A new composite lognormal-Pareto type II regression model to analyze household budget data via particle swarm optimization

Hande Konşuk Ünlü

Accepted: 30 November 2021 / Published online: 30 January 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

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 distributions at specified threshold(s), are alternative models. When data contain covariate information, composite regression models can be used. In the existing literature, there is not much work done on this topic. The only study is Gan and Valdez (N Am Actuar J 22(4):554–573, 2018) paper. In this study, a novel Lognormal-Pareto type II composite regression model is proposed. Particle swarm optimization (PSO) is performed to obtain model parameters of the proposed model. The proposed model is applied to model monthly consumption expenditure and affecting factors. The data is obtained from the National Household Budget Survey, which is conducted annually by the Turkish Statistical Institute (TurkStat). Since the sampling design of the Household Budget Survey is stratified two-stage cluster sampling, the parameters are estimated under weighted data by updating the proposed model and PSO. Additionally, the proposed regression model performance is compared with Lognormal, Lomax, Gamma, and Gamma–Pareto type II regression models. The results demonstrate that the proposed model provides an improved fit to data.

Keywords Composite regression model · Particle swarm optimization · Complex survey design · Household budget survey · Lognormal regression model · Lomax regression model

1 Introduction

Researchers are often working on finding models that fit well to data where a single component distribution does not provide a good fit due to certain properties of the data, such as heterogeneity, multimodality, heavy-tailedness, etc. Finite mixture models have been widely used by researchers for modeling data involving heterogeneous sub-populations. Modeling such data with a single distribution often yields an inadequate fit to whole data. Whether a uni-modal, bi-modal, or even multi-modal distribution is used depends on the degree of heterogeneity among all sub-populations (Shen 2011). Especially after the publication of the monograph by McLachlan and Basford (1988), researchers working in various fields have noticed the potential usefulness of mixture models for inference and clustering (McLachlan and Peel 2000b). 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 (van Dijk 2009). 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), the mixture of t-distributions by Peel and McLachlan (2000), the mixture of skew normal and skew t-distributions by Lee and McLachlan (2013), and the mixture of generalized hyperbolic distributions by Browne and McNicholas (2015). For detailed information, interested readers can also refer to the books written for example by Everitt and Hand (1981), McLachlan and Peel (2000a) and Frühwirth-Schnatter (2006).

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 is called as finite mixture regression model Dietz and Böhning (1995). 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), a two-component finite mixture normal regression model by Yau et al. (2003), and a two-component finite mixture generalized linear mixed effect regression model by Hall and Wang (2005).

Composite models constructed by joining the two (or more) weighted distributions at a given threshold value(s) can be interpreted as a finite mixture model with mixing weights $c$ and $(1 - c)$ (Dominicy and Sinner 2017). 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), composite Lognormal-Pareto models by Scollnik (2007), composite exponential-Pareto models by Teodorescu and Vernic (2009), composite Pareto models by Teodorescu and Vernic (2013), and composite Stoppa models by Calderín-Ojeda and Kwok (2016).

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) proposed the use of two composite regression models consisting of three components to model Singapore auto claims data which capture both fat-tail behavior and zero-inflation of data and the policyholders heterogeneity.

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

The rest of the paper is organized as follows. Section 2 presents 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 Sect. 3. Section 4 presents a simulation study to assess the performance of PSO. The application of the proposed model to household budget survey data is given in Sect. 5. A brief discussion of the results of this study is given in Sect. 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 are skewed with a heavier right tail than classic models, including the gamma, the lognormal, the Weibull, or the Pareto (Bølviken 2014).

Assuming that the random variable $Y$ 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 its probability density function can be expressed as follows:

$$f(y) = \begin{cases} rf_1(y), & y \leq \theta \\ rf_2(y), & y > \theta, \end{cases}$$

where $\theta$ is threshold and $r = \frac{1}{\int_0^\theta f_1(y) \, dy + \int_\theta^\infty f_2(y) \, dy}$ 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 the threshold (Cooray and Ananda 2005).

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

$$f(y) = \begin{cases} cf_1^*(y), & y \leq \theta \\ (1 - c)f_2^*(y), & y > \theta, \end{cases}$$

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 by, respectively,

$$f_1^*(y) = \frac{f_1(y)}{\int_0^\theta f_1(y) \, dy} = \frac{f_1(y)}{F_1(\theta)}, \quad y \leq \theta$$

$$f_2^*(y) = \frac{f_2(y)}{\int_\theta^\infty f_2(y) \, dy} = \frac{f_2(y)}{1 - F_2(\theta)}, \quad y > \theta.$$
\[ \sigma > 0. \] The second component in Eq. 3 is Pareto type II distribution with scale parameter \( \alpha > 0 \), shape parameter \( \lambda > 0 \) and location parameter \( \theta > 0 \). The location parameter of Lognormal distribution and the scale parameter of the Pareto type II distribution are modeled using a regression. Then, the composite Lognormal-Pareto type II regression model is:

\[
f(y_i) = \left\{ \begin{array}{ll}
c \Phi \left( \frac{\ln(\theta) - \mu_i(\mathbf{\beta}_1, \mathbf{X}_i)}{\sigma} \right)^{-1} & \frac{1}{\sigma y_i \sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{\ln(y_i) - \mu_i(\mathbf{\beta}_1, \mathbf{X}_i)}{\sigma} \right)^2 \right), y_i \in (0, \theta] \\
(1-c) \frac{\lambda}{\alpha_i(\mathbf{\beta}_2, \mathbf{X}_i)} \left[ 1 + \frac{y_i - \theta}{\alpha_i(\mathbf{\beta}_2, \mathbf{X}_i)} \right]^{-(\lambda+1)} & y_i \in (\theta, \infty).
\end{array} \right.
\]

Here, \( c \) is mixing weight, \( y = (y_i, i = 1, 2, \ldots, n) \) are the values of the response variable. Let \( \mathbf{X} = (\mathbf{X}_{ij}, i = 1, 2, \ldots, n, j = 0, 1, \ldots, p) \) denotes the \( n \times (p + 1) \) matrix of the values of the \( p \) explanatory variables, with the first column assumed to be a 1 to accommodate the estimation of an intercept and \( \mathbf{X}_i = (1, \mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{ij}, \ldots, \mathbf{X}_{ip}) \) is vector of the \( p \) explanatory variable values for the \( i \)th observation. \( \mathbf{\beta} = (\mathbf{\beta}_1, \mathbf{\beta}_2) \) denotes the component specific parameter vectors. The explanatory variables are incorporated through the link function \( g(\cdot) \). In this setting, the location parameter \( \mu \) and the scale parameter \( \alpha \) are related to the linear predictors through identity and exponential link functions as \( \mu_i(\mathbf{\beta}_1, \mathbf{X}_i) = \mathbf{\beta}_1^T \mathbf{X}_i [1] \{ y_i \leq \theta \} \) and \( \alpha_i(\mathbf{\beta}_2, \mathbf{X}_i) = \exp(\mathbf{\beta}_2^T \mathbf{X}_i [1] \{ y_i > \theta \}) \) respectively.

