CHAOS
Theory and Applications
in Applied Sciences and Engineering

e-ISSN: 2687-4539
RESEARCH ARTICLE
Vol.6 / No.1 / 2024 / pp.63-72
https://doi.org/10.51537/chaos.1422400

Future Prediction for Tax Complaints to Turkish Ombudsman by Models from Polynomial Regression and Parametric Distribution

Mehmet Niyazi Çankaya [*,1] **and Murat Aydın** [α,2]

*Faculty of Applied Sciences, Department of International Trading and Finance, Uşak University, Uşak, Turkiye, αFaculty of Applied Sciences, Department of Accounting Finance and Management, Uşak University, Uşak, Turkiye.

ABSTRACT The aim of this study is to forecast the amount of tax complaints filed with the Turkish Ombudsman in the future and whether or not policymakers require a specific tax Ombudsman. The polynomial regression for discrete data set is proposed to fit the number of events of tax complaints in the period from years 2013 to 2021. The artificial data set is generated by models which are polynomial regression and parametric distribution. The location, scale and shape parameters are determined according to the smallest value between the observed and predicted dependent variable. After determining the smallest value for the tried values of shape parameter and the parameters of polynomial regression, the best value determined by grid search for shape parameter is around 1.07. Thus, the heavy-tailed from of exponential power distribution is gained. The artificial data sets are generated and sorted from the smallest to biggest ones. The maximum values are around 700 and 800 which can be regarded as future prediction because the distance among observations is taken into account by models from polynomial regression and parametric distribution. Since the polynomial regression and the parametric models are used simultaneously for modelling, the distance among observations can also be modelled by parametric model as an alternative approach provided.

KEYWORDS
Estimation
Inference
Public economics
Parametric models
Simulation

INTRODUCTION

Estimation is a challenging topic that needs to be improved by advancing the tools in the statistical literature. Many data sets in the applied sciences should be modelled efficiently. For example, the number of Ombudsman who hear complaints from citizens about failures, actions and decisions by public authorities is discrete data such that natural numbers are used to represent these kind of data sets. The main aim of the Ombudsman is to fight against abuse of rights, omissions, wrong decisions and delays for citizens. As the institution of the ombudsman is important, a design for estimating the number of Ombudsman will be an important issue in the near future. A combination of polynomial regression as a parametric model based on the regression case and the parametric distribution, for example the exponential power distribution, based on the distributional form of dependent variable or any variable can be proposed to set an approach for forecasting (Mineo and Ruggieri 2005). Note that if the data set is discrete, the discrete models such as binomial, generalized form of binomial, etc. can be used to fit the data set. If the data set is continuous, the continuous parametric model such as exponential power distribution and its variants such as skew, model, trimodal family for a known parametric models can be generated and used. The compound forms of distributions are also derived to model the data set more efficiently as far as we can do (Balakrishnan and Nevzorov 2004).

In the working principle of nature providing the observed values after the experiment has been performed, it is not easy to imply that a data set can be only one parametric model. There is a hard indeterminacy in the nature of data formation. For this reason, the regression form can be a bridge for us to fit the data if we insist on driving the tools as alternative objectives in this study. Since estimation is a fluctuation around function f used for the representation of parametric model such as exponential power distribution, that is, we can get $\hat{f}$ representing the estimated form of f due to the

Manuscript received: 19 January 2024,
Revised: 4 March 2024,
Accepted: 20 March 2024.

[1]mehmet.cankaya@usak.edu.tr
[2]murat.aydin@usak.edu.tr (**Corresponding author**).

finite sample points of f, it is logical to suggest that a regression form can be used for data to perform a modelling instead of using directly parametric model to fit the data set (Vila *et al.* 2020, 2022).

Especially, since we have small sample size of data, it can be a gate for us to overcome the problem about the case where we have few data that will be needed to fit precisely as far as we can achieve the joint work between polynomial regression as a parametric model and the parametric distribution when compared by the non-parametric forms (Härdle *et al.* 2004; Hunter 2023). Thus, we can have an applicable form when we use the computational tool for this marriage. Especially when the sample size is small, the parametric models cannot be very powerful because we do not have enough data where the data set comes from or it is very rough to know how the real data set has occurred while the experiment is being conducted.

Note that a data set can show a regression or a polynomial movement/pattern. Since we are proposing to use the regression equation in order to model the data set observed over time, the time series form can be suggested as a regression case. On the other hand, the bulk of the data at each time throughout the time period cannot be a fixed variance. In particular, it is reasonable to observe that a non-identical distributed data throughout time is indispensable observed in the nature of the data set. That is, there may be heteroscedasticity (Mokhtari *et al.* 2022). Such non-identical movement or heteroscedasticity can be modelled by using the peakedness parameter p in the exponential power distribution when the regression case is used. On the other hand, there is a struggle between the chosen parametric model for the distribution of the error term of the regression equation to determine whether or not there is heteroscedasticity in reality. There may be another reason to imply the existence of heteroscedasticity if we change the analytical form of the regression model.

In the regression equation, the square of the error term, known as the estimated variance, is used to generate the artificial dataset, that is, we want to estimate the scale parameter for the dataset using the regression approach with parametric model based on the peakedness parameter p. This approach is an alternative if we want to estimate the scale parameter for the data set. Since the discrepancy can be detected by the error term in the regression, the generation for artificial data can also be done by using the peakedness parameter of the exponential power distribution (Mineo and Ruggieri 2005). On the other hand, it is reasonable to expect that the peakedness (p) and scale (σ) parameters can interact, because they are parameters which are responsible for changing the shape of the function. Note that the interaction can lead to occur the heteroscedasticity. The main problem is about the future prediction of tax complaints when the polynomial regression and parametric model are used together. Thus, we can perform an efficient fitting by using not only a trend as regression (location) but also the error terms of polynomial regression model which are modelled by exponential power distribution. It should be noted that heteroscedasticity can be modelled as an alternative approach if the parametric distribution is used.

Taxation is an important and effective way to manage government resources fairly. There are many studies to model the tax of governments which use the distribution assumption to model the tax managed by governments. The role of tax is investigated and the different suggestion on the tax regulation in the management modelling or planning in the government have been carried out in the different and directions which are still investigating what the government policy should be or some markers in the financial markets push or drive the policy management to increase the efficiency and correctness on the process improvement (Bala and Biswas 2005). It is generally accepted that the tax management is very problematic and the Ombudsman is an inevitable community for the set and built law system to touch the correct and effective decision in the timing period in the tax system. We believe that the Ombudsman should be supported and improved by means of using different directions. Thus, the role of tax management can be improved and more accurate decision can be reached by the responsible drivers in the system of government policy.

