6 Comparison of Statistical Methodologies for Philanthropic Studies

Marta Pittavino

1 Introduction

Philanthropic data related to charitable donations is often not homogeneously distributed. Outliers, corresponding to small and large donation amounts, are most certainly present. The reasons for these extreme values are various. They can be linked to both demographic and economic backgrounds and psychological personal characteristics such as the propensity to donate (Adena, 2021; Bernardic et al., 2021). Another example is provided by Ugazio in 2019 (Ugazio, 2019), where preferences for philanthropy are explained to estimate moral and financial subjective values.

While understanding why charitable donations are heterogeneous is important, a key question is how to detect and investigate this diversity. Identifying an efficient methodology that is useful for answering which variable mainly affects the donations is a powerful tool.

The key research questions addressed in this chapter are mainly two:

- What are suitable methodologies to analyze for philanthropic studies?
- How is it possible to proceed, in terms of steps, with the data analysis?

The answers to the research questions above directly impact data-driven decision-making for policy development. The philanthropic sector is subject to important decisions that affect society, and often, these decisions are based on the outcomes of analytical procedures. The results can vary depending on the technique used and the adopted methodology, influencing stakeholders differently. This chapter provides potential methodologies and critical steps for effective decision-making processes.

The aim and novelty of this work is the description and comparison, with potential pedagogical usage (rule of thumb), of suitable statistical methodologies that can be used for the detailed analysis of philanthropic data. Following this scope, this chapter will mainly focus on the methods and results rather than a presentation and investigation of data set.

This work summarizes the results from direct analysis done with philanthropic data (Lideikyte-Huber et al., 2021; Lideikyte-Huber & Pittavino, 2022; Pittavino & Lideikyte-Huber, 2024), where charitable donations were

DOI: 10.4324/9781003546986-7

represented as a form of deduction from tax returns, and the incentives for contributing to them were mainly linked to fiscal benefits (OECD, 2020).

Tax incentives for charitable giving are a prevalent feature of legal systems worldwide (OECD, 2020). The primary goal of such incentives, at least from an economic perspective, is to boost donations: it is, for instance, argued that the system that grants tax incentives for charitable donations increases transparency in the philanthropic sector (Brakman & Dean, 2023). However, legislative proposals are often vague on this point. Indicating that they want to increase charitable giving in general, they often fail to say how exactly donors' giving behavior is expected to change because of the legal standards establishing tax incentives for charitable giving or which donors these reforms intend to benefit (Lideikyte Huber et al., 2021).

This framework consists of the setup for real-world applications to show the results of the method comparison involving tax data from the Canton of Geneva spanning the years 2009–2011. The sample data set used for the models' comparison is drawn from original data from taxpayers' returns for the period 2009–2011, confidentially shared by the Tax Administration of the Canton of Geneva (TACG) for previous studies (Lideikyte Huber & Pittavino, 2022; Lideikyte Huber et al., 2021; Pittavino & Lideikyte Huber, 2024). Using a bootstrapping technique without replacement (Efron & Tibshirani, 1993), comprises a randomly drawn subsample from the original dataset comprising 100,000 observations was drawn for a specific illustrative scope.

The year 2009 has been specifically selected for the legal framework related to this, during which the threshold for donations and tax incentives was increased from 5% to 20% (Federal Act of December 14, 1990 on Direct Federal Taxation, AS 1991 1184, Lideikyte Huber et al., 2021).

Depending on the objective and the nature of the philanthropic data, several statistical methodologies can be used to analyze the data. Estimation and prediction are the two main objectives in the majority of studies, each having different focuses and specific methods to address this need to be implied. Based on this, section 2 describes the regression models with two different types of estimation techniques (classical, LM, and robust, M-estimation) and two classes of forecasting models (the ETS and ARIMA). The regression models described belong to both the standard (LM) and modern (M-estimation) methods, which are commonly used in econometrics. The philanthropic data often satisfies the underlying assumption of normality by fitting well with this type of technique. Otherwise, if the normality assumption is not satisfied, a mathematical transformation can be used to ensure it. When collecting data over time, the forecasting models help to describe the future scenarios. The specific two classes that will be illustrated have properties that include both time-dependent characteristics (ETS) and another one independent from time (ARIMA).

Section 3 includes a rule of thumb for analyzing the data, consisting of possible steps to follow. Section 4 illustrates the results of the comparison

between the previously described methods. Section 5 provides discussions and conclusions on the previously presented topics.

2 Methods

When in the philanthropic data set, there are multiple explanatory variables, which might be demographic (i.e., age, gender), economic (i.e., income, wealth), or psychological (i.e., the propensity to generosity), the objective is to understand the variables that mainly influence the donations and the outcome of interest, and describe the possible relationships between variables; the linear regression methods are mainly used. The classic least-squares (LS) and the robust M-estimation techniques are characterized in the following subsection. The regression models described belong to both the standard and modern ones and are commonly used in econometrics. The philanthropic data often satisfies the underlying assumption of normality by fitting well with this type of technique. Otherwise, if the normality assumption is not satisfied, a mathematical transformation can be used to ensure it.