Let \( E(Y|Y \leq \theta) \) and \( E(Y|Y > \theta) \) denote, respectively, the contribution of the Lognormal distribution and Pareto type II distribution to the expected value of the response variable. The expectation of the response variable with probability density function given in Eq. 4 can be obtained as \( E(Y) = E(Y|Y \leq \theta) + E(Y|Y > \theta) \).

\[
E(Y|Y \leq \theta) = c \exp(\mu^*) \frac{\Phi \left( \frac{\ln(\theta) - \mu^* - \sigma^2}{\sigma} \right)}{\Phi \left( \frac{\ln(\theta) - \mu^*}{\sigma} \right)} \exp \left( \frac{\sigma^2}{2} \right), \quad (5a)
\]

\[
E(Y|Y > \theta) = (1-c) \left( \theta + \frac{\alpha^*}{\lambda - 1} \right), \quad \lambda > 1, \quad (5b)
\]

where \( \mu^* \) represents the mean of \( \mu_i(\mathbf{\beta}_1, \mathbf{X}_i) \) and \( \alpha^* \) represents the mean of \( \alpha_i(\mathbf{\beta}_2, \mathbf{X}_i) \).

As seen from the right-hand side of Eq. 5a, the contribution of the Lognormal distribution to the expected value is approximately proportional to \( \exp(\mu^*) \). The term \( \Phi \left( \frac{\ln(\theta) - \mu^* - \sigma^2}{\sigma} \right) / \Phi \left( \frac{\ln(\theta) - \mu^*}{\sigma} \right) \) is the ratio of two probabilities and the value of this ratio is approximately equal to 1. This expression does not affect the proportionality. As seen from the right-hand side of Eq. 5b, the effect of Pareto type II distribution has an additive contribution \( \alpha^* \) on the expected value. Consequently, the effect of explanatory variables on the response variable \( Y \) could be interpreted proportionally and additively for the Lognormal and Pareto type II regression parts, respectively.

Since the parameter estimates of the proposed model given in Eq. 4 will be obtained based on the maximum likelihood method, the log-likelihood function given the observed data is written in Eq. 6.

\[
\log L = \sum_{i=1}^{n} I_{(0,\theta)}(y_i) \log(c) - \frac{1}{2} \log(2\pi) - \log(\sigma) - \log(y_i) - \frac{1}{2} \left( \frac{\log(y_i) - \mu_i(\mathbf{\beta}_1, \mathbf{X}_i)}{\sigma} \right)^2 \\
- \log \left( \Phi \left( \frac{\log(\theta) - \mu_i(\mathbf{\beta}_1, \mathbf{X}_i)}{\sigma} \right) \right) + I_{(\theta,\infty)}(y_i) \log(1-c) + \log(\lambda) - \log(\alpha_i(\mathbf{\beta}_2, \mathbf{X}_i)) - (\lambda + 1) \log \left( 1 + \frac{y_i - \theta}{\alpha_i(\mathbf{\beta}_2, \mathbf{X}_i)} \right) \right), \quad (6)
\]

It should be noted that it is very difficult to solve score functions obtained by derivation of the log-likelihood of the proposed model to find the parameter estimates. In this case, heuristic-based algorithms such as genetic algorithm, PSO algorithm, etc. are widely used in parameter estimation for complex problems. In this study, PSO is performed to obtain the model parameters of the proposed model. The log-likelihood function given in Eq. 6 is used as a fitness function in the PSO algorithm.

### 3 Particle swarm optimization

Heuristic methods such as Genetic Algorithm, Differential Evolution Algorithm and PSO algorithm inspired by the events in nature are powerful optimization methods used to solve a complex system. PSO, developed by Kennedy and Eberhart (1995), was inspired by the social behavior of bird flocking or a school of fish. It has been observed that the actions (e.g., to search food and have safe location) of the animals’ moving in herd, often randomly, 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 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, an individual is called a “particle” (e.g., each bird in swarm) in the swarm. The change of particle position in the search space (i.e., parameter 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 (Özsağlam and Çunkaş 2008).

The PSO starts with a set of possible solutions called a “swarm”, which are randomly generated from the parameter space at the beginning of the algorithm. In the swarm, each solution is defined as a “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 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, 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 the whole swarm (Saravanan 2006; Shi and Eberhart 1999).

Let \( X_{kj}(t) = (x_{k1}, x_{k2}, \ldots, x_{kd}) \) denote the location and \( v_{kj} = (v_{k1}, v_{k2}, \ldots, v_{kd}) \) represent the velocity of the \( k \)th particle (\( k = 1, 2, \ldots, K \)) and \( j \)th position (\( j = 1, 2, \ldots, d \)) at the \( t \)th iteration in the \( d \)-dimensional search space. For the \( (t+1) \)th iteration, the velocity and location of the \( k \)th particle and \( j \)th position are updated by the following formulas:

\[
v_{kj}(t+1) = v_{kj}(t) + c_1 r_{1,kj}(t) [p_{best,kj}(t) - x_{kj}(t)] + c_2 r_{2,kj}[g_{best,j}(t) - x_{kj}(t)] \quad (7a)
\]

\[
X_{kj}(t+1) = X_{kj}(t) + v_{kj}(t+1) \quad (7b)
\]

Here, \( c_1 \) and \( c_2 \) are accelerating factors used to scale cognitive and social components respectively, and \( r_{1,kj} \) and \( r_{2,kj} \) are random numbers generated from the standard uniform distribution, \( U[0, 1] \). \( p_{best,kj} \) (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 (Acitas et al. 2019; Amoshahy et al. 2016; Engelbrecht 2013; Parsopoulos et al. 2002).

Velocity regulation is very important in the PSO algorithm. There are many methods to limit the velocity in the literature. One of them is inertia weight \( (w) \), which was introduced to limit velocity of the particles by Shi and Eberhart (1998). The modified velocity formula by adding inertia weight to the Eq. 7a is given below:

\[
v_{kj}(t+1) = w v_{kj}(t) + c_1 r_{1,kj}(t) [p_{best,kj}(t) - x_{kj}(t)] + c_2 r_{2,kj}[g_{best,j}(t) - x_{kj}(t)] . \quad (8)
\]

When the inertia weight, \( w > 1 \), the velocities of particles increase with time to the maximum speed and the swarm diverges. On the other hand, small inertia values facilitate the local search but weaken the global search capability of PSO (Karaboğa 2017; Shi and Eberhart 1999). Inertia weight was taken as constant in the early studies published in the literature. Later in the paper by Eberhart and Shi (2000), a linear decreasing value for the inertia weight \( (w) \) was proposed.

The algorithm consists of the following steps:

1. The initial swarm is created with randomly generated starting locations and velocities for prespecified swarm size \( (K) \).
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 and the location of the particles are updated at each iteration by using Eq. 8 and Eq. 7b, respectively.
7. The steps 1–6 are repeated until stopping criteria is satisfied.

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 make 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 is decided to use PSO 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 a two-stage process. In the first stage, the simulation study is designed in order to determine optimal hyper parameters of PSO algorithm. In the second stage, another simulation study is designed to obtain model parameters given in Eq. 4 using PSO with its hyper parameters which were chosen in stage 1.