These improvements, such as rotations, directions, suggestions, etc., make an automatic control and checked feedback in the working principle of the set system, which should be improved simultaneously, no matter when it finally stops. In other words, it should be a system which is a rounding around itself and it should be controlled by independent communities which can work out independently and put their suggestions briefly touching every point of the picture carefully and do not cover any potential uncovered parts in the tax system management. Such a situation can be achieved by carrying out the stricter effective analysing procedure which will not be influenced and touched. The reality of the current system should have its clarifications in order to discover the hidden parts in the system. For this purpose, the Ombudsman is a key institute for us to round up the system and so the quality control can be tried to be guaranteed (Serrano 2007).

Basically, this institution listens to taxpayers' complaints and solves their problems. It also improves the organisation of the tax service. In America, for example, the name is even more different. The Taxpayer Advocate Service is an institution in the US that intervenes when the Internal Revenue Service does not want to do so. There is at least one local Taxpayer Advocate in every state. There is no such thing as a tax ombudsman in US tax law. In Spain, as in other countries, there is a tax ombudsman who depends on the government. He balances the relationship between the taxpayer and the administration. In the Law of Taxpayers' Rights there are five functions of the Tax Ombudsman. These are; informative action, democratic control, alternative dispute resolution, and improving the moving legal system (Bala and Biswas 2005).

The organization of the paper is as follows: Section for preliminaries introduces the parametric model and the estimation method. In next section, the problem is solved by using models from polynomial regression and parametric distribution. The numerical results and figures are given as a separate section. The last section is divided for the conclusions.

PRELIMINARIES

Parametric model

The normal distribution is commonly used and it is a popular distribution. The generalisation of the normal distribution and the different probable directions are proposed by (Çankaya 2018). The exponential power distribution is one of them and it has a peakedness parameter being responsible for determining the peak of the function. The empirical distribution of the data can be a way to observe how the shape of the function behaves.

The analytical expression of the exponential power distribution is given by the following form:

$$f(x;\mu,\sigma,p)=\frac{1}{2\sigma p^{1/p}\Gamma(1+1/p)}\exp\{-\left|\frac{x-\mu}{p^{1/p}\sigma}\right|^{p}\} \qquad (1)$$

$\mu\in\mathbb{R},\sigma>0,p>0$ represents parameters for location, scale and peakedness of the function, respectively. The exponential power distribution is the special form of asymmetric bimodal exponential power distribution (Mineo and Ruggieri 2005; Çankaya 2018).

Estimation method: Maximum likelihood estimation

If there is a distribution of the error terms in the regression model, or the error terms are assumed to have a distribution such as normal, Student t, exponential power, asymmetric bimodal exponential power distribution (Çankaya 2018) etc., then the error terms can be a member of a parametric model. If a distribution is used in the regression model as the location of the parametric model, then maximum likelihood estimation is preferred to estimate the parameters of the regression model. There are important properties which are efficiency, consistency, minimum variance, etc. when maximum likelihood estimation method is used (Lehmann and Casella 2006).

The function `lmp` in 'normalp' at RStudio 2023.09.1+494 free open statistical software is used to fit a regression model with the dependent variable y and the independent variables $x_1, x_2, \cdots, x_k$. It can be used when the errors are distributed as an exponential power distribution (Mineo and Ruggieri 2005; Lehmann and Casella 2006). The maximum likelihood estimation method gives us advantage of using the assumed parametric model to estimate the regression parameters when the errors are distributed as the corresponding parametric model. Each observation can be considered as an output, i.e. there can be a regression expression that can be applied to find the relationship between the observations in the period. The next section provides the detailed discussion and methodological contribution to assess the distribution of the variable y and the variable ε. Thus, the regression expression, as a fixed part that tries to represent how a real relationship exists between variables, can be used to determine the distribution of the error term ε.

DESCRIPTION OF THE PROBLEM

There is a potential overlap between the chosen regression model, with its corresponding distribution of the error term of the regression model, and the chosen kernel smoothing techniques which are based on the parametric or semi-parametric approaches. In modelling, not only the assumed regression model but also the distribution of the error terms are two components that influence each other. Thus, if we can determine what the distribution of the error term can be, then the distribution of the observation term y as a dependent variable will also be determined. Each data is the replication of the previous case or there is a potential dependence among the previous cases of the data set, as in the case of the random walk in the stochastic process (Iacus *et al.* 2008). The data set can be reorganised by using the regression approach which we can consider to apply. It is logical to expect that a data set can be represented by a polynomial approach. Since a polynomial approach can be performed on the data set, the future prediction can be performed by polynomial regression. In the context of the artificial data set, it is reasonable to perform a random number generation procedure to observe which data can be artificially observed.

In order to fit the data set via the proposed function, the polynomial regression approach can be used. Secondly, a parametric model for error term of regression can be suggested. After the regression case can be done, the peakedness parameter of the exponential power distribution can be determined by using the grid search approach. Since we perform such an approach to determine the value of the peakedness parameter p for peakedness, the computational cost of simultaneously estimating the location, scale and peakedness parameters can be solved as an alternative approach for modelling. This is an important contribution when the sample size is small and we need to use a parametric model to generate the artificial dataset. Why do we need to generate an artificial dataset? The future prediction or the probable numerical values for the questionnaire of phenomena can be determined, as an alternative approach if the regression case is not the only solution for the future prediction.

In other words, it is logical to observe that each data can be a member of a polynomial movement in the forthcoming situation in an experiment. Even if we assume that each data is independently distributed, it is reasonable to perform a polynomial motion among the data set. Independence can be a restrictive approach, or an alternative comment, is that it is already well known that a number can be made to belong to a polynomial function. A rounding around a data can be produced by another data, which shows that it is possible to carry out modelling using a polynomial approach based on the regression case. In the statistical literature, this approach should be preferred when the computational cost increases as the number of parameters to be estimated increases. In our approach, the first step was to determine the peakedness parameter using the grid search algorithm. The second step is the estimation of the location and scale parameters when the peakedness parameter p is given as a fixed value determined by using the case of regression with polynomial motion. Finally, random numbers are generated for the estimated values of the position and scale parameters. The maximum likelihood estimation method is used to obtain the estimators of the parameters when the fixed value of the peakedness parameter p determined by the polynomial approach is given.

