural Networks, vol. 4, pp. 1942–1948. IEEE (1995)
19. Lee, S.X., Mclachlan, G.J.: On mixtures of skew normal and skew t-distributions. Advances in Data Analysis and Classification **7**(3), 241–266 (2013)
20. McLachlan, G.J., Basford, K.E.: Mixture models: Inference and applications to clustering. Marcel Dekker, New York (1988)
21. McLachlan, G.J., Peel, D.: Finite Mixture Distributions. John Wiley Sons, Inc., Canada (2000)
22. McLachlan, G.J., Peel, D.: Finite mixture models, *Probability and Statistics – Applied Probability and Statistics Section*, vol. 299. Wiley, New York (2000)
23. ÖZSAĞLAM, M.Y., ÇUNKAŞ, M.: Optimizasyon problemlerinin çözümü için parçaçık sürü optimizasyonu algoritması. Politeknik Dergisi **11**(4), 299–305 (2008)
24. Parsopoulos, K.E., Vrahatis, M.N., et al.: Particle swarm optimization method for constrained optimization problems. Intelligent Technologies–Theory and Application: New Trends in Intelligent Technologies **76**(1), 214–220 (2002)
25. Peel, D., McLachlan, G.J.: Robust mixture modelling using the t distribution. Statistics and computing **10**(4), 339–348 (2000)
26. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2020). URL `http://www.R-project.org/`
27. Saravanan, R.: Manufacturing optimization through intelligent techniques. CRC Press (2006)
28. Scollnik, D.P.: On composite lognormal-pareto models. Scandinavian Actuarial Journal **2007**(1), 20–33 (2007)
29. Shen, J.: Finite mixture regression models and applications: Detection limit and goodness-of-fit test. Ph.D. thesis, The School of Public Health University of Medicine and Dentistry of New Jersey and the Graduate School—New Brunswick (2011)
30. Shi, Y., Eberhart, R.C.: Empirical study of particle swarm optimization. In: Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406), vol. 3, pp. 1945–1950. IEEE (1999)
31. Teodorescu, S., Vernic, R.: On composite pareto models. Mathematical Reports **15**(65), 11–29 (2013)

32. Teodorescu, S., Vernic, R., et al.: Some composite exponential-pareto models for actuarial prediction. Romanian Journal of Economic Forecasting **12**(4), 82–100 (2009)
33. Tooze, J.A., Grunwald, G.K., Jones, R.H.: Analysis of repeated measures data with clumping at zero. Statistical methods in medical research **11**(4), 341–355 (2002)
34. Yau, K.K., Lee, A.H., Ng, A.S.: Finite mixture regression model with random effects: application to neonatal hospital length of stay. Computational statistics & data analysis **41**(3-4), 359–366 (2003)

Table 1: Comparison of Performance According to Different Inertia Weights

Inertia Weight	**AIC**	**Total Elapsed Time**
0.1	1695.0239	0.0807
0.3	1694.3400	0.0919
0.5	1696.7118	0.0958
0.7	1701.5213	0.0965
0.9	1703.7745	0.0971
1.1	1708.9244	0.0905
Dinamic Inertia[1]	1649.8723	0.9405

[1]Linearly decreasing values from 1.4 to 0.4.

Table 2: The parameter estimates of simulation setting for defined sample size (N=300) and swarm size (20 and 40)

	Sample Size=300			
Parameters	$\mathbb{K} = 20$		$\mathbb{K} = 40$	
	Estimates	**MSE**	**Estimates**	**MSE**
θ	13.0288	17.4269	12.5873	13.9573
σ	0.2157	0.0169	0.1926	0.012
$\beta_{1,0}$	1.1398	0.4065	1.067	0.3475
$\beta_{1,1}$	-0.1937	0.0193	-0.2117	0.0174
$\beta_{1,2}$	0.264	0.038	0.283	0.0324
$\beta_{1,3}$	0.6485	0.1492	0.6966	0.1228
λ	1.144	1.0708	1.1336	0.8763
$\beta_{2,0}$	0.2667	4.5075	0.087	4.8465
$\beta_{2,1}$	0.539	0.5503	0.5259	0.5818
$\beta_{2,2}$	0.4969	0.3416	0.5137	0.4165
$\beta_{2,3}$	-0.3111	0.6298	-0.2452	0.6372
AIC	1651.8909		1632.1402	
BIC	1692.6325		1672.8818	
Total Absolute Bias	7.4167		6.9128	
Total Elapsed Time	0.734		3.1439	