2.1 The Classic Linear Regression Least Squared (LS) Method

The linear regression (Faraway, 2002, 2004, 2016) is a statistical model that estimates the linear relationship between a scalar response: Y_i (i.e., the charitable donations), representing the dependent variable, and one or more explanatory variables: X_{i1}, …, X_{ip}, (i.e., the regressors or independent variables), with $i = 1, \ldots, n$ the statistical units. When only one regressor X_{i1} is present, it is called simple linear regression, and it is mathematically represented with the model formulation below:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

with ε_i = "epsilon_i," $\varepsilon_i \sim N(0,\sigma^2)$, the resulting errors are assumed to be independent and identically distributed (iid) following a Gaussian distribution with a null mean and variance σ^2 (Pittavino et al., 2017a, 2017b). The linear regression is modeled through ε_i, representing the error variable, an unobserved random variable adding "noise" to the linear relationship.

If one or more variables are present: X_{i1}, …, X_{ip}, it is named multiple linear regression, and the model formulation is as follows, $Y_i \mid X_i \sim N(\mu_x,\sigma^2)$:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

With $\varepsilon_i \sim N(0,\sigma^2)$, independent and identically distributed (iid) errors following a Gaussian distribution with mean 0 and variance $\sigma^{2.}$ In the linear regression methodology, the conditional mean of the response given the values of the explanatory variables is assumed to be an affine function of those values,

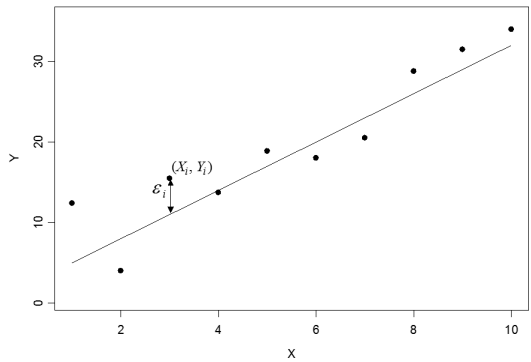

Figure 6.1 Graphical representation of the least square (LS) technique.

summarized by $Y_i \mid X_i \sim N(\mu_x,\sigma^2)$. The peculiarity of the regression analysis, rather than the multivariate analysis, is the focus on the conditional probability distribution of the response given the values of the predictors instead of the joint probability distribution.

The classical method for getting estimates from linear regression is the least squares (LS) technique, a parameter estimation method based on minimizing the sum of the squares of the errors, known as residuals, resulting from each equation. A residual ε_i is the difference between the observed value Y_i and the fitted value provided by the model: $\beta_0 + \beta_1 X_{i1}$. This mathematical procedure is illustrated in Figure 6.1, which shows "the residuals" ε_i and the intuition behind the minimization of the sum of squares. While from the mathematical point of view, the LS technique can be written as: $\min_{\beta_0\beta_1} \sum_{i=1}^{n} \varepsilon_i^2$, this equation is feasible if and only if: $\min_{\beta_0\beta_1} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_{i1}\right)^2$.

The last equation results in a system of equations called "normal equations," from which the exact closed-form or analytical expression of the LS estimates is derived. The LS estimates are efficient, meaning they have the smallest possible variance, indicating a minor deviance between the estimated value and the true value. The software used for the model implementation and estimation is the software R (*R Development Core Team, 2024*), with the *lm* R package from the base *stats* tools in R, which is useful for fitting this type of statistical model.

The linear regression with the classical LS procedure assigns the same weight to each data point, resulting in efficient estimates. For this reason, when extreme values, also known as outliers, are present, they influence the estimation procedure. In philanthropy, outliers are common, especially among generous individuals making high donations.

2.2 The Robust Linear Regression M-estimation Method

Another type of estimation technique within the linear regression analysis is also possible: robust regression analysis. In this case, the Least Median of Squares (LMS) method is used instead of the classical LS method. The LMS method can be mathematically written as: $med\,\varepsilon_i^2$, where the median of the squared errors is minimized instead of the sum (or average) of the squared errors. This minimization of the median of the squared errors assigns different weights for each data point, allowing the exclusion of strong influence by extreme values. Therefore, the LMS is not influenced by extreme values, which is one of the main differences from the classic LS estimation, ensuring the robustness of the estimator.

The drawback of this type of estimation is the lack of an analytic expression for the LMS estimate. Searching algorithms are implemented to find the convergence results, implying empirical results instead of efficient ones, and to reach this goal, a specific class of estimators, the M-estimators, is used: $\min_{\beta_0\beta_1} \sum_{i=1}^{n} \rho\left(\frac{\varepsilon_i}{\sigma}\right)$, (Huber & Ronchetti, 2009).

When there is a special case of the function ρ that coincides with the extremes values $\frac{\varepsilon_i}{\sigma}$: $\rho\left(\frac{\varepsilon_i}{\sigma}\right) = \left(\frac{\varepsilon_i}{\sigma}\right)^2$, then the result is exactly the LS estimate, providing an efficient result.