Econometricians search for regression models to fit the data set based on the time series sense. Statisticians make the overlap between the chosen regression model and the error term (ε) of the regression model. Both scientists try to find the best strategy for modelling the data set. The advantage of being a statistician may be more beneficial because a statistician focuses on the distribution of the error term. The distribution of the error term in the equation (2) corresponds to the distribution of the variable y which represents the observed value. Even if the nature of the observed value of tax complaints is discrete, the polynomial regression model can also be proposed to fit the number of events in the period. In this case, since the events $y(t) := y_t$, where t represents time, depend on time as in the case of the random walk in the stochastic process, there can be a potential correspondence from the discrete data to the continuous data.

For further discussion, the tax complaints occur due to many reasons based on the continuous observations from the government money process. The currency and the economic indicators are responsible tools that lead to have the continuous observations. That is, in other words, when we make a projection where the discrete data comes from, it is observable to detect that the continuous observations touch occurring the discrete data sets. Even if the binomial regression or the corresponding counterparts are used to model the discrete data sets, the continuous case is more flexible to generate artificial data sets; because the peakedness parameter is more important to determine the distribution of the data set. On the other hand, the continuous data already represents the discrete data as a neighbourhood framework. In addition, the binomial distribution converges to the normal distribution with peakedness parameter $p = 2$ in the exponential power distribution. In this direction, there can be a transfer from the discrete data to the continuous data (Sicuro *et al.* 2015).

It should be noted that even if the discrete data are analysed in the regression case, the continuous distribution can be used for the application; because there is a correspondence between the polynomial function and the gamma function (Alzer and Grinshpan 2007).

In this sense, as discussed above, the peakedness parameter plays a key role in determining the distribution of the data. The exponential power distribution is symmetric around the location parameter and so the general tendency of the economic indicators is assumed to be symmetric due to the nature of the experiments; because the government applies the tax rule equally balanced on the people in the country and the symmetric distribution can play the role for modelling the data sets output by many reasons occurred on the economic indicators (Haberman 1989; Coles *et al.* 2001; Mineo and Ruggieri 2005; Çankaya 2018; Çankaya and Arslan 2020).

Materials and Methods

Regression model Regression models are generally used to set the relationship between at least two variables. The nature of the dependent and independent variables can be determined according to what a researcher investigates. When the observations from the experiments are measured, they are analysed according to the researcher's objective. In our framework, a polynomial regression equation is used to model the observations by using the sequence occurred over time. In other words, in mathematical terms, the set of data or observations can be expressed in terms of the movement over time.

The assumed regression model for representation of reality is as follow:

$$y = a_0 + a_1 x^{p_1} + a_2 x^{p_2} + a_3 x^{p_3} + \varepsilon \tag{2}$$

where ε is a random variable assumed to have an exponential power distribution. The parameters p_1, p_2 and p_3 are responsible for the different degrees of the polynomial function in the equation (2). a_0, a_1, a_2 and a_3 are regression parameters estimated using the `lmp` function with different trial values of the peakedness parameter p.

The sampling form of the equation (2) is given by

$$y_t = a_0 + a_1 x_t^{p_1} + a_2 x_t^{p_2} + a_3 x_t^{p_3} + \varepsilon_t,\ t = 1, 2, \cdots, n \tag{3}$$

where $x_t = t$ as an explanatory variable representing the time (year) and n is the number of sample size.

It should be noted that since the polynomial movement among the data sets is assumed to be expressed by the polynomial regression case, it should be preferred to model the upcoming events. It is important to note that the distribution of the error term ε begins to play a role in determining the value of the peakedness parameter in the parametric model used. In the general setting, it is logical to propose a parametric model that has a peakedness parameter. Thus, using the role of the peakedness parameter will give us an advantage in determining the tail movement and thus we can have a chance for future prediction instead of using the regression case in future prediction. Such an approach makes an alternative suggestion/contribution to the statistical literature to determine the value of the peakedness parameter p as an alternative approach. On the other hand, the peakedness parameter p and the powers p_1, p_2 and p_3 in the regression model can play same role when there is a conceptual equivalence and the definition of the regression, i.e., $\mathbb{E}(Y/X = x)$ with the non-fixed value of scale parameter (or with heteroscedasticity), takes into account, i.e., $Y \sim D(\mathbb{E}(Y/X = x_t), \sigma(X = x_t))$. Since the number of samples is small, it will not be easy to determine the almost exact peakedness of the assumed parametric model. For this reason, we have provided an alternative approach to determine the peakedness parameter from the data set. The assumed model for the distribution of ε is the exponential power one.

The nature of the occurred phenomena tricks our approach, because tax complaints can have a heavy tailed distribution due to the nature of the tax complaints (Jenkins 2017).

Algorithmic schema for computational procedure The following steps show the schematic algorithm how the computational procedure is conducted to reach the value of peakedness of parameter p if the polynomial regression in equation (2) is used (see Appendices).

1. Determine the peakedness parameter p by using `lmp` function in the `normalp` package in RStudio 2023.09.1+494 software
2. Try different values of the parameters p to get the different probable the smallest difference between the predicted y, i.e. $\hat{y}$, and the observed y as data
3. Use the polynomial regression in equation (2)
4. Set a vector for the tried values of peakedness parameter p
5. For each values of p_1, p_2 and p_3 which are $(0, 75, 1.25]$, $(1.75, 2.25]$ and $(2.75, 3.25]$, respectively, the values of peakedness parameter p are determined according to the smallest value, i.e. predicted error ($\hat{\varepsilon} = y - \hat{y}$), of the distance between the predicted y, i.e. $\hat{y}$, and the observed y. If an appropriate p value which satisfies the smallest value for $\hat{\varepsilon}$ is determined, then the determined value of p in the 125 000 times due to `set1=50`, `set2=50` and `set3=50` makes the probable appropriate values of p, which is obtained by each values of p_1, p_2 and p_3

If the number of degree of power parameter in the polynomial regression in equation (3) is increased according to `for` loop given above, the different forms of the values of p_1, p_2 and p_3 in regression equation can be tried to model the movement among the observations. The role of sensitivity of higher order powers p_2 and p_3 should be applied to fit the observations more precisely, because the degree of polynomial regression can be versatile due to the chosen values of p_2 and p_3 especially. Note that it is possible to apply different polynomial regression with higher order polynomial power; however, the dependence structure among the right hand side of regression equation (3) can start to be a problem, leading to a multicollinearity problem. The dependence can be tricked according to the chosen values of parameters p_1, p_2 and p_3. To avoid the more biased estimation for the parameters a_0, a_1, a_2 and a_3 due to the probable structure of the dependence among the variables x^{p_1}, x^{p_2} and x^{p_3}, we continue to follow the regression equation in (3). That is, the new variable x^{p_4} or other variables have not been added to the regression model, because the degree of perturbation should be avoided for mathematical reasons as well (Montgomery *et al.* 2021).