Table 3: The parameter estimates of simulation setting for defined sample size (N=500) and swarm size (20 and 40)

	Sample Size=500			
Parameters	$\mathbb{K} = 20$		$\mathbb{K} = 40$	
	Estimates	**MSE**	**Estimates**	**MSE**
θ	13.1073	20.2107	12.2596	11.6345
σ	0.2088	0.0148	0.1838	0.0099
$\beta_{1,0}$	1.1543	0.3898	1.0675	0.3040
$\beta_{1,1}$	-0.1886	0.0193	-0.2093	0.0165
$\beta_{1,2}$	0.2574	0.0369	0.2803	0.0311
$\beta_{1,3}$	0.6348	0.1448	0.6930	0.1111
λ	1.1226	0.8334	1.1239	0.7935
$\beta_{2,0}$	0.2813	3.4771	0.1034	3.6314
$\beta_{2,1}$	0.5218	0.4450	0.4990	0.5016
$\beta_{2,2}$	0.4766	0.3271	0.5330	0.2868
$\beta_{2,3}$	-0.2768	0.4973	-0.2555	0.5125
AIC	2748.5058		2711.2291	
BIC	2794.8665		2757.5898	
Total Absolute Bias	7.2494		6.2714	
Total Elapsed Time	0.4233		3.2149	

Table 4: The parameter estimates of simulation setting for defined sample size (N=1000) and swarm size (20 and 40)

	Sample Size=1000			
Parameters	$\mathbb{K} = 20$		$\mathbb{K} = 40$	
	Estimates	**MSE**	**Estimates**	**MSE**
θ	12.7440	14.7185	11.8923	7.1641
σ	0.2041	0.0135	0.1783	0.0086
$\beta_{1,0}$	1.1602	0.3924	1.0610	0.2983
$\beta_{1,1}$	-0.1911	0.0182	-0.2123	0.0152
$\beta_{1,2}$	0.2587	0.0369	0.2804	0.0298
$\beta_{1,3}$	0.6317	0.1460	0.7008	0.1023
λ	1.1229	0.9290	1.1288	0.8102
$\beta_{2,0}$	0.2989	2.3017	0.1156	2.4509
$\beta_{2,1}$	0.4954	0.3822	0.5062	0.3955
$\beta_{2,2}$	0.4828	0.1931	0.5158	0.1945
$\beta_{2,3}$	-0.2792	0.3881	-0.2667	0.3926
AIC	5481.7268		5409.7275	
BIC	5535.7121		5463.7128	
Total Absolute Bias	6.6097		5.5792	
Total Elapsed Time	0.8619		3.3834	

Table 5: The summary statistics for weighted and unweighted data

	Unweighted Statistics		Weighted Statistics	
	Count	**Percent**	**Count**	**Percent**
Gender				
Male	423	36.8	1602617	42.7
Female	726	63.2	2152018	57.3
Education level				
Not completed any school	332	28.9	992776	26.4
Primary education	471	41.0	1464698	39.0
Secondary education	118	10.3	463461	12.3
Bachelor's degree and higher	228	19.8	833701	22.2
Marital status				
Single	1085	94.4	3533899	94.1
Married	64	5.6	220736	5.9
Employment status				
Not working	735	64.0	2151497	57.3
Working	414	36.0	1603138	42.7
Whether to have private health insurance or not				
No	1054	91.7	3335590	88.8
Yes	95	8.3	419045	11.2
Ownership status of any real-estate				
No	317	27.6	1320054	35.2
Yes	832	72.4	2434581	64.8
Smoking habit				
No	836	72.8	2576204	68.6
Yes	313	27.2	1178431	31.4
Habit of buying daily newspaper				
No	1094	95.2	3522847	93.8
Yes	55	4.8	231788	6.2
Having cable TV				
No	1047	91.1	3314034	88.3
Yes	102	8.9	440601	11.7
Habit of playing games of chance				
No	1102	95.9	3561653	94.9
Yes	47	4.1	192982	5.1
Habit of eating out				
No	729	63.4	2081599	55.4
Yes	420	36.6	1673036	44.6
Habit of shopping via internet				
No	334	29.1	845344	22.5
Yes	815	70.9	2909291	77.5
Having credit card				
No	789	68.7	2282910	60.8
Yes	360	31.3	1471726	39.2
Savings habit				
No	433	37.7	1502920	40.0
Yes	716	62.3	2251716	60.0