Overall, an M-estimator is robust if the function ρ limits the extreme values $\frac{\varepsilon_i}{\sigma}$ (criteria).

The function Tukey's bisquare (or biweight) is one possible example of a function satisfying the previous limitation criteria and is often used:

$$\rho_c\left(\frac{\varepsilon_i}{\sigma}\right) = \begin{cases} \frac{6}{c}\left[\left(\frac{\varepsilon_i}{c\sigma}\right)^6 - 3\left(\frac{\varepsilon_i}{c\sigma}\right)^2 + 3\left(\frac{\varepsilon_i}{c\sigma}\right)^2\right] & \text{if } \left|\frac{\varepsilon_i}{\sigma}\right| < c \\ \frac{6}{c} & \text{if } \left|\frac{\varepsilon_i}{\sigma}\right| \geq c, \end{cases}$$

The robust M-estimation regression method is implemented using the robust R package (*Wang et al., 2023*), and it can include both options, with and without interaction, to compare the findings from the classical LS estimation technique.

2.3 The Forecasting Error-Trend-Seasonality (ETS) Models

When the aims of philanthropic studies are projections and predictions of future figures (i.e., donations and donors), appropriate forecasting methods must be applied to achieve the objective.

This subsection describes the class of methods called Error-Trend Seasonality (ETS). When collecting data over time, the forecasting models help to describe the future scenario, and the specific two classes that will be illustrated have properties that include both time-dependent characteristics (ETS) and another independent of time (ARIMA).

The Error-Trend-Seasonality models (Hyndman & Athanopoulos, 2021), which from now on will be simply called ETS models, represent an evolution of the standard Moving Average (MA) estimation technique for the trend component of the time series, with the inclusion of specific weights that base and update the forecast not only on the last observations but also by taking into account the time series history; hence, the weighted average origin. The ETS models are used when the whole time series, representing the time span of the philanthropic data, is available, and the model's performance is based on the time length already collected and the future time window to forecast. This type of model takes its name from the specific part of the time series they are predicting.

Generally, the philanthropic time-series data are decomposed into three habitual components (E, T, S):

- The Error (E) indicates the difference between the true and the estimated values;
- The Trend (T) represents the increasing or decreasing pattern of the data;
- The Seasonality (S) shows the presence of higher or lower changes in the data behavior.

As the direct name suggests, in this methodology, each time-series component mentioned above is singularly estimated using either an additive or a multiplicative decomposition, resulting in a total of 30 possible model combinations, fully illustrated in Figures 6.2 and 6.3.

ADDITIVE ERROR MODELS

Trend	Seasonal		
	N	**A**	**M**
N	$y_t = \ell_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$	$y_t = \ell_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/\ell_{t-1}$
A	$y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + b_{t-1})$
$\mathbf{A_d}$	$y_t = \ell_{t-1} + \phi b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = \phi b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + \phi b_{t-1})$
M	$y_t = \ell_{t-1}b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t/\ell_{t-1}$	$y_t = \ell_{t-1}b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}b_{t-1}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b_{t-1})$
$\mathbf{M_d}$	$y_t = \ell_{t-1}b_{t-1}^{\phi} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/\ell_{t-1}$	$y_t = \ell_{t-1}b_{t-1}^{\phi} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}b_{t-1}^{\phi}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b_{t-1}^{\phi})$

Figure 6.2 The summary table for the 15 additive error models for the error-trend-seasonality model.

MULTIPLICATIVE ERROR MODELS

Trend	Seasonal		
	N	A	M
N	$y_t = \ell_{t-1}(1+\varepsilon_t)$ $\ell_t = \ell_{t-1}(1+\alpha\varepsilon_t)$	$y_t = (\ell_{t-1} + s_{t-m})(1+\varepsilon_t)$ $\ell_t = \ell_{t-1} + \alpha(\ell_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m}(1+\varepsilon_t)$ $\ell_t = \ell_{t-1}(1+\alpha\varepsilon_t)$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$
A	$y_t = (\ell_{t-1} + b_{t-1})(1+\varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1+\alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1} + s_{t-m})(1+\varepsilon_t)$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m}(1+\varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1+\alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$
A_d	$y_t = (\ell_{t-1} + \phi b_{t-1})(1+\varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1+\alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1} + s_{t-m})(1+\varepsilon_t)$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}(1+\varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1+\alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$
M	$y_t = \ell_{t-1}b_{t-1}(1+\varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}(1+\alpha\varepsilon_t)$ $b_t = b_{t-1}(1+\beta\varepsilon_t)$	$y_t = (\ell_{t-1}b_{t-1} + s_{t-m})(1+\varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}b_{t-1}s_{t-m}(1+\varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}(1+\alpha\varepsilon_t)$ $b_t = b_{t-1}(1+\beta\varepsilon_t)$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$
M_d	$y_t = \ell_{t-1}b_{t-1}^{\phi}(1+\varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi}(1+\alpha\varepsilon_t)$ $b_t = b_{t-1}^{\phi}(1+\beta\varepsilon_t)$	$y_t = (\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})(1+\varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}b_{t-1}^{\phi}s_{t-m}(1+\varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi}(1+\alpha\varepsilon_t)$ $b_t = b_{t-1}^{\phi}(1+\beta\varepsilon_t)$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$

Figure 6.3 Summary table for the 15 multiplicative error models for the error-trend-seasonality model.