Since the nature of the polynomial approach can have the negative estimated values for the parameters, the forecast in the forthcoming numbers for the occurred events from tax complaints cannot be determined by using the regression approach directly. On the other side of the picture, since we have few data sets for applying the regression case, the rank problem can occur due to the number of regression parameters starting to be close to the number of sample size $n = 9$ in our case (Stanimirović 2017). For this reason, we prefer to use two steps for the approach in the future prediction instead of doing the prediction using the regression equation.

NUMERICAL RESULTS

Tools: polynomial regression and correlation

The numerical results with illustrative representations are provided to observe how the regression equation produces the results which are the estimated values for correlation, standard deviation and their empirical probability density function (pdf) computed by means of the `EnvStats` package with kernel functions (such as Refs. (Härdle *et al.* 2004; Hunter 2023)) in RStudio 2023.09.1+494 software.

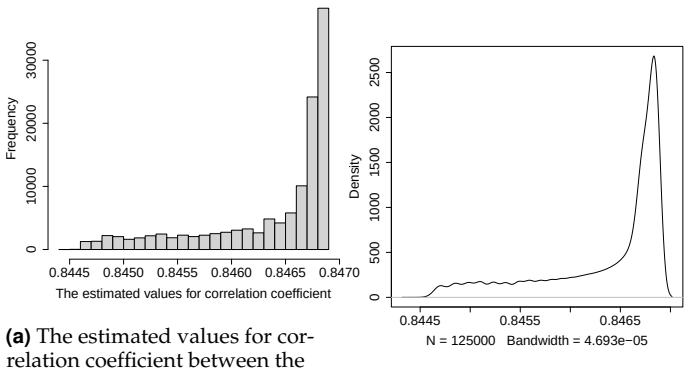
The parameters a_0, a_1, a_2 and a_3 in regression model at equation (3) can be estimated and provided by the following form:

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x^{p_1} + \hat{a}_2 x^{p_2} + \hat{a}_3 x^{p_3} \tag{4}$$

where x is independent variable having numerical values from 1 to the sample size n and p_1, p_2 and p_3 are power parameters which are responsible to make a flexible fitting on the data.

The equation (4) is used to get the predicted $\hat{y}$. The correlation values are computed by using the correlation formula for the observed y and the predicted $\hat{y}$ values (Lehmann and Casella 2006). According to the codes in appendix, the estimated values for the correlation coefficient are given by figure 1. Figure 1 informs us for the estimated values of correlation coefficients between the observed values of y and the estimated values of $\hat{y}$ given by equation 4.

The performance of future prediction depends on the degree of the values of correlation. That is, we have success at the degree %85 for trusting the numerical values generated artificially. For this aim, the values at the figure 2a should be preferable to represent the non-identically case of distributed data set, i.e., if the estimated values of scale parameter are big, then we can have values being far from the bulk of the data.



(a) The estimated values for correlation coefficient between the predicted and the observed values of y

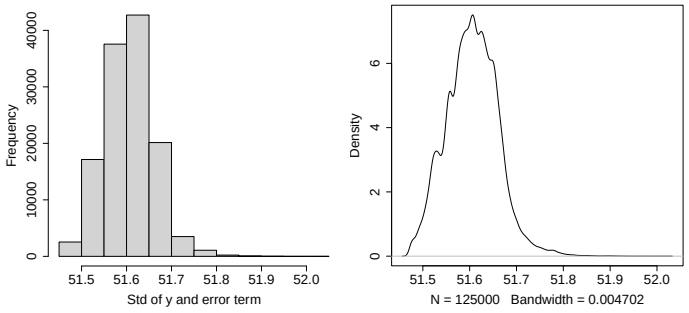
(b) Empirical pdf of the estimated values for correlation coefficient

Figure 1 $\hat{\rho}$: The estimated values for the correlation coefficient

Figures 1a and 1b represent the histograms and the smoothed form of empirical pdf according to the frequency when the bandwidth from kernel estimation in `EnvStats` package is determined automatically as nearly as being small. The same illustrations are given by forthcoming figures 2-5 and 7.

The statistics for scale parameter as a dispersion measure

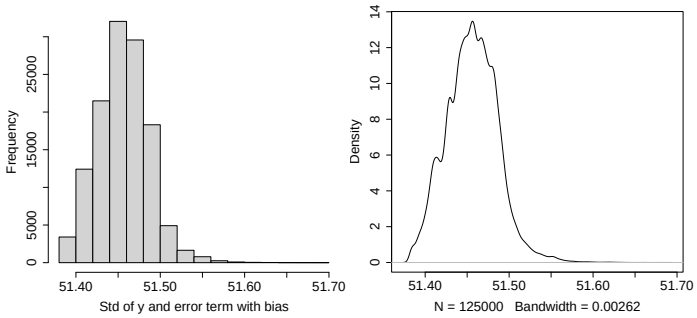
The following figures represent the numerical values generated at random from the exponential power distribution with the determined value of the peakedness parameter $p = 1.07$. These numerical values are sorted from smallest to largest. They are then plotted according to these sorted values. Each sorted value is replicated and the number of replication is 10 000. Since we have the sorted values, we have the advantage of being able to plot the values that are the maximum and the previous ones which are represented by (-1), (-2), (-3), etc. Thus, the probable values which are maximum and the previous values before maximum can be observed. It is important to note that since we are replicating, the random data may have different values for each replication. For such a design, there may be some cases where there are the same numerical results in the simulation given by Figure 8.