Table 5: The summary statistics for weighted and unweighted data (continued)

	Unweighted Statistics	**Weighted Statistics**
Monthly Expenditure Consumption		
Mean+SD	2503.32 ± 2220.60	2832.84 ± 2334.91
Min-Max	109.00 - 24798.19	109.00 - 24798.19
Median (IQR)	1945.13 (1723.75)	2260.23 (1958.45)
Age		
Mean ± SD	58.97 ± 17.60	55.97 ± 18.47
Min-Max	22.00 - 97.00	22.0 - 97.00
Median (IQR)	62.00 (26.00)	59.00 (33.00)
Annual Income		
Mean ± SD	29512.22 ± 28992.45	33822.48 ± 31546.51
Min-Max	1800.00 - 393316.00	1800.00 - 393316.00
Median (IQR)	20480.00 (22120.00)	24950.00 (27304.00)

SD:Standard Deviation, Min:Minimum, Max:Maximum, IQR: Interquartile Range

Table 6: The weighted and unweighted parameter estimates of proposed Lognormal-Pareto Type II Composite Regression model for Household Budget Data

		Unweighted Estimates		**Weighted Estimates**	
		$\mathbb{K}=20$	$\mathbb{K}=40$	$\mathbb{K}=20$	$\mathbb{K}=40$
	Threshold, θ	36.8089	36.8410	9.8688	9.8795
LOGNORMAL	σ	0.4894	0.4488	0.4062	0.4868
	Intercept_1	2.0278	1.9449	1.5309	1.9980
	Age_1	-0.0026	-0.0015	0.0022	0.0016
	Savings Habit_1 (Yes)	0.2081	0.1767	-0.0063	0.0033
	Employment_1 (Working)	-0.0401	-0.0331	0.1372	0.0556
	Annual income_1	0.00003	0.00003	0.00001	0.00001
	Ownership_1 (Yes)	0.1134	0.0540	0.1808	0.0999
	Shopping_1 (Yes)	0.0228	0.0725	0.0977	0.1199
	Smoking Habit_1 (Yes)	0.0050	0.0113	0.0367	0.0388
	Newspaper_1 (Yes)	-0.0235	0.0086	-0.0414	0.0339
	Cable TV_1 (Yes)	0.1937	0.0669	0.1174	0.0800
	Chance Games_1 (Yes)	0.2984	0.1469	-0.0168	-0.0083
	Eating out_1 (Yes)	0.1055	0.1923	0.1130	0.1321
	$\text{Health insurance}_1$ (Yes)	0.0465	0.0088	-0.2138	-0.0565
	Credit Card_1 (Yes)	0.1303	0.1159	0.0477	0.0412
	Gender_1 (Female)	0.0680	0.0678	-0.0137	-0.0134
	Education_1 (Primary)	0.1435	0.1084	0.0906	0.1024
	Education_1 (Secondary)	0.0450	0.0052	0.0765	0.0555
	Education_1 (Bachelor)	0.0341	0.1129	-0.0294	-0.0534
	Marital_1 (Married)	-0.0541	-0.0784	0.1365	-0.0564
PARETO TYPE II	λ	7.8656	1.1341	1.7075	1.9284
	Intercept_2	4.3013	1.6936	29.4277	23.5244
	Age_2	-0.0059	-0.0129	-0.0028	-0.0017
	Savings Habit_2 (Yes)	0.4167	0.4484	0.4858	0.1828
	Employment_2 (Working)	-0.0870	-0.0989	0.0849	0.0288
	Annual income_2	0.00003	0.000003	0.000001	0.000002
	Ownership_2 (Yes)	0.3174	0.2697	0.2920	0.3266
	Shopping_2 (Yes)	-0.3676	-0.1703	-0.2874	0.0062
	Smoking Habit_2 (Yes)	0.0699	-0.0235	0.0289	0.0457
	Newspaper_2 (Yes)	0.0242	0.0910	0.1780	0.1732
	Cable TV_2 (Yes)	0.0956	0.1447	0.0603	0.1741
	Chance Games_2 (Yes)	0.1553	0.2164	0.3399	0.2336
	Eating out_2 (Yes)	0.2081	0.2682	0.3156	0.2792
	$\text{Health insurance}_2$ (Yes)	0.4287	0.1367	0.1890	0.2429
	Credit Card_2 (Yes)	-0.0950	0.1875	0.1099	0.1267
	Gender_2 (Female)	-0.0514	-0.4681	0.0084	0.0129
	Education_2 (Primary)	0.3832	0.1853	0.1858	0.1589
	Education_2 (Secondary)	0.1287	0.2742	0.2141	0.2190
	Education_2 (Bachelor)	0.6694	0.5828	0.5097	0.4923
	Marital_2 (Married)	-0.0781	0.0864	0.0341	-0.0300
	AIC	8761.5902	8766.3907	4265864.3855	4228302.0147
	BIC	8968.5028	8973.3032	4266403.0640	4228840.6933
	Total Elapsed Time	1.5	4	1.61	4.17