Figure 6.2 shows the 15 possible combinations resulting from the inclusion of an additive error in the model, while Figure 6.3 refers to the 15 possible combinations resulting from the insertion of a multiplicative error in the model.

The Additivity (A) is chosen when the changes happen in time with a linear and slight variation, while the Multiplicativity (M) is chosen when the changes occur in time with an exponential and considerable variation. Only in the Trend case is a Damping Effect (Ad, Md) resulting from sudden changes in the data pattern for a short period possible.

The ETS models are also equivalently named Exponential Smoothing methods due to the parameter α present in each model, independently of the type of combinations with the other components. This parameter gives the rate *1-α* of the exponential decrease of the weights, decreasing in the past, hence the name of the methodology.

Since the philanthropic data, unless they are the results of specific charitable campaigns, do not have a seasonal effect, such as business products do on a particular period, only the trend component is modeled. For this reason, four different types of ETS models (from Simple Exponential Smoothing (SES) to Holt's Models), considering three different trend effects (i.e., additive, additive with damped, and multiplicative), have been fitted for the models' comparison.

In the *R* software, these types of models are implemented using the *ets* function from the *fpp2* and *fpp3* packages, which are essential tools for working with this type of model.

2.4 The Forecasting AutoRegressive Integrated Moving Average (ARIMA) Models

The last class of models we will describe is the AutoRegressive Integrated Moving Average (ARIMA) models, which represent an extension of

the AutoRegressive Moving Average (ARMA) models (Hyndman & Athanopoulos, 2021). They differ from regression models as they do not explicitly identify the link between the outcome expressed in terms of time y_t and the covariates x_t. They rather use linear models to express the link between y_t and its history y_{t-1}.

When working with ARIMA models, the concept of stationarity has to be introduced. The stationary time series has a constant pattern over time, with unpredictable patterns in the short term. In particular, the main moments do not depend on time t.

Formally, the time series y_t is stationary if:

- The Expected Value of y_t does not depend on t;
- The Variance of y_t: Var(yt) does not depend on t;
- The Covariance between y_t and the next *m* point in time y_{t+m}: Cov(yt, yt+m) does not depend on t (only on the *lag m*).

If the time series y_t is not stationary, a way to reach this property is to work with the differentiated series of y_t, indicated as y'_t, where $y'_t = y_t - y_{t-1}$. This differentiation part takes the name of the ARIMA models as an additional characteristic of the ARMA models.

The entire and compact model formulation of the ARIMA model can be found in the mathematical equation below:

$$y'_t = c + \phi_1 y'_{t-1} + \ldots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} 1 + \ \ldots \ + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

The ϕ parameters represent the AutoRegressive (AR) model part, where autoregression states that the current observation y_t can be predicted by a linear combination of the past observations y_{t-p}. At the same time, the θ parameters represent the Moving Average (MA) model part, where the current observation y_t can be predicted by a linear combination of the past forecast errors ε_{t-q}.

The software R implements them through the arima or autoarima function present in the package fpp2.

3 The Data with a Rule of Thumb and Steps for Analysing the Data

This section describes the dataset and the sampling technique, along with the procedure and important steps to follow that are useful for analyzing philanthropic data.

The sample data set used for the models' comparison is drawn from original data from taxpayers' returns for the period 2009–2011, confidentially shared by the Tax Administration of the Canton of Geneva (TACG) for previous studies (Lideikyte-Huber et al., 2021; Lideikyte-Huber & Pittavino, 2022; Pittavino & Lideikyte-Huber, 2024). The selected variables provide information on the entire population of taxpayers in the Canton of Geneva;

however, for the following illustrative example, a smaller subsample of 100,000 observations has been drawn.

A different dataset was provided for each year under study. An entire description of the dataset is provided in Lideikyte-Huber et al. (2021); the selected six variables particularly used in the present study are described and listed below with their original names provided in brackets:

- "Coded ID" (`"identifiant"`): a coded ID for each taxpayer. This variable allows tracking the same taxpayer over time. Each fiscal year, the same coded ID is used for a given taxpayer;
- "Global net taxable income" (`"revenu_net_imposable_taux"`): the net taxable income for cantonal tax purposes (after all deductions) applied to set the tax rate; this includes the totality of any foreign income;
- "Net taxable income in Geneva" (`"revenu_net_imposable_GE"`): the net taxable income in the canton of Geneva. In 2010 and 2011, the canton of Geneva introduced several changes to its income tax law (e.g., the extension of the deduction for family expenses). To a certain extent, those changes influenced the definition of taxable income for cantonal tax purposes;
- "Gross wealth" (`"fortune_brute"`): global wealth of the taxpayer;
- "Fortune_imposable" (`"fortune_imposable"`): taxable wealth;
- "Deductions for donations" (`"versements_benevoles"`): the amount of deduction (if any) for charitable giving admitted for cantonal tax purposes.