(a) $\sqrt{Var(y) + \widehat{Var}(\hat{\varepsilon})}$ from $\hat{y}$ and $\hat{\varepsilon}$ at the chosen values of parameter p

(b) Empirical pdf of $\sqrt{Var(y) + \widehat{Var}(\hat{\varepsilon})}$

Figure 2 Histogram and empirical pdf for $n - k$ in computation of the estimated error, $\hat{\varepsilon}$



(a) $\sqrt{Var(y) + \widehat{Var}(\hat{\varepsilon})}$ from $\hat{y}$ and $\hat{\varepsilon}$ at the chosen values of parameter p

(b) Empirical pdf of $\sqrt{Var(y) + \widehat{Var}(\hat{\varepsilon})}$

Figure 3 Histogram and empirical pdf for n in computation of the estimated error, $\hat{\varepsilon}$

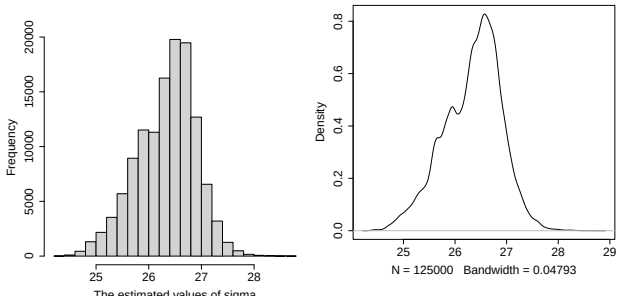
The distance between the observed variable y and the predicted variable $\hat{y}$ is defined as the variation. It is also called as error term ε. The sampling form of ε, i.e., $\hat{\varepsilon}$, is given by the following expression for the exponential power distribution:

$$\hat{\varepsilon} = \frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^{\bar{p}} \tag{5}$$

$$\hat{\varepsilon} = \frac{1}{n-k} \sum_{t=1}^{n} (y_t - \hat{y}_t)^{\bar{p}} \tag{6}$$

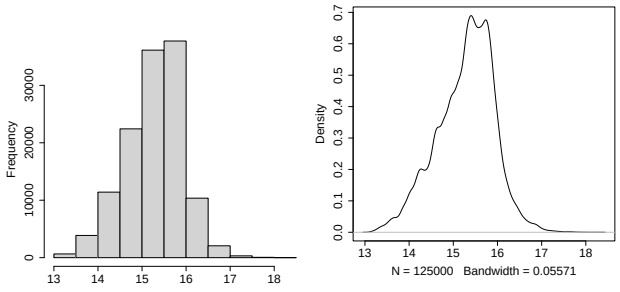
Comparing figures 2 and 3, the values in figure 3 are smaller than those in figure 2 because the formula for the error term, ε, is

$\frac{1}{n}\sum_{t=1}^{n}(y_t-\hat{y}_t)^p$ for exponential power distribution. For the figure (2), $\frac{1}{n-k}\sum_{t=1}^{n}(y_t-\hat{y}_t)^p$, where k is the number of the estimated parameter. Note that the error term can also be considered as a scale parameter. Then we have the estimated values of the parameter σ, given by the figures 4 and 5.



(a) $\sqrt[p]{\frac{1}{n-k}\sum_{i=1}^{n}(y_i-\hat{y}_i)^p}$ **(b)** Empirical pdf of $\sqrt[p]{\frac{1}{n-k}\sum_{i=1}^{n}(y_i-\hat{y}_i)^p}$

Figure 4 Histogram of $\hat{\sigma}$ and empirical pdf for $n-k$ in computation of the estimated error, $\hat{\varepsilon}$



(a) $\sqrt[p]{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^p}$ **(b)** Empirical pdf of $\sqrt[p]{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^p}$

Figure 5 Histogram of $\hat{\sigma}$ and empirical pdf for n in computation of the estimated error, $\hat{\varepsilon}$

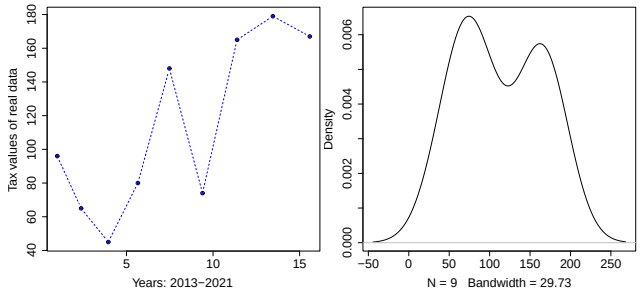
Since the determined value for the parameter p is around 1.07 (see figure 7a), the behaviour of the exponential power distribution is a heavy-tailed one. Thus, the future prediction may be more representative for the target in which we can safely use it, taking into account the predictive performance of our approach for the probable cases in the future. Since the number of sample size n is 9, it cannot be enough to suggest numerical values from the regression model in equation (2), because the polynomial movement cannot be enough to evaluate how the future occurs. It is possible to use other methods based on the mode movement that can also be taken into account to model and analyse the real data set. In this case, there can be parametric and non-parametric models that can produce the light-tailed movement. In such a case, we can have the numerical values that cannot be far from the location as the central tendency of the empirical data set. In our statistical analysis, the year 2100 can also be proposed for the future prediction.

In each replication, the number of samples is 100. According to the estimated values used for the location and scale parameters, we can have the negative values due to the nature of the parametric model used, which is the exponential power distribution with $p=$ 1.07. Since the awareness of the population about the Ombudsman system is reflected in different areas of the tax and financial system, the numbers that represent the case of the Turkish Ombudsman consulted for the tax complaints can be increased. It is surprisingly important to note that the results in the 2013-2021 period provide the analysis results that give the heavy-tailed function, which can provide an advantage for future prediction even if the numerical values are discrete.

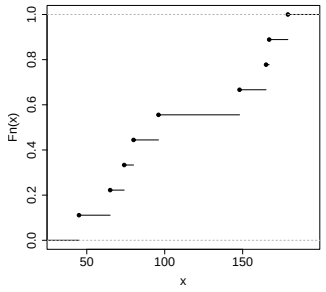
It is known that the discrete data can show the representation of a continuous case if the number of replications of events in the experiments is increased or the big law of large numbers is applied as an asymptotic behaviour. For example, the binomial distribution approaches the normal distribution (Lehmann and Casella 2006). The figure 6a showing the scatter plot shows a polynomial movement, which may be one approach we propose to model the data set.