Table 7: The summary statistics for low and high expenditure class

	Low Expenditure Class		High Expenditure Class	
	Count	**Percent**	**Count**	**Percent**
Gender				
Male	88017	21.2	1514599	45.4
Female	327656	78.8	1824363	54.6
Education level				
Not completed any school	245895	59.2	587806	17.6
Primary education	145297	35.0	1319401	39.5
Secondary education	17113	4.1	446348	13.4
Bachelor's degree and higher	7369	1.8	985407	29.5
Marital status				
Single	402105	96.7	3131795	93.8
Married	13569	3.3	207167	6.2
Employment status				
Not working	375105	90.2	1776392	53.2
Working	40569	9.8	1562570	46.8
Whether to have private health insurance or not				
No	415673	100.0	2919917	87.4
Yes	0	0.0	419045	12.6
Ownership status of any real-estate				
No	85189	20.5	1234865	37.0
Yes	330484	79.5	2104097	63.0
Smoking habit				
No	378055	90.9	2198149	65.8
Yes	37619	9.1	1140813	34.2
Habit of buying daily newspaper				
No	408540	98.3	3114307	93.3
Yes	7133	1.7	224654	6.7
Having cable TV				
No	415673	100.0	2898360	86.8
Yes	0	0.0	440601	13.2
Habit of playing games of chance				
No	415673	100.0	3145980	94.2
Yes	0	0.0	192982	5.8
Habit of eating out				
No	406457	97.8	1675142	50.2
Yes	9217	2.2	1663819	49.8
Habit of shopping via internet				
No	183638	44.2	661706	19.8
Yes	232035	55.8	2677256	80.2
Having credit card				
No	403203	97.0	1879707	56.3
Yes	12471	3.0	1459255	43.7
Savings habit				
No	112361	27.0	1390559	41.6
Yes	303313	73.0	1948403	58.4

Table 7: The summary statistics for low and high expenditure class (continued)

	Low Expenditure Class	**High Expenditure Class**
Age		
Mean ± SD	67.54 ± 12.59	54.53 ± 18.58
Min-Max	31.00 - 93.00	22.0 - 97.00
Median (IQR)	69.00 (15.00)	57.00 (33.00)
Annual Income		
Mean ± SD	11453.82 ± 9462.30	36607.19 ± 32236.02
Min-Max	1800.00 - 73800.00	3555.00 - 393316.00
Median (IQR)	9600.00 (9600.00)	27428.00 (28142.00)