In both stages, the number of explanatory variables is taken as 4, component specific parameter vectors are 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 is specified as 10. An explanatory variable matrix is generated using a normal distribution with location parameter \( \theta \) and shape parameter \( \sigma \) taken as 4, component specific parameter vectors are 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 is specified as 10. An explanatory variable matrix is generated using a normal distribution with mean 2 and standard deviation 0.5. The values of the dependent variables being lower and higher than the threshold \( \theta \) are generated, respectively, from a Lognormal distribution with location parameter \( \mu_i, (\beta_1, X_i) = \beta_1^T X_i \) and shape parameter \( \sigma = 0.1 \) and from a Pareto type II distribution with scale parameter \( \alpha_i, (\beta_2, X_i) = \exp (\beta_2^T X_i), \) shape parameter \( \lambda = 2 \) and location parameter \( \theta = 10. \) Here, the mixing weight is taken as 0.50.

In the first stage of the simulation study, a grid search is implemented in order to optimize inertia weight parameters, swarm size, maximum number of iterations, and the accelerating coefficients \( c_1 \) and \( c_2 \) of the PSO algorithm. The inertia weight values are taken fixed as \( (0.1, 0.3, 0.5, 0.7, 0.9, 1.1) \) in the first stage. In addition, inertia weight values are taken as dynamically decreasing linearly from 1.4 to 0.4. The other hyper-parameters of the PSO algorithm used in the first stage are given below:

- Sample size is taken as 300, 500 and 1000.
- Swarm size is taken as 20 and 40.
- The accelerating coefficients are taken equal to:
  - \( c_1 = c_2 = 1.49 \)
  - \( c_1 = c_2 = 2.00 \)
- The maximum number of iterations are taken as 100 and 250.

Each scenario is repeated 500 times. The simulation is stopped when the PSO algorithm reaches the maximum iteration number. For each scenario, the Akaike Information Criteria (AIC), total absolute bias, and total elapsed time are calculated. The scenarios for different inertia weight values are compared according to their AIC values.

The PSO algorithm is given in greater detail in Sect. 3 and all data generation processes are carried out using R ver.4.0.0 (R Core Team 2020) with self-written code. The results of the first stage of the simulation study are given in Tables 1 and 2.

As seen in Tables 1 and 2, total elapsed times for different inertia weight values are found to be similar while the other hyper-parameters, sample size, swarm size, maximum number of iterations, \( c_1 \) and \( c_2 \) are not changed.

When the sample size is 300 or 500, the swarm size is 20 or 40, the maximum number of iterations is 100 or 250, the accelerating coefficients are \( c_1 = c_2 = 1.49, \) and the inertia weight value with the lowest AIC value is found to be 0.5. This inertia weight value also has the lowest total absolute bias value for all simulation settings.

In Table 1, the total absolute bias values of each inertia weight value are generally decreasing when the maximum number of iterations increase from 100 to 250 under the corresponding sample and swarm size. The total absolute bias values of each inertia weight value are decreasing when the swarm size increases from 20 to 40 under the corresponding sample and maximum number of iterations.

As seen in Table 2, when the AIC values are compared, it is concluded that the inertia weight value 0.1 has the lowest AIC value in most of the simulation settings. The inertia weight value with the minimum AIC value has been found to be 0.3 when the sample size is 300 or 500, swarm size is 40, and maximum number of iterations is 250.

The total absolute bias values of each inertia weight value generally increase as the maximum number of iterations increases from 100 to 250, where the corresponding swarm size is 20. However, the total absolute bias values of each inertia weight value generally decrease as the maximum number of iterations increases from 100 to 250, where the corresponding swarm size is 40.

According to the results of Tables 1 and 2, there is a slight decrease in AIC values when the swarm size is increased from 20 to 40 or the maximum number of iterations increased from 100 to 250. Therefore, the values of swarm size and maximum number of iterations are selected as 20 and 40, respectively for the second stage of simulation study.

Comparing the results of Tables 1 and 2 under the same sample size, the AIC values corresponding to \( c_1 = c_2 = 1.49 \) are better when the swarm size is 40 and the maximum number of iterations is 250.

According to the above results, the hyper-parameters which will be used in order to obtain parameter estimates of the model in Eq. 4 at the second stage of the simulation study are determined as:

- Swarm size is taken as 40.
- The accelerating coefficients are taken equal to \( c_1 = c_2 = 1.49 \).
- The maximum number of iterations is taken as 250.
- The inertia weight value is taken as 0.5.

In addition to the parameters given above, the sample size is taken as \( N = 300, 500 \) and 1000 and the mixing weight is
Comparison of performance of PSO algorithm for \( N = 300, 500 \) and \( c_1 = c_2 = 1.49 \) with respect to different hyper parameters

| Inertia weight | Maximum Iteration = 100 | Maximum Iteration = 250 |
|----------------|-------------------------|-------------------------|
|                | AIC         | TET        | TAB        | AIC         | TET        | TAB        |
| Sample size = 300, \( K = 20 \) |             |            |            |             |            |            |
| 0.1            | 1619.0017  | 0.0906    | 6.1768    | 1618.2495  | 0.4565    | 6.1505    |
| 0.3            | 1611.5130  | 0.0913    | 6.0335    | 1610.6098  | 0.4589    | 5.9479    |
| 0.5            | 1601.8139  | 0.0922    | 5.8791    | 1599.9624  | 0.4603    | 5.7745    |
| 0.7            | 1638.1819  | 0.0906    | 7.0524    | 1638.5923  | 0.4582    | 6.9913    |
| 0.9            | 1675.3860  | 0.0920    | 8.3311    | 1673.8192  | 0.4594    | 8.3078    |
| 1.1            | 1693.2067  | 0.0918    | 8.7431    | 1693.5688  | 0.4627    | 8.7057    |
| Dynamic Inertia\(^1\) | 1653.2072  | 0.0894    | 7.9659    | 1660.0042  | 0.4562    | 7.5709    |
| Sample size = 300, \( K = 40 \) |             |            |            |             |            |            |
| 0.1            | 1591.2223  | 0.3131    | 5.4391    | 1590.6953  | 1.7075    | 5.4234    |
| 0.3            | 1577.1771  | 0.3134    | 4.9850    | 1577.9296  | 1.7149    | 5.0619    |
| 0.5            | 1572.8125  | 0.3133    | 4.8917    | 1568.0444  | 1.7126    | 4.7611    |
| 0.7            | 1615.9128  | 0.3102    | 6.1003    | 1613.5411  | 1.7063    | 5.9551    |
| 0.9            | 1656.0878  | 0.3114    | 7.5244    | 1654.5885  | 1.7079    | 7.5003    |
| 1.1            | 1674.2394  | 0.3118    | 8.3056    | 1672.7948  | 1.7134    | 8.2932    |
| Dynamic Inertia\(^1\) | 1610.6678  | 0.3079    | 6.6320    | 1622.7182  | 1.7030    | 6.9075    |
| Sample size = 500, \( K = 20 \) |             |            |            |             |            |            |
| 0.1            | 2697.1204  | 0.0932    | 6.1212    | 2696.0563  | 0.4830    | 6.0965    |
| 0.3            | 2674.7074  | 0.0936    | 5.5641    | 2678.4426  | 0.4821    | 5.6142    |
| 0.5            | 2663.5095  | 0.0949    | 5.4350    | 2658.7930  | 0.4833    | 5.1891    |
| 0.7            | 2724.9524  | 0.0932    | 6.7560    | 2720.1830  | 0.4813    | 6.4474    |
| 0.9            | 2788.3858  | 0.0948    | 8.7647    | 2789.8381  | 0.4830    | 8.7133    |
| 1.1            | 2822.1536  | 0.0948    | 9.7783    | 2821.2222  | 0.4865    | 9.8330    |
| Dynamic Inertia\(^1\) | 2751.6598  | 0.0916    | 7.6875    | 2769.7596  | 0.4802    | 8.3434    |
| Sample size = 500, \( K = 40 \) |             |            |            |             |            |            |
| 0.1            | 2644.7134  | 0.3222    | 4.9562    | 2643.4459  | 1.7346    | 4.9062    |
| 0.3            | 2624.8537  | 0.3223    | 4.4652    | 2621.9506  | 1.7444    | 4.4186    |
| 0.5            | 2609.0750  | 0.3223    | 4.1561    | 2599.8100  | 1.7436    | 3.7586    |
| 0.7            | 2686.5341  | 0.3192    | 5.5574    | 2682.1006  | 1.7368    | 5.4026    |
| 0.9            | 2757.2209  | 0.3205    | 7.2688    | 2758.1993  | 1.7369    | 7.4547    |
| 1.1            | 2790.6697  | 0.3213    | 8.7514    | 2788.7616  | 1.7447    | 8.5696    |
| Dynamic Inertia\(^1\) | 2676.2041  | 0.3167    | 6.2033    | 2703.8507  | 1.7342    | 6.9184    |