The sampling technique used is the bootstrap sampling technique (Efron, 1993). Bootstrapping is a procedure for estimating the distribution of an estimator by sampling, without replacement in this specific case, one's data or a model estimated from the data. This technique allows the estimation of the sampling distribution of almost any statistic using random sampling methods.

The procedure and important steps for analyzing philanthropic data are explained from this point on.

First, independent of the study's objectives, the data set has to be analyzed with an exploratory data analysis (EDA). The main summary statistics (i.e., mean, sd, min, max, median, length, and missing values) must be computed; this is part of the exploratory stage of the data analysis. This initial step helps to identify some features connecting the data.

Afterward, the Pearson linear correlation coefficients pxy have to be calculated between all the pairs of variables. This coefficient, giving values between –1 and 1, is essential for understanding if two variables are highly correlated, which might lead to multicollinearity issues. This is part of the confirmatory stage of the process.

The linear correlation coefficients take the following mathematical expression:

$$pxy = \frac{covariance(X,Y)}{\sqrt{var(X)var(Y)}} - 1 \leq pxy \leq 1.$$

It is estimated through the Pearson moment correlation coefficients:

$$r_{X,Y} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}},$$

with each of the fraction components taking the expressions below:

$$S_{XY} = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right),\ S_{XX} = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \text{ and } S_{YY} = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

Pearson's correlation coefficient measures the strength of a linear association between two variables and determines whether there is a linear component of association between two continuous variables. From this computation, if the resulting Pearson's correlation is close to the extreme value (–1 or 1) by showing very high multicollinearity between two pairs of variables, it is necessary to proceed with a further check in order to measure the amount of variance explained by each one of them for the resulting model. The statistical measure allowing this computation is the Variance Inflation Factor (VIF). The VIF is applied to select variables for variable *j* and with R^2_j: the model coefficient of determination, where *Xj* is regressed by: $VIF_j = \frac{1}{1-R_j^2}$.

Once the VIF gives the results, only the variables with a value lower than five are kept; all the others indicate high multicollinearity issues and will, therefore, be deleted from the analysis.

The philanthropic data is now ready to be analyzed with the classical LS estimation and/or the robust M-estimation technique if the scope is the identification of variables influencing charitable donations. Otherwise, if the primary goal is predicting future values, either ETS or ARIMA models can be implied for this case.

Following what is described above, Table 6.1 reports the steps in a synthetic, helpful way for the data analysis.

4 Results

4.1 Comparison between the Linear Regression Estimation Techniques

The two linear regression techniques are compared to understand the most suitable models for charitable giving, both from a theoretical and applied point of view. A simulated dataset is used for the application. A randomly

Table 6.1 Illustrating the steps for the data analysis and the rule of thumb for the procedure

Steps for the data analysis	*Rule of thumb for the procedure*
1	An exploratory data analysis (EDA) should be performed, with the computation of the main summary statistics (e.g., mean, sd, min, max, median, length, and missing values).
2	The Pearson's correlation coefficient has to be computed between all pairs of variables.
3	The presence of potential multicollinearity has to be identified.
4	Following 3., the Variance Inflation Factor (VIF) has to be applied to select variables. For a variable *j* and with R^2j: the model coefficient of determination, where *Xj* is regressed.
5	Among all the initial variables, only the ones having a VIF lower than 5 will be kept.
6	Based on the results from 5 and if the purpose is the estimation and identification of the main variables influencing charitable donations, the classical LS estimation linear regression analysis is performed.
7	If the number of outliers and extreme values is particularly high, the robust M-estimation linear regression model is also fitted for completion and as sensitivity analysis.
8	Comparison between 6. and 7. to identify the best fit, model, and significant variable(s) is performed. Based on the outcome of the estimates from model 6 or model 7 and from the one minimizing R^2j: the model coefficient of determination, the best model is chosen.
9	Always based on the results from 5 and if the purpose is the forecasting of future model behaviors, the ETS models and ARIMA models are performed.
10	Comparison between the two models implemented in point 9 is performed to identify the best fit model and future forecasts. The model comparison is done based on the error metrics that will produce the lower results, and depending on these, the best model is chosen.

drawn subsample from tax data from the Canton of Geneva spanning 2009–2011 was selected for a merely specific illustrative scope.

Table 6.2 is a summary table illustrating the main characteristics of the two linear regression estimation techniques; on the left, the main features for the classic LS estimation are listed, while on the right, the ones for the robust M-estimation technique are present. The last line indicates the name of the R package (R Development Core Team, 2024), which is suggested to be used for the model implementation and the R library where it can be found.