SD:Standard Deviation, Min:Minimum, Max:Maximum, IQR: Interquartile Range

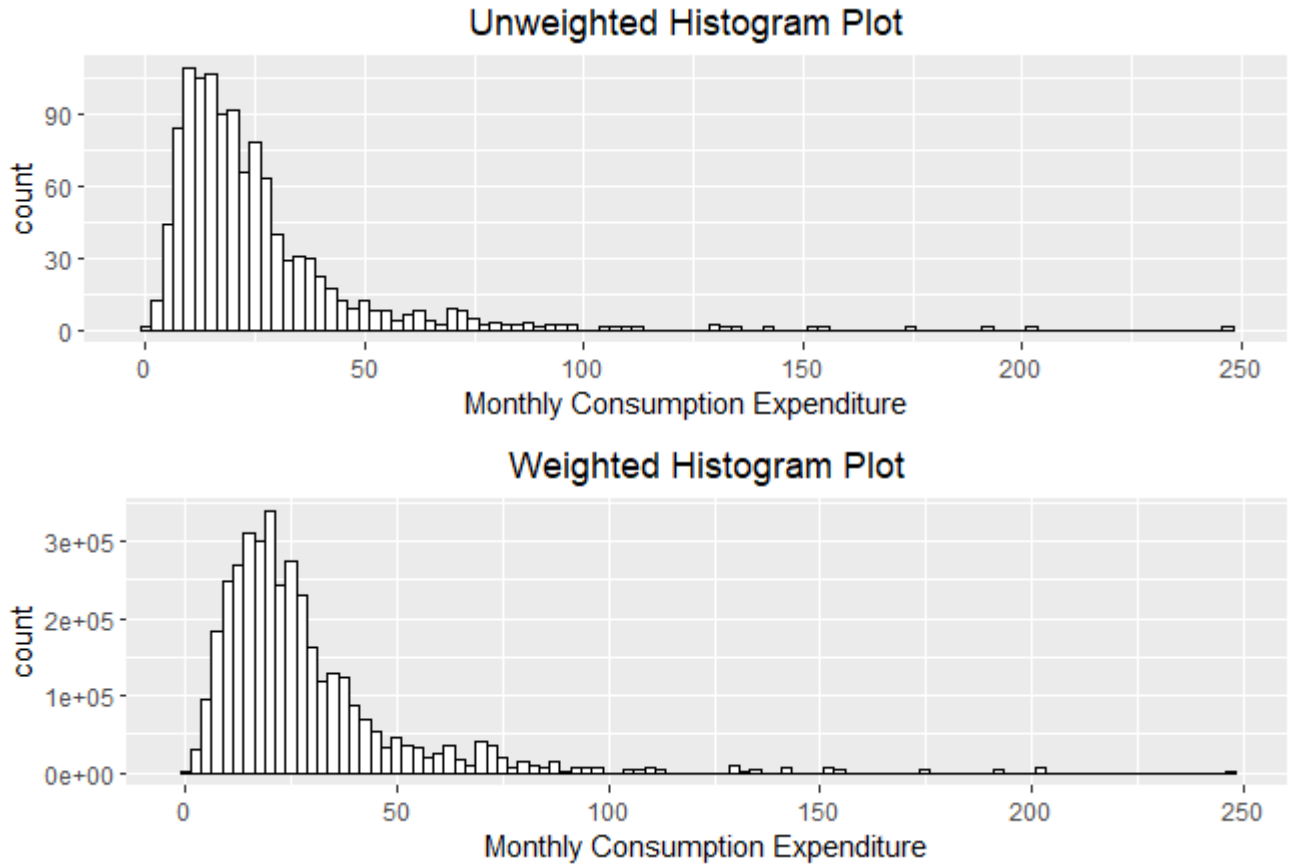

Fig. 1: The weighted / unweighted histograms for the monthly consumption expenditure