\(^1\) Linearly decreasing values from 1.4 to 0.4. 
\( TET \) total elapsed time (minute), \( TAB \) total absolute bias

taken as \( c = 0.3, 0.5 \) and 0.7. Each scenario was replicated 1500 times. The simulation was stopped when the PSO algorithm reached to the maximum iteration number. For each scenario, Bayesian Information Criteria (BIC), mean of total absolute bias and total elapsed time are calculated.

The results of the proposed model obtained for different sample sizes and mixing weights are compared. Using the simulated data, the parameters of the Lognormal regression, Lomax regression, and Gamma regression models are also estimated. The BIC values for these models are compared with the proposed model, Lognormal-Pareto type II composite regression model. The results are given in Tables 3, 4 and 5.

The parameter estimates and their MSE values for different sample sizes of 300, 500 and 1000, under different mixing weights, \( c = 0.30, 0.50 \) and 0.70, are given respectively in Tables 3, 4 and 5. The results indicate that the total absolute bias values are decreasing as the sample size is increasing. Under different mixing weight parameter value, MSEs of the parameters are generally decreasing as the sample size is increasing. The only exception is the increase observed in the estimation of MSEs of the threshold parameter when the sample size changed from 300 to 500 and the mixing ratio
was 0.70. Besides, when the mixing ratio is 0.70, an increase is observed in the MSE values of the parameter estimates of the Pareto regression model, which is the second part of the composite regression model. This may be due to the insufficient sample size above the threshold parameter, \( \theta \), obtained depending on the mixing weight.

The ‘VGAM’ package in the R ver.4.0.0 (R Core Team 2020) software is used to estimate the parameters of the Lognormal, Lomax, and Gamma regression models using simulated data. The probability density function with their mean structures of these models are given below:

Lognormal Regression Model:

\[
f(y_i) = \frac{1}{\sqrt{2\pi\sigma y_i}} \exp\left(-\frac{1}{2} \left( \ln(y_i) - \mu, \frac{y_i}{\sigma} \right)^2\right), \quad y_i > 0
\]

\[
\mu_i(y, X_i) = y^T X_i = y_0 + y_1 x_{i1} + y_2 x_{i2} + y_3 x_{i3}
\]

where \( \sigma > 0 \) is shape parameter, \( \mu > 0 \) is scale parameter, \( X = (X_{ij}, i = 1, 2, \ldots, n, j = 0, 1, 2, 3) \) denotes the matrix of explanatory variables, with the first column assumed to be a 1 to accommodate the estimation of an intercept and \( y^T = (y_0, y_1, y_2, y_3) \) is parameter vector.
Table 3 The parameter estimates of simulation setting for $N = 300, 500$ and $1000$ and mixing weight $= 0.30$

| Parameters | Real values | Sample size = 300 | Sample size = 500 | Sample size = 1000 |
|------------|-------------|--------------------|--------------------|--------------------|
| $\theta$   | 10          | 11.6380            | 9.4004             | 11.1891            |
| $\sigma$   | 0.1         | 0.1684             | 0.0091             | 0.1586             |
| $\beta_{1,0}$ | 0.6931     | 0.9343             | 0.2146             | 0.8844             |
| $\beta_{1,1}$ | -0.2877    | -0.2589            | 0.0106             | -0.2704            |
| $\beta_{1,2}$ | 0.4055      | 0.3429             | 0.0171             | 0.3553             |
| $\beta_{1,3}$ | 0.9163      | 0.7997             | 0.0709             | 0.8336             |
| $\lambda$  | 2           | 1.7344             | 1.2906             | 1.6982             |
| $\beta_{2,0}$ | 0.4055      | 0.4285             | 2.3366             | 0.4340             |
| $\beta_{2,1}$ | 0.8329      | 0.6911             | 0.2032             | 0.6688             |
| $\beta_{2,2}$ | 0.5596      | 0.5107             | 0.1823             | 0.5128             |
| $\beta_{2,3}$ | -0.5108     | -0.3935            | 0.1865             | -0.3917            |
| Total elapsed time | 1.7080 | 1.7173             | 2.3770             | 2.3770             |
| Total absolute bias | 4.5144 | 3.5642             | 3.0290             | 3.0290             |

Table 4 The parameter estimates of simulation setting for $N = 300, 500$ and $1000$ and mixing weight $= 0.50$

| Parameters | Real values | Sample size = 300 | Sample size = 500 | Sample size = 1000 |
|------------|-------------|--------------------|--------------------|--------------------|
| $\theta$   | 10          | 11.5386            | 6.0786             | 11.2233            |
| $\sigma$   | 0.1         | 0.1402             | 0.0039             | 0.1333             |
| $\beta_{1,0}$ | 0.6931     | 0.9046             | 0.1185             | 0.8712             |
| $\beta_{1,1}$ | -0.2877    | -0.2481            | 0.0051             | -0.2547            |
| $\beta_{1,2}$ | 0.4055      | 0.3403             | 0.0106             | 0.3501             |
| $\beta_{1,3}$ | 0.9163      | 0.7812             | 0.0450             | 0.8031             |
| $\lambda$  | 2           | 1.7394             | 2.8902             | 1.5761             |
| $\beta_{2,0}$ | 0.4055      | 0.1677             | 2.8716             | 0.1130             |
| $\beta_{2,1}$ | 0.8329      | 0.6556             | 0.2802             | 0.6650             |
| $\beta_{2,2}$ | 0.5596      | 0.5621             | 0.2074             | 0.5470             |
| $\beta_{2,3}$ | -0.5108     | -0.2945            | 0.2899             | -0.3088            |
| Total elapsed time | 2.0971 | 2.5578             | 2.6200             | 2.6200             |
| Total absolute bias | 4.6815 | 3.9422             | 3.1977             | 3.1977             |

Lomax Regression Model:

$$f(y_i) = \left(1 + \frac{y_i}{\alpha_i(\gamma, X_i)}\right)^{-\lambda}, \quad y_i > 0$$

$$\alpha_i(\gamma, X_i) = \exp \left(\gamma^T X_i\right) = \exp(\gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i3})$$

where $\lambda > 0$ is shape parameter, $\alpha > 0$ is scale parameter, $X = (X_{ij}, i = 1, 2, \ldots, n, j = 0, 1, 2, 3)$ denotes the matrix of explanatory variables, with the first column assumed to be a 1 to accommodate the estimation of an intercept and $\gamma^T = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ is parameter vector.