Figure 6.4 shows the results from the comparison between the weights from the classic LS estimation and the robust M-estimation. As Table 6.1 already mentions, while the weights in the classic LS estimation are the same for each data point, in the robust M-estimation, the data with the highest leverage, which deviates from the standard data pattern, reach a null weight.

Table 6.2 Summary of the main characteristics of the two linear regression estimation techniques

Classic LS estimation	*Robust M-estimation*
Efficient exact method	Empirical, not efficient method
Not robust estimates	Robust and convergent
Influenced by outliers	Not influenced by outliers
Same weights for each data point (Figure 6.4)	Different weights for each data point (Figure 6.4)
In R: lm function in the stats library	*In R: lmRob function in the robust library*

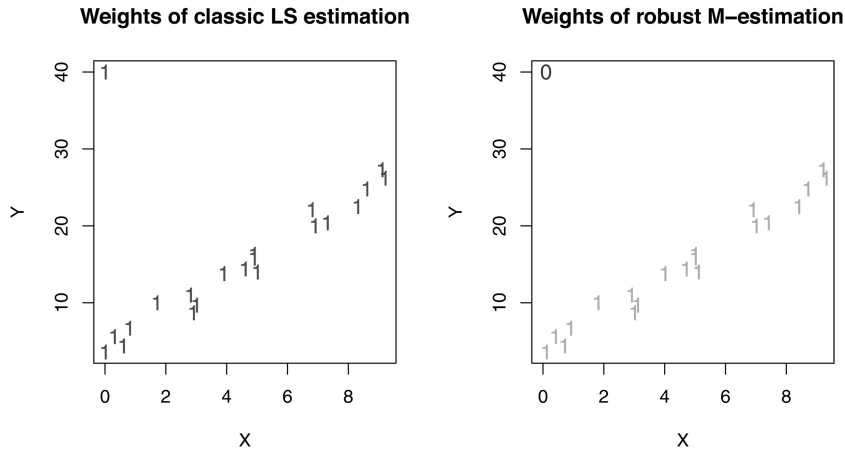

Figure 6.4 Results from the comparison between the weights from the classic LS estimation and the robust M-estimation.

Table 6.3 shows the estimates from the bivariable linear regression model. The two variables are income and wealth, and the resulting adjusted model coefficient of determination R^2_{Adj} is consistently low for the robust M-estimation technique. This suggests that when outliers are present, this is the most suitable linear regression methodology.

Table 6.3 shows the beta estimates: β and the p-values: p for the variables Inc (Income) and Wth (Wealth), with the model adjusted R^2: R^2_{Adj} resulting from the standard bivariable linear regression model, implemented both in the classic LS estimation and the robust M-estimation case.

4.2 Comparison between the Forecasting Models

In the forecasting context, when the model comparison is performed, error metrics are used for model selection instead of model estimates. Following this, five different error metrics: AIC, AICc, BIC, RMSE, and MAPE

Table 6.3 Estimates from the bivariable linear regression model

	Classic LS estimation regression models results					*Robust M-estimation regression models results*				
	β*Inc*	*pInc*	β*Wth*	*pWth*	*R2Adj*	β*Inc*	*pInc*	β*Wth*	*pWth*	*R2Adj*
2009	6.5×10^{-2}	<0.05	$23.3\times10^{-}$	<0.05	0.70	4.5×10^{-2}	<0.05	$13.3\times10^{-}$	<0.05	0.95
2010	6.1×10^{-2}	<0.05	$23.7\times10^{-}$	<0.05	0.76	4.1×10^{-2}	<0.05	$13.7\times10^{-}$	<0.05	0.92
2011	1.5×10^{-1}	<0.05	$23.1\times10^{-}$	<0.05	0.65	4.3×10^{-2}	<0.05	$13.1\times10^{-}$	<0.05	0.98

(Hyndman & Athanopoulos, 2021) have been calculated, and the smallest ones have been applied to select the best model for the forecast.

The first three metrics (AIC, AICc, BIC) are characterized by penalizing the model parameters and providing comparable results, while the last two (RMSE, MAPE) are based on different mathematical transformations (root and absolute percentage) of the related model errors/residuals.

The principle for all five is the same: the lowest value represents the best model.

Using simulations on a reduced sample from an original data set, the forecasts have been calculated for charitable giving projections and the number of donors' predictions. The focus is mainly on model performance based on error metrics. The results of the forecast estimates are also provided as an illustrative scope.

Five forecasting methods, including the ETS model, with several trend effects, have been implemented and compared using error metrics to identify the most suitable one for predicting the amounts of deductions with the best data fit. ARIMA models have also been fitted to the data, however, they were like Simple Exponential Smoothing (SES, an ETS model with no trend and no seasonality) by consistently producing the same output. Therefore, they have been discarded from the charitable donations forecast. Conversely, they have been incorporated into donors' forecasts.

The model that performed the best, with the lowest AIC, BIC, and MAPE error metrics, was model 4) in Table 6.4, which is related to an ETS model with an additive damped effect and a multiplicative error to incorporate the increasing nature of the donations over time. The estimated results of this forecast over the predicted period are reported in Table 6.5, where the previously mentioned model with the 95% prediction intervals is shown.