Illustrative purposes for observing behaviour of artificial and real data sets

Figure 6 shows the scatter plot, the empirical probability and the empirical cumulative distributions of real data set with sample size $n=9$. Figure 6b shows that there can be a bimodality on the data set. However, even if data is discrete and shows a bimodality, this is an extra situation needed to investigate. In our approach, we keep to follow polynomial regression and one-mode parametric model called as exponential power distribution, because the tax and the related part being Ombudsman can have a movement based on the time series.



(a) Scatter plot of real data in years 2013-2021 **(b)** Empirical pdf (Density) of real data



(c) Empirical cdf (F_n(x)) of real data

Figure 6 Illustrative representation for real data

Figure 7 shows the histogram, empirical and cumulative distributions of the determined values of the peakedness parameter p of the exponential power distribution. Cumulative distribution function (cdf) is the cumulated form of probability density function (pdf). Note that the bimodality in figure 6b can also be modelled by bimodal distribution. However, the main aim is to determine the tail behaviour in order to where the maximum values can be around. It should be noted that the empirical distribution of the determined values of the parameter p can be modelled using the smooth kernel estimation method (Härdle *et al.* 2004); however, the chosen kernel plays the role of determining the probabilities.

Instead of using the location estimation for the parameter p in its probable empirical distribution, we use the mean of the determined values of the peakedness parameter p with a sample size of 125 000; because we generate the artificial data set when the parameter p is close to upper values of 1, which will not affect the more accurately generated probable random numbers (see figure 8). On the other hand, since the maximum likelihood estimation method is used, the distribution of the determined values of p is expected to be asymptotically normal, which may allow to use the arithmetic mean as a statistic for the values of p (see figure 7b) (Lehmann and Casella 2006).

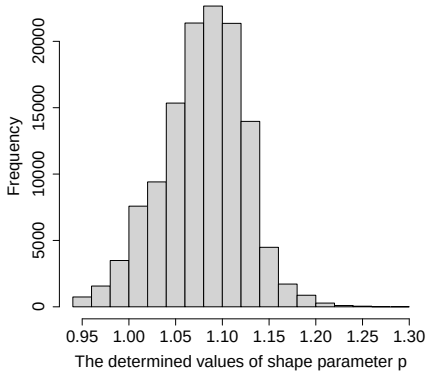
On the other hand, the values for the parameter p is around 1.07, which means that the synthetic data can get values from the tails, that is, it is possible to observe the values which is bigger than the real values, 179 as a maximum value. In such scenario, we add the role of distance among observations while performing the analysing on the data set. If the values of p tends going to 1 which leads to get reaching more degree of heavy tails for the exponential power distribution, it is reasonable to observe the values which can be bigger than 800. The codes in the appendices can be used to generate the synthetic data sets.

Figures 8 and 9 show the different numbers of estimated values generated by the exponential power distribution. In figure 9, the estimated value of the scale parameter is larger than in figure 8, which is why the estimated values from the simulation for the Ombudsman are around 800. An additional comment is that after sorting the synthetic dataset from smallest to largest, the previous values that come before the maximum value of the dataset are also given by the y-axis of Figures 8 and 9, labelled with the last value (-1) of the synthetic. (-1), (-2) and (-3) represent the ordered data. Note that the data set is sorted from the smallest to the biggest one, the last three values are chosen and they are represented by (-1), (-2) and (-3).

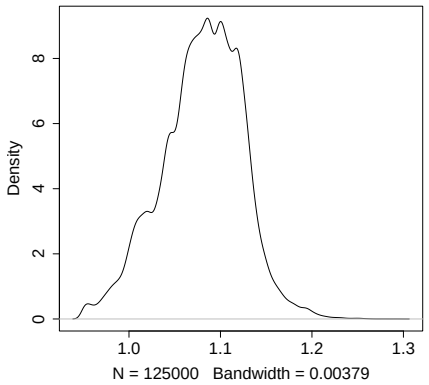
In addition, the fact that the trending slope looks away from the x-axis can be interpreted as an evidence that these artificially generated data will increase the necessity of the Ombudsman in the future. Note that even though the maximum value of real data sets is the number 179, the generated values for the artificial data sets are close to 800 as a maximum value; because we suggest to use the role of scale parameter as a dispersion measure which provides very important indicator for determining the behaviour of the data sets in any phenomena at the applied field of science. Thus, the role of scale parameter is an inevitable situation to touch more precisely the process in phenomena.

Note that the peakedness parameter p plays role as well importantly. Thus, the scale and peakedness playing role for determining tail behaviour of the function are in the class to determine the shape of function (Lehmann and Casella 2006; Arslan and Genç 2009; Çankaya 2018; Çankaya *et al.* 2019).

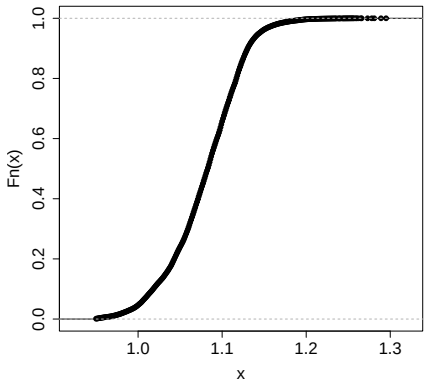
In figures 1-5 and 7, it should be noted that for the sake of the fact that representation of the frequency and the smoothed form can be more feasible, the histogram and smooth form of pdf are given separately at different scaling form of the cartesian coordinates.