Gamma Regression Model:

$$f(y_i) = \exp\left(-\frac{y_i}{\lambda_i(\gamma, X_i)}\right)^{-1}$$

$$\lambda_i(\gamma, X_i) = \exp \left(\gamma^T X_i\right) = \exp(\gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i3})$$

where $\nu > 0$ is shape parameter, $\lambda > 0$ is scale parameter, $X = (X_{ij}, i = 0, 1, 2, \ldots, n, j = 0, 1, 2, 3)$ denotes the matrix of explanatory variables, with the first column assumed to be a 1 to accommodate the estimation of an intercept and $\gamma^T = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ is parameter vector.
As seen in Table 6, the Lognormal regression model has the lower BIC value than Lomax and Gamma regression models under the same sample size and same mixing weight. Comparing the results of Lognormal, Lomax, and Gamma regression models with those of the proposed Lognormal-Pareto type II regression model, it is seen that the difference between the BIC values between Lognormal regression model and the proposed model, \( \text{BIC}_\Delta = \text{BIC}_{\text{LogN}} - \text{BIC}_{\text{Prop}} \), the difference between Lomax regression model and the proposed model, \( \text{BIC}_\Delta = \text{BIC}_{\text{Lomax}} - \text{BIC}_{\text{Prop}} \), and the difference between Gamma regression model and the proposed model, \( \text{BIC}_\Delta = \text{BIC}_{\text{Gamma}} - \text{BIC}_{\text{Prop}} \) for all sample sizes are greater than 10 (\( \text{BIC}_\Delta >> 10 \)). Hence, it is concluded that the proposed model provides a better fit to the data than Lognormal, Lomax, and Gamma regression models.

### 5 Household budget data

The proposed model is applied to analyze monthly consumption expenditure and affecting factors. The data is obtained from National Household Budget Survey in 2018, annually conducted by the 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 1st and December 31st, 2018 on 1296 households is used. The sampling design of the Household Budget Survey is a stratified two-stage cluster sampling method. Therefore, parameter estimates are obtained by considering sampling weights. The microdata of 2018 Household Budget Survey consists of three data sets.
which are Household, Individual and Consumption Expenditure. These three data sets are linked to each other with one-to-one matching on the "Household ID", which is common in all data sets in order to select variables to be used in the analysis. After a mapping process, the data are 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 are 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
- Having 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 lottery, 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
- Using credit card or not: no (reference), yes
- Savings habit: no (reference), yes

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

As mentioned before, the parameters are estimated under weighted data by taking the sampling method into consideration. Besides, parameter estimates under unweighted data are also obtained and compared with the weighted results. Similar to the simulation study, parameter estimates are also obtained under weighted and unweighted data using Lognormal, Lomax and Gamma regression models. In addition, a Gamma-Pareto type II composite regression model is also applied and compared with the proposed model.

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

Summary statistics calculated under weighted and unweighted data for the variables to be used in the regression model are given in Table 7.

When the results in Table 7 are examined, it can be seen that despite some differences, the summary statistics obtained for weighted and unweighted data are generally similar. As the data is obtained from the stratified two stage cluster sampling, summary statistics obtained for weighted data are interpreted. The average age of the participants in the study is 55.97 with standard deviation 18.47. 42.7% of them are male, 39.0% of them have primary education and 94.1% of them are single. More than half of the participants are not working (57.3%), and about two-thirds of the participants have their own property. 6.2% of them have the habit of buying daily newspapers. Only 10% of them have private insurance and one third of them have a smoking habit. In addition, 11.7% have a cable TV subscription and 5.1% is playing games of chance. The average monthly consumption expenditure of the participants is 2832.84 with standard deviation 2334.91 and varies between 109 and 24798.19. The average annual income of individuals is $33822.48 Turkish Lira (TL) with standard deviation $31546.51 and varies between $1800 and $393316. 60% of the participants have a savings habit, 39% of them have a credit card and 77.5% of them have the habit of online shopping.

The pairwise correlations among the explanatory variables are investigated. According to the results of the correlation analysis, it is observed that no pair of explanatory variables is highly correlated. Consequently, it is accepted that there is no problem of multi-collinearity in the Lognormal-Pareto type II composite regression model.

Before obtaining the parameter estimates, hyper-parameter optimization is performed for the parameters of the PSO algorithm as in the first step of the simulation study. According to the first-stage of the simulation results, it is seen that more consistent estimations are obtained when the swarm size is 40 and the maximum number of iterations is 250. Hence, the swarm size and maximum number of iterations are taken as 40 and 250, respectively. A grid search is implemented for the remaining hyper-parameters:

- The inertia weight values are taken fixed as (0.1, 0.3, 0.5, 0.7, 0.9, 1.1).
- In addition, inertia weight value is taken dynamically decreasing linearly from 1.4 to 0.4.
- The accelerating coefficients are taken equal to:
  - $c_1 = c_2 = 1.49$
  - $c_1 = c_2 = 2.00$

The results of the hyper-parameter optimization are given in Table 8. As seen in Table 8, when the accelerating coefficients are taken as $c_1 = c_2 = 1.49$, the inertia weight value
A new composite lognormal-Pareto type II regression model to analyze… 2401

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

![Unweighted Histogram Plot](image1)

![Weighted Histogram Plot](image2)

Table 7 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    |
| **Having 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     |
Table 7 continued

|                                | Unweighted statistics | Weighted statistics |
|--------------------------------|-----------------------|--------------------|
|                                | Count  | Percent | Count  | Percent |
| **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    | 3561635| 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    |
| **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 8 Results of hyper parameters optimization for real data set

| Inertia Weight | c1 = c2 = 1.49 | c1 = c2 = 2.00 |
|----------------|----------------|----------------|
|                | AIC            | BIC            | TET | AIC            | BIC            | TET |
| 0.1            | 29438372.32    | 29438911.00    | 226.44 | 29430359.50  | 29130898.18  | 158.49 |
| 0.3            | 29284967.92    | 29285506.60    | 183.31 | 29273248.54  | 29273787.22  | 159.29 |
| 0.5            | 29507355.69    | 29507894.37    | 163.63 | 29840290.06  | 29840828.74  | 158.24 |
| 0.7            | 29112599.38    | 29113138.06    | 163.44 | 29261064.99  | 29061603.66  | 158.62 |
| 0.9            | 29234548.93    | 29235087.61    | 160.64 | 29874078.00  | 29874616.68  | 158.47 |
| 1.1            | 30001049.97    | 30001588.65    | 162.62 | 30248944.61  | 30249848.32  | 158.14 |
| Dynamic Inertia1 | 29576172.81  | 29576711.49    | 168.87 | 29461700.78  | 29462239.46  | 159.34 |

1 Linearly decreasing values from 1.4 to 0.4.
TET total elapsed time (minute)

with the minimum AIC value has been found to be 0.7. Similarly, the inertia weight value with the minimum AIC value has also been found to be 0.7, when the accelerating coefficients are taken as c1 = c2 = 2.00. Comparing the results of different accelerating coefficients under the swarm size is 40 and the maximum number of iterations is 250; the AIC values
A new composite lognormal-Pareto type II regression model to analyze...