Table 6.4 compares the five forecasting models (simple exponential smoothing (SES, equivalent to ARIMA model), ETS model with and without damped effect for the trend) implemented for donations' projections and their error metrics.

No more than ten years have been forecasted because only information from the previous three years was available, and it is advisable not to go too far back in time to limit the uncertainty.

Table 6.4 Comparison of the five forecasting models

	Errors				
Forecasting models	*AIC*	*AICc*	*BIC*	*RMSE*	*MAPE*
1) SES	388.17	391.56	389.32	11'028'511	16.50
2) ETS (A,A,N)	380.38	392.36	381.35	6'458'915	10.24
3) ETS (A,Ad,N)	382.36	403.36	386.73	6'455'513	9.90
4) ETS (Z,Ad,N)	**375.96**	398.96	***380.35***	6'511'750	***9.15***
5) ETS (M,M,N)	384.56	394.52	386.51	8'064'654	12.48

Table 6.5 Projections of donations over the following ten years (2012–2021), resulting from the model ETS (M, Ad, N) with 95% prediction interval

Years	*Forecast*	*Lo 95% PI*	*Hi 95% PI*
2012	79'170'590	56'338'283	102'002'902
2013	83'438'938	59'375'665	107'502'215
2014	87'621'917	62'352'299	112'891'543
2015	91'721'242	65'269'399	118'173'084
2016	95'738'576	68'128'156	123'348'996
2017	99'675'560	70'929'737	128'421'389
2018	103'533'808	73'675'284	133'392'336
2019	107'314'890	76'365'919	138'263'865
2020	111'020'353	79'002'737	143'037'964
2021	114'651'702	81'586'824	147'716'583

For the donors' forecast, given the data pattern, which is constantly increasing over time due to the increase in the population and its inhabitants, a multiplicative error for a model with an additive damped trend would not be supported, yielding similar results as the model 3) in Table 6.4 with an additive error instead. The best model to forecast the donors' prediction was an ARIMA model double differentiated, with neither an autoregressive nor a moving average component.

Table 6.6 compares the five forecasting models (simple exponential smoothing and ETS model with and without damped effect for the trend, several errors, and ARIMA model) implemented for donors' projections and their error metrics. While in Table 6.7 are shown the projections of donors over the following ten years (2012–2021), resulting from the model 5) ARIMA (0, 2, 0) with 95% prediction intervals.

5 Discussion and Conclusion

Modeling charitable giving requires considering different econometric problems, such as heavy left censoring, right-skewed distribution, and very heterogeneous behavioral responses, that have important implications

Table 6.6 Comparison of the five forecasting models

	Errors				
Forecasting models	*AIC*	*AICc*	*BIC*	*RMSE*	*MAPE*
1) SES	212.88	214.74	211.25	3467.838	8.29
2) ETS(A,A,N)	204.04	214.80	206.99	2018.52	5.65
3) ETS(A,Ad,N)	205.56	224.66	207.03	1914.64	5.19
4) ETS(Z,Ad,N)	202.38	214.71	204.70	1964.35	3.69
5) ARIMA(0,2,0)	**161.16**	163.73	162.36	1600.37	**2.18**

Table 6.7 Projections of donors over the following ten years (2012–2021), resulting from the model ARIMA (0, 2, 0) with 95% prediction intervals

Years	*Forecasts*	*Lo 95% PI*	*Hi 95% PI*
2012	53483	50016.19	56955.81
2013	55580	47823.27	63338.73
2014	57675	44695.15	70658.85
2015	59774	40769.06	78778.94
2016	61868	36137.19	87602.81
2017	63966	30866.11	97065.89
2018	66062	25006.64	107117.36
2019	68158	18599.18	117716.82
2020	70254	11676.84	128831.16
2021	72348	4267.38	140432.62

for determining the correct estimation strategy (Bönke et al., 2013, Dean, 2023).

This chapter aimed to provide a fast and comprehensible guide for analyzing philanthropic data, starting from the principle that estimation and forecasting are the main objectives in this field. Moreover, the model comparison was conducted with an illustrative scope, moving from the theoretical to a more applied context, where simulations can be implemented.

These results have a direct impact on data-driven decision-making for policy development. The philanthropic sector is subject to important decisions that affect society, often based on the outcomes of analytical procedures. The results can vary depending on the technique used and the adopted methodology, influencing stakeholders in different ways. This chapter provides potential methodologies and crucial steps that can be applied for effective decision-making processes.

This study is based on and takes inspiration from recent previous studies on philanthropy (Lideikyte-Huber & Pittavino, 2022; Lideikyte-Huber et al., 2021; Pittavino & Lideikyte-Huber, 2024). These analyses and case studies helped set the scene and provided experience and a summary of the most commonly used statistical methodologies.