(a) Histogram of p



(b) Empirical pdf (Density) of the determined values of parameter p



(c) Empirical cdf (F_n(x)) of the determined values of parameter p

Figure 7 Illustrative representation for the values of parameter p computed by `lmp` function with the smallest estimated error, $\hat{\varepsilon}$

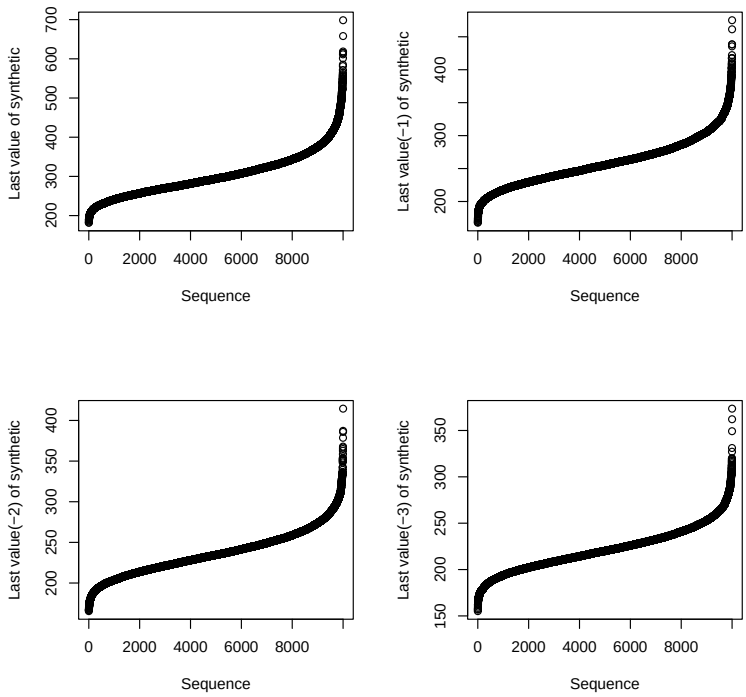


Figure 8 Case 1: The last values and previous ones of the replicated synthetic data for the ordered form of data for 10000 replication with $\hat{\sigma}$ from values represented by figure 2a

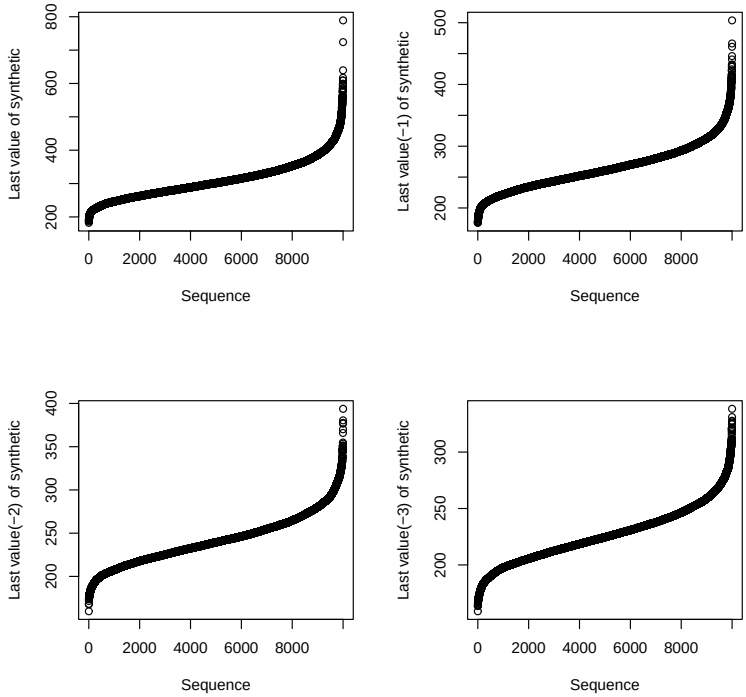


Figure 9 Case 2: The last values and previous ones of the replicated synthetic data for the ordered form of data for 10000 replication with $\hat{\sigma}$ from values represented by figure 3a

CONCLUSION

The polynomial regression models providing a relationship among observations have been proposed to help determining the distribution of the observations as well. That is, the distribution of the error term ε corresponds to the distribution of the variable y. Such an approach is important; because, when the small number of sample size is given, it is not an easy task to propose a parametric model to analyse the dataset accurately. The correlation between the observed y and the predicted y can be increased if the values of p from exponential power distribution and the power parameters p_1, p_2 and p_3 in regression model can play same role in fitting data well. Note that the parameters p, p_1, p_2 and p_3 are conceptually in same framework when the heteroscedasticity is taken into account. Such an approach provides a novelty for the study.

For example, using a distribution for discrete data cannot be an easy task to perform a precise modelling on the data set. The determined parametric model based on the regression models has been a special case of the exponential power distribution. Thus, we can have an advantage to generate the artificial data set to perform a prospective overview for modelling. If the peakedness parameter is around 1.07, the function is called the heavy-tailed form for the exponential power distribution. Since we have a heavy-tailed distribution, it is observable to get the numerical values which can come from tail parts of the function. For this reason, the synthetically generated numerical values are around 700 and 800 at most. Thus, we have suggested that the probable projection for the future prediction can provide numbers for the tax complaints upto year 2100. Future studies are in progress to suggest different materials, such as applying the heavy-tailed distributions and estimation methods for future projections of tax complaints in the Ombudsman (Çankaya 2021; De Gregorio *et al.* 2023).

APPENDIX

The main codes in the numerical evaluation for computation and regression case

The smallest value of $|y - \hat{y}|$ is used to determine the best value for the peakedness parameter p, because the best prediction performance can be gained, which means that the distribution of observation y can be determined by means of the distribution of error term which is equivalent to the observation y. The values of p_1, p_2 and p_3 are generated by using the following schema.

```
set1=50;set2=50;set3=50;pwl1=0.75;pwl2=1.75;pwl3=2.75;
indx=replicate(set1, numeric(set2));
ppval=replicate(set1, numeric(set2));
indx3ar<-array(c(indx, indx), dim = c(set1,set2,set3));
ppval3ar<-array(c(ppval, ppval), dim = c(set1,set2,set3));
pvariable=seq(0.95,1.35,0.001);
replication = length(pvariable);
for (i1 in 1:set1)
{
pwl1 = pwl1 + 0.01;
for (i2 in 1:set2)
{
pwl2 = pwl2 + 0.01/set1;
for (i3 in 1:set3)
{
pwl3 = pwl3 + 0.01/(set1*set2);
for (i in 1:replication)
{
regp<-lmp(y~x, p = pvariable[i]);
coefa <- regp$coefficients;
ypredict[i,] <- xx %*%
matrix(c(coefa[1],coefa[2],coefa[3],coefa[4]),4,1);
errory[i,] <- abs(y -  ypredict[i,]);
meanerrory[i] <- sum(errory[i,])/n;
}
indx3ar[i1,i2,i3]=min(which(sumerrory == min(sumerrory)));
ppval3ar[i1,i2,i3]=pvariable[indx3ar[i1,i2,i3]];
}
}
}
```