Table 9 The weighted and unweighted parameter estimates of proposed Lognormal-Pareto type II Composite Regression model for Household Budget Data

| Threshold, $\theta$ | Unweighted estimates | Weighted estimates |
|---------------------|-----------------------|--------------------|
|                      | $\hat{\beta}$ | $\exp(\hat{\beta})$ | $\hat{\beta}$ | $\exp(\hat{\beta})$ |
| Lognormal            |                     |                    |                    |                     |
| $\sigma$             | 0.5078              | 0.4727             |
| Intercept$^1$        | 1.6829              | 5.3813             |
| Age$^1$              | 0.0023              | 1.0023             |
| Savings Habit$^1$ (Yes) | 0.233           | 1.2624             |
| Employment$^1$ (Working) | 0.0002         | 1.0002             |
| Annual income$^1$    | 0.00002             | 1.00002            |
| Ownership$^1$ (Yes)  | 0.1771              | 1.1937             |
| Shopping$^1$ (Yes)   | 0.0721              | 1.0747             |
| Smoking Habit$^1$ (Yes) | -0.0094        | 0.9907             |
| Newspaper$^1$ (Yes)  | 0.228               | 1.2561             |
| Cable TV$^1$ (Yes)   | 0.0818              | 1.0852             |
| Chance Games$^1$ (Yes) | 0.2544        | 1.2897             |
| Eating out$^1$ (Yes) | 0.0618              | 1.0638             |
| Health insurance$^1$ (Yes) | 0.0933       | 1.0978             |
| Credit Card$^1$ (Yes) | 0.1234           | 1.1313             |
| Gender$^1$ (Female)  | 0.1082              | 1.1143             |
| Education$^1$ (Primary) | 0.092           | 1.0964             |
| Education$^1$ (Secondary) | 0.2218        | 1.2484             |
| Education$^1$ (Bachelor) | 0.3527          | 1.4229             |
| Marital$^1$ (Married) | 0.2559           | 1.2916             |
| Pareto Type II       |                     |                    |                    |                     |
| $\lambda$           | 1.3057              | 18.3679            |
| Intercept$^2$        | 4.3564              | 77.9777            |
| Age$^2$              | -0.0511             | 0.9502             |
| Savings Habit$^2$ (Yes) | -0.2222        | 0.8008             |
| Employment$^2$ (Working) | 0.512            | 1.6686             |
| Annual income$^2$    | -0.00001            | 1.0000             |
| Ownership$^2$ (Yes)  | 0.2541              | 1.2893             |
| Shopping$^2$ (Yes)   | -0.0656             | 0.9365             |
| Smoking Habit$^2$ (Yes) | 0.0179           | 1.0181             |
| Newspaper$^2$ (Yes)  | -0.0258             | 0.9745             |
| Cable TV$^2$ (Yes)   | -0.4159             | 0.6598             |
| Chance Games$^2$ (Yes) | -8.5415         | 0.0002             |
| Eating out$^2$ (Yes) | 0.6934              | 2.0005             |
| Health insurance$^2$ (Yes) | 0.4607          | 1.5851             |
| Credit Card$^2$ (Yes) | -0.7073           | 0.493              |
| Gender$^2$ (Female)  | -0.0548             | 0.9467             |
| Education$^2$ (Primary) | 0.17              | 1.1854             |
| Education$^2$ (Secondary) | 0.6207         | 1.8603             |
| Education$^2$ (Bachelor) | 0.8086           | 2.2448             |
| Marital$^2$ (Married) | -0.6003           | 0.5486             |

corresponding to $c_1 = c_2 = 1.49$ are obtained better. Consequently, according to these results, the hyper-parameters, which will be used in order to obtain parameter estimates of the proposed model, are determined as:

- Swarm size is taken as 40.
- The accelerating coefficients are taken equal to $c_1 = c_2 = 1.49$.
- The maximum number of iterations are taken as 250.
• The inertia weight value is taken as 0.7.

After the hyper-parameters of the PSO algorithm are decided, the PSO algorithm is run 10 times by taking different random seeds. The solution set with the lowest AIC value is determined as the best result.

The parameter estimates for the Lognormal-Pareto type II composite regression model, which are obtained for weighted and unweighted data using variables considered as related to the average monthly consumption expenditure, are given in Table 9. It is observed that the histograms of the monthly consumption expenditure for weighted and unweighted data are very similar as it can be seen in Fig. 1. Similarly, the summary statistics obtained under weighted and unweighted data are also found similar. However, when the parameter estimates given in Table 9 are examined, it is observed that the results of weighted and unweighted data are very different from each other. In particular, the estimate of the threshold value ($\theta$) for weighted data is around 30, while the estimate for the unweighted data is around 63. However, since the sampling method is stratified two-stage cluster sampling, parameters should be estimated under weighted data. The reason to obtain parameter estimates for weighted and unweighted data separately is to show that parameter estimates for weighted data are quite different from the parameter estimates for unweighted data, although both distributions are found quite similar in Fig. 1.

Similar to the simulation study, parameter estimates are also obtained by using Lognormal, Lomax, and Gamma regression models. Additionally, Lognormal-Pareto type II composite regression model is also compared to Gamma-Pareto type II composite regression model. Suppose, the first component in Eq. 3 is the Gamma distribution with scale parameter $k > 0$ and shape parameter $\nu > 0$. The second component in Eq. 3 is Pareto type II distribution with scale parameter $\lambda > 0$, shape parameter $\theta > 0$ and location parameter $\lambda > 0$. The probability density function of Gamma-Pareto type II composite regression model is given below:

$$
 f(y_i) = \begin{cases} 
 \frac{1}{k_1 (\beta_1, X_i)^\nu} \left( \frac{\theta}{\nu} \right) y_i^{\nu-1} \exp\left( -\frac{y_i}{k_1 (\beta_1, X_i)} \right) & , y_i \in (0, \theta] \\
 \frac{\lambda}{\alpha_i (\beta_2, X_i)} \left[ \frac{y_i - \theta}{\alpha_i (\beta_2, X_i)} \right]^{-(\lambda+1)} & , y_i \in (\theta, \infty) 
\end{cases}
$$

Here, $c$ is mixing weight, $\gamma\left(\nu, \frac{\theta}{k_1 (\beta_1, X_i)}\right)$ represents the lower incomplete gamma function, $y_i = (y_i, i = 1, 2, \ldots, n)$ are the values of the response variable. Let $X = (X_{ij}, i = 1, 2, \ldots, n, j = 0, 1, \ldots, p)$ denotes the $n \times (p + 1)$ matrix of the values of the $p$ explanatory variables, with the first column assumed to be a 1 to accommodate the estimation of an intercept and $X_i = (1, x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{ip})$ is vector of the $p$ explanatory variable values for the $i$th observation. $\beta = (\beta_1', \beta_2')$ denotes the component specific parameter vectors. The explanatory variables are incorporated through the link function $g(\cdot)$. In this setting, the scale parameters $k$ and $\nu$ are related to the linear predictors through inverse and exponential link functions as $k_i (\beta_1, X_i) = \exp(\beta_1' X_i) [I(y_i \leq \theta)]$ and $\alpha_i (\beta_2, X_i) = \exp(\beta_2' X_i [I(y_i > \theta)])$ respectively. The parameter estimates of Gamma-Pareto type II composite regression model are obtained using PSO algorithm in R version 4.0.0 (R Core Team 2020) with self-written code.