When the purpose is estimation, robust regression techniques with M-estimation prove to be the most effective for handling this type of data. Outliers prefer a robust rather than an efficient statistical methodology. Since these kinds of extreme observations are often common in philanthropic data sets, techniques that they do not influence are preferable.

Conversely, when the goal is prediction, ARIMA models, especially for constantly increasing patterns, tend to better incorporate the peculiar characteristics of the data, leading to more accurate forecasts in the future. This result is due to the flexibility of this type of model, which allows predictions without too many restrictions and impositions on the original modeling part.

References

Adena, M. (2021). Tax-price elasticity of charitable donations: Evidence from the German taxpayer panel. In H. Peter & G. Lideikyte Huber (Eds.), *The Routledge handbook of taxation and philanthropy* (pp. 219–235). http://doi.org/10.4324/9781003139201-17.

Bernardic, U., Lebreton, M., Lideikyte Huber, G., Peter, H., & Ugazio, G. (2021), Behavioural philanthropy: Harnessing behavioural sciences to design more effective tax incentives for philanthropy. In H. Peter & G. Lideikyte Huber (Eds.), *The Routledge handbook of taxation and philanthropy* (pp. 354–376). http://doi.org/10.4324/9781003139201-17.

Bönke, T., Massarrat-Mashhadi, N., & Sielaff, C. (2013). Charitable giving in the German welfare state: Fiscal incentives and crowding out. *Public Choice*, *154*(1–2), 39–58.

Brakman, R. D., & Dean, S. A. (2023). *For-profit philanthropy: Elite power and the threat of limited liability companies, donor-advised funds, and strategic corporate giving*. Oxford University Press.

Dean, S. A. (2023). *For-profit philanthropy: Elite power and the threat of limited liability companies, donor-advised funds, and strategic corporate giving*. Oxford University Press.

Duquette, N. J. (2019). Do share-of-income limits on tax-deductibility of charitable contributions affect giving?. *Economics Letters*, *174*, 1–4. https://doi.org/10.1016/j.econlet.2018.10.009

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC. ISBN 0-412-04231-2.

Faraway, J. J. (2002). *Practical regression and ANOVA using R* (Vol. 168). Taylor & Francis and University of Bath.

Faraway, J. J. (2004). *Linear models with R*. Taylor & Francis. Pub. Location New York Imprint Chapman and Hall/CRC. https://doi.org/10.4324/9780203507278

Faraway, J. J. (2016). *Extending the linear model with R, generalized linear, mixed effects and nonparametric regression models*. Taylor & Francis. Pub. Location New York Imprint Chapman and Hall/CRC. https://doi.org/10.1201/9781315382722.

Federal Act of 14 December 1990 on Direct Federal Taxation (DFTA), AS 1991 1184.

Hyndman, R. J., & Athanopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts.

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). Wiley.

Lideikyte-Huber, G., & Pittavino, M. (2022). Who donates and how? New evidence on the tax incentives in the canton of Geneva, Switzerland. *Journal of Empirical Legal Studies (JELS)*, *19*(3), 758–797. http://doi.org/10.1111/jels.12322.

Lideikyte Huber, G., Pittavino, M., & Peter, H. (2021). Tax incentives for charitable giving: Evidence from the canton of Geneva, Switzerland. In H. Peter & G. Lideikyte Huber (Eds.), *The Routledge handbook of taxation and philanthropy* (pp. 253–267). http://doi.org/10.4324/9781003139201-17.

OCDE. (2020). *Taxation and philanthropy* (OECD Tax Policy Studies, No. 27). OCDE Publishing. https://doi.org/10.1787/df434a77-en, 2020.

Pittavino, M., Dreyfus, A., Heuer, C., Benschop, J., Wilson, P., Collins-Emerson, J., Torgerson, P. R., & Furrer, R. (2017a). Comparison between generalized linear modelling and additive Bayesian network; identification of factors associated with the incidence of antibodies against Leptospira interrogans sv Pomona in meat workers in New Zealand. *Acta Tropica*, *173*, 191–199. https://doi.org/10.1016/j.actatropica.2017.04.034.

Pittavino, M., Dreyfus, A., Heuer, C., Benschop, J., Wilson, P., Collins-Emerson, J., Torgerson, P. R., & Furrer, R. (2017b). Data on Leptospira interrogans sv Pomona infection in meat workers in New Zealand. *Data in Brief*, *13*, 587–596. https://doi.org/10.1016/j.dib.2017.05.053.

Pittavino, M., & Lideikyte-Huber, G. (2024). *Forecasting Geneva's donors and their charitable deductions*, Methodological and Applied Statistics and Demography IV, edited by Pollice, A. and Mariani. 435–441. ISBN: 9783031644467.

R Development Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org.

Ugazio, G. (2019). Personalized philanthropy. *Expert Focus*, *3*, 121–124.

Wang, A. J., Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Maronna, R., Zivot, E., Rocke, D., Martin, D., Maechler, M., & Konis, K. (2023). *Methods for robust statistics, a state of the art in the early 2000s, notably for robust regression and robust multivariate