Estimation of error term distributed as the exponential power

Let us provide the codes showing how the equations (5) and (6) are adopted to the simulation in free open source statistical software RStudio 2023.09.1+494.

```
for (i1 in 1:set1)
{
pwl1 = pwl1 + 0.01;
for (i2 in 1:set2)
{
pwl2 = pwl2 + 0.01/set1;
for (i3 in 1:set3)
{
pwl3 = pwl3 + 0.01/(set1*set2);
indx3ar[i1,i2,i3]=min(which(sumerrory == min(sumerrory)));
ppval3ar[i1,i2,i3]=pvariable[indx3ar[i1,i2,i3]];
regp <- lmp(y~x, p = ppval3ar[i1,i2,i3]);
coefp <- regp$coefficients;
y_last_predict <- xx %*%
matrix(c(coefp[1],coefp[2],coefp[3],coefp[4]),4,1);
cor_vals3ar[i1,i2,i3] <- cor(y,y_last_predict);
var_n_eps3ar[i1,i2,i3] <-
sum(abs(y -  y_last_predict)^ppval3ar[i1,i2,i3]) / n;
sig_n_eps3ar[i1,i2,i3] <-
(sum(abs(y -  y_last_predict)^ppval3ar[i1,i2,i3]) / n)^
(1/ppval3ar[i1,i2,i3]);
var_eps3ar[i1,i2,i3] <-
sum(abs(y -  y_last_predict)^ppval3ar[i1,i2,i3]) /
(n - (dim(x)[2]+1));
sig_eps3ar[i1,i2,i3] <-
(sum(abs(y -  y_last_predict)^ppval3ar[i1,i2,i3]) /
(n - (dim(x)[2]+1)))^(1/ppval3ar[i1,i2,i3]);
}
}
}
```

Random number generation for the design

The `normalp` package is used to generate random number. The estimated values for location, scale and peakedness parameters are plug into the function `rnormp` given by:

```
rnormp(n, mu, sigma, p , method = c("def", "chiodi"))
```

Note that `mu`, `sigma` and `p` are parameters which are estimated by using the empirical distribution produced by the regressional form.

Acknowledgments

We appreciate the editorial board's and reviewers' valuable comments on the paper.

Availability of data and material

Not applicable.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

LITERATURE CITED

Alzer, H. and A. Z. Grinshpan, 2007 Inequalities for the gamma and q-gamma functions. Journal of Approximation Theory **144**: 67–83.

Arslan, O. and A. I. Genç, 2009 The skew generalized t distribution as the scale mixture of a skew exponential power distribution and its applications in robust estimation. Statistics **43**: 481–498.

Bala, S. K. and P. K. Biswas, 2005 Tax-ombudsman in bangladesh: an analytical review of the regulatory framework. Cost and Management **33**: 27–40.

Balakrishnan, N. and V. B. Nevzorov, 2004 *A primer on statistical distributions*. John Wiley & Sons.

Çankaya, M. N., 2018 Asymmetric bimodal exponential power distribution on the real line. Entropy **20**: 23.

Çankaya, M. N. and O. Arslan, 2020 On the robustness properties for maximum likelihood estimators of parameters in exponential power and generalized t distributions. Communications in Statistics-Theory and Methods **49**: 607–630.

Çankaya, M. N., A. Yalçınkaya, Ö. Altındağ, and O. Arslan, 2019 On the robustness of an epsilon skew extension for burr iii distribution on the real line. Computational Statistics **34**: 1247–1273.

Çankaya, M. N., 2021 Derivatives by ratio principle for q-sets on the time scale calculus. Fractals **29**: 2140040.

Coles, S., J. Bawa, L. Trenner, and P. Dorazio, 2001 *An introduction to statistical modeling of extreme values*, volume 208. Springer.

De Gregorio, J., D. Sanchez, and R. Toral, 2023 Entropy estimators for markovian sequences: A comparative analysis. arXiv preprint arXiv:2310.07547 .

Haberman, S. J., 1989 Concavity and estimation. The Annals of Statistics pp. 1631–1661.

Härdle, W., M. Müller, S. Sperlich, A. Werwatz, *et al.*, 2004 *Nonparametric and semiparametric models*, volume 1. Springer.

Hunter, D. R., 2023 Unsupervised clustering using nonparametric finite mixture models. Wiley Interdisciplinary Reviews: Computational Statistics p. e1632.

Iacus, S. M. *et al.*, 2008 *Simulation and inference for stochastic differential equations: with R examples*, volume 486. Springer.

Jenkins, S. P., 2017 Pareto models, top incomes and recent trends in uk income inequality. Economica **84**: 261–289.

Lehmann, E. L. and G. Casella, 2006 *Theory of point estimation*. Springer Science & Business Media.

Mineo, A. and M. Ruggieri, 2005 A software tool for the exponential power distribution: The normalp package. Journal of Statistical Software **12**: 1–24.

Mokhtari, F., R. Rouane, S. Rahmani, and M. Rachdi, 2022 Consistency results of the m-regression function estimator for stationary continuous-time and ergodic data. Stat **11**: e484.

Montgomery, D. C., E. A. Peck, and G. G. Vining, 2021 *Introduction to linear regression analysis*. John Wiley & Sons.

Serrano, F., 2007 The taxpayer's rights and the role of the tax ombudsman: an analysis from a spanish and comparative law perspective. Intertax **35**.

Sicuro, G., P. Tempesta, A. Rodríguez, and C. Tsallis, 2015 On the robustness of the q-gaussian family. Annals of Physics **363**: 316–336.

Stanimirović, I., 2017 *Computation of generalized matrix inverses and applications*. CRC Press.

Vila, R., L. Alfaia, A. F. Menezes, M. N. Çankaya, and M. Bourguignon, 2022 A model for bimodal rates and proportions. Journal of Applied Statistics pp. 1–18.

Vila, R., L. Ferreira, H. Saulo, F. Prataviera, and E. Ortega, 2020 A bimodal gamma distribution: properties, regression model and applications. Statistics **54**: 469–493.

How to cite this article: Çankaya, M. N., and Aydın, M. Future Prediction for Tax Complaints to Turkish Ombudsman by Models from Polynomial Regression and Parametric Distribution *Chaos Theory and Applications*, 6(1), 63-72, 2024.