As seen in Table 10, using Lognormal regression model, AIC$_w$ and BIC$_w$ values for weighted data are found as 80190799.46 and 80191062.23, respectively. Using Lomax regression model, AIC$_w$ and BIC$_w$ values for the weighted data are found as 54178799.27 and 54179062.04, respectively. Using Gamma regression model, AIC$_w$ and BIC$_w$ values for the weighted data are found as 48632112.26 and 48632213.19, respectively. According to the results of the Gamma-Pareto type II composite regression model, AIC$_w$ and BIC$_w$ values for the weighted data are found as 38998900.92 and 38999439.60, respectively. There is a significant difference between the BIC values of Lognormal-Pareto type II composite regression model and Lognormal, Lomax, and Gamma-Pareto type II composite regression models. This result confirms that the most compatible model with the data is the proposed Lognormal-Pareto type II composite regression model.

|                | AIC    | BIC    |
|----------------|--------|--------|
| Lognormal-Pareto type II | 29666730.38 | 29667269.06 |
| Gamma-Pareto type II      | 38998900.92  | 38999439.60  |
| Lognormal                 | 80190799.46  | 80191062.23  |
| Lomax                     | 54178799.27  | 54179062.04  |
| Gamma                     | 48632112.26  | 48632213.19  |
Table 11 The summary statistics for low and high expenditure class

|                        | Low expenditure class |                         | High expenditure class |                         |
|------------------------|-----------------------|-------------------------|------------------------|-------------------------|
|                        | Count                 | Percent                 | Count                  | Percent                 |
| **Gender**             |                       |                         |                        |                         |
| Male                   | 1053541               | 39.7                    | 549076                 | 49.9                    |
| Female                 | 1600999               | 60.3                    | 551019                 | 50.1                    |
| **Education level**    |                       |                         |                        |                         |
| Not completed any school | 791995             | 29.8                    | 41705                  | 3.8                     |
| Primary education      | 1147479               | 43.2                    | 317219                 | 28.8                    |
| Secondary education    | 318244                | 12.0                    | 145217                 | 13.2                    |
| Bachelor’s degree and higher | 396822     | 14.9                    | 595954                 | 54.2                    |
| **Marital status**     |                       |                         |                        |                         |
| Single                 | 2515339               | 94.8                    | 1018560                | 92.6                    |
| Married                | 139201                | 5.2                     | 81535                  | 7.4                     |
| **Employment status**  |                       |                         |                        |                         |
| Not working            | 1752464               | 66.0                    | 399033                 | 36.3                    |
| Working                | 902076                | 34.0                    | 701062                 | 63.7                    |
| **Whether to have private health insurance or not** | | | | |
| No                     | 2461780               | 92.7                    | 873810                 | 79.4                    |
| Yes                    | 192760                | 7.3                     | 226285                 | 20.6                    |
| **Ownership status of any real-estate** | | | | |
| No                     | 828498                | 31.2                    | 491557                 | 44.7                    |
| Yes                    | 1826043               | 68.8                    | 608538                 | 55.3                    |
| **Smoking habit**      |                       |                         |                        |                         |
| No                     | 1938873               | 73.0                    | 637331                 | 57.9                    |
| Yes                    | 715668                | 27.0                    | 462764                 | 42.1                    |
| **Habit of buying daily newspaper** | | | | |
| No                     | 2549914               | 96.1                    | 972933                 | 88.4                    |
| Yes                    | 104626                | 3.9                     | 127162                 | 11.6                    |
| **Having cable TV**    |                       |                         |                        |                         |
| No                     | 2498205               | 94.1                    | 815828                 | 74.2                    |
| Yes                    | 156335                | 5.9                     | 284267                 | 25.8                    |
| **Habit of playing games of chance** | | | | |
| No                     | 2564355               | 96.6                    | 997298                 | 90.7                    |
| Yes                    | 90185                 | 3.4                     | 102797                | 9.3                     |
| **Habit of eating out**|                       |                         |                        |                         |
| No                     | 1787433               | 67.3                    | 294166                 | 26.7                    |
| Yes                    | 867107                | 32.7                    | 805929                 | 73.3                    |
| **Habit of shopping via internet** | | | | |
| No                     | 700297                | 26.4                    | 145047                 | 13.2                    |
| Yes                    | 1954243               | 73.6                    | 955048                 | 86.8                    |
| **Having credit card** |                       |                         |                        |                         |
| No                     | 1947951               | 73.4                    | 334959                 | 30.4                    |
| Yes                    | 706590                | 26.6                    | 765136                 | 69.6                    |
| **Savings habit**      |                       |                         |                        |                         |
| No                     | 924248                | 34.8                    | 578671                 | 52.6                    |
| Yes                    | 1730292               | 65.2                    | 521424                 | 47.4                    |
| **Age**                |                       |                         |                        |                         |
| Mean ± SD              | 59.40 ± 17.74         |                         | 47.68 ± 17.55          |                         |
The parameter estimates under weighted data are presented in Table 9. The threshold value $\theta$ for the Lognormal-Pareto type II composite regression model when the swarm size is 40, the maximum number of iterations is 250 and $c_1 = c_2 = 1.49$ is estimated as 30.626. Households, whose monthly consumption expenditure is equal to or less than 30.626 TL will be called “low expenditure class”, those above 30.626 TL will be called “high expenditure class”. This classification will be used in following comments.

When parameter estimates in the weighted data in Table 9 are examined, it is seen that the age variable has a decreasing (0.9996 times) effect on the amount of average monthly consumption expenditure in low expenditure class. Similarly, it has a decreasing (0.9971 units) effect on the amount of high expenditure classes. For low expenditure class, the average monthly consumption expenditure of women compared to men is 1.24659 times larger. For high expenditure class, women spend 1.248376 units more than men.

For low expenditure classes, the average monthly consumption expenditure of people who have the level of primary, secondary education and bachelor degree compared to those who do not complete any school is 1.1298, 1.097 and 1.114 times higher, respectively. For high expenditure class, the average monthly consumption expenditure is increasing as the level of education is increasing.

For both expenditure classes, the average monthly consumption expenditure is lower for married individuals compared to single ones (For low expenditure class: 0.9211 times; for high expenditure class: 0.9481 units). For low expenditure classes, the average monthly consumption expenditure is higher for individuals who have the habit of buying daily newspaper, have smoking habit and have the habit of online shopping while for high expenditure classes, the average monthly consumption expenditure is lower for individuals who have the same habits.

For both expenditure classes, the average monthly consumption expenditure is higher for individuals who have a savings habit, have the habit of dining out, are working, have their own property, have Cable TV, have the habit of playing games of chance, have credit card and have private health insurance.

Summary statistics calculated for low and high expenditure classes are given in Table 11.

### 6 Conclusion

The main purpose of this study is to propose a novel composite regression model for heavy-tailed data. The proposed model consists of two components which are Lognormal and Pareto type II distributions. Differently from Gan and Valdez’s study (Gan and Valdez 2018), the proposed model provides an important contribution to the existing literature. Since the PSO algorithm is capable of dealing with a large number of parameters, is a powerful tool for solving complex systems and the algorithm’s elapsed time is quite short, in this study, the PSO algorithm is used to obtain parameter estimation of the proposed model. The PSO algorithm is implemented in two stages. The first stage is designed to find optimal hyper-parameters of the PSO algorithm. The second stage of the simulation setting is performed to obtain parameter estimates of the proposed model using the PSO algorithm with the best values of the hyper-parameters according to the first stage of the simulation results. In addition, the proposed composite Lognormal-Pareto type II regression model is compared to those of Lognormal, Lomax, and Gamma regression models using BIC values. The results of the simulation study show strong support in favor of the proposed model.

To demonstrate applicability, monthly consumption expenditure and affecting factors using the National Household Budget Study conducted by the Turkish Statistical Institute (TurkStat), are modeled via the proposed composite Lognormal-Pareto type II regression model. To set up a regression model for providing better fit to the data is challenging. As it can be seen in Fig. 1, monthly consumption expenditure have heavy-tailed behavior. Therefore, standard regression models are deficient in capturing the relationship between monthly consumption expenditure and its related
variables. The results of Household Budget Study data show that the proposed regression model indicates more favorable fit than Lognormal, Lomax, Gamma and Gamma-Pareto type II regression models.

Additionally, the sampling design of Household Budget Study is stratified two-stage clustering sampling. Accordingly, it is another challenge to obtain parameter estimates under-weighted data. To address this challenge, both the model equation and PSO algorithm are updated by taking into account sampling weights. Finally, parameters are estimated for weighted data and the effect of explanatory variables on average monthly consumption expenditure is 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 like to thank Dr. Ayten Yıgiter for valuable comments to help improve the presentation of this paper. I would also like to thank Dr. Derek S. Young for final check and language editing.

Funding Not applicable.

Availability of data and material The data was provided Turkish Statistical Institute (TurkStat) with restricted use. I am not allowed to share the data for this research. I would like to thank Dr. Ayten Yıgiter for final check and language editing.

Funding Not applicable.

Conflict of interest Not applicable.

Declarations

Conflict of interest Not applicable.

References

Acitas S, Aladag CH, Senoglu B (2019) A new approach for estimating the parameters of Weibull distribution via particle swarm optimization: an application to the strength of glass fibre data. Reliab Eng Syst Saf 183:116–127

Amoshay M, Shamsi M, Sedaghi MH (2016) A novel flexible inertia weight particle swarm optimization algorithm. PLoS ONE 11(8):e0161558

Bolviken E (2014) Computation and modelling in insurance and finance. Cambridge University Press

Browne RP, McNicholas PD (2015) A mixture of generalized hyperbolic distributions. Can J Stat 43(2):176–198

Calderin-Ojeda E, Kwok CF (2016) Modeling claims data with composite Stoppa models. Scand Actuar J 2016(9):817–836

Cohen AC (1967) Estimation in mixtures of two normal distributions. Technometrics 9(1):15–28

Cooray K, Ananda MM (2005) Modeling actuarial data with a composite lognormal-pareto model. Scand Actuar J 2005(5):321–334

Dietz E, Böhning D (1995) Statistical inference based on a general model of unobserved heterogeneity. In: Statistical modelling, pp 75–82

Dominicy Y, Sinner C (2017) Distributions and composite models for size-type data. In: Advances in statistical methodologies and their application to real problems, p 159

Eberhart RC, Shi Y (2000) 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

Engelbrecht AP (2013) 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

Everitt BS, Hand DJ (1981) Finite mixture distributions. Springer, Berlin

Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Springer Science Business Media, LLC, New York

Gan G, Valdez EA (2018) Fat-tailed regression modeling with spliced distributions. N Am Actuar J 22(4):554–573

Hall DB, Wang L (2005) Two-component mixtures of generalized linear mixed models for cluster correlated data. Stat Model 5(1):21–37

Karaboga D (2017) Yapay Zeka Optimizasyon Algorimalari. Nobel Akademi Yayıncılık

Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of ICNN’95-international conference on neural networks, vol 4, pp 1942–1948. IEEE

Lee SX, McLachlan GJ (2013) On mixtures of skew normal and skew t-distributions. Adv Data Anal Classif 7(3):241–266

McLachlan GJ, Basford KE (1988) Mixture models: inference and applications to clustering. Marcel Dekker, New York

McLachlan GJ, Peel D (2000a) Finite mixture distributions. Wiley

McLachlan GJ, Peel D (2000b) Finite mixture models, probability and statistics—applied probability and statistics section, vol 299. Wiley, New York

Özsağlam MY, Çunkaş M (2008) Optimizasyon problemlerinin çözümlü için parçaçık sürü optimizasyonu algoritması. Politeknik Dergisi 11(4):299–305

Parxopoulos KE, Vrahatis MN et al (2002) Particle swarm optimization method for constrained optimization problems. Intell Technol Theor Appl New Trends Intell Technol 7(6):214–220

Peel D, McLachlan GJ (2000) Robust mixture modelling using the t distribution. Stat Comput 10(4):339–348

R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/

Saravanan R (2006) Manufacturing optimization through intelligent techniques. CRC Press

Scolnik DP (2007) On composite lognormal-pareto models. Scand Actuar J 2007(1):20–33

Shen J (2011) 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

Shi Y, Eberhart R (1998) A modified particle swarm optimizer. In: 1998 IEEE international conference on evolutionary computation proceedings. IEEE world congress on computational intelligence (Cat. No. 98TH8360), pp 69–73. IEEE

Shi Y, Eberhart RC (1999) 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

Teodorescu S, Vernic R et al (2009) Some composite exponential-pareto models for actuarial prediction. Rom J Econ Forecast 12(4):82–100

Teodorescu S, Vernic R (2013) On composite pareto models. Math Rep 15(65):11–29

Tooze JA, Grunwald GK, Jones RH (2002) Analysis of repeated measures data with clamping at zero. Stat Methods Med Res 11(4):341–355

van Dijk B (2009) Essays on finite mixture models. Ph.D. thesis, Erasmus University Rotterdam, Tinbergen Institute
Yau KK, Lee AH, Ng AS (2003) Finite mixture regression model with random effects: application to neonatal hospital length of stay. Comput Stat Data Anal 41(3–4):359–366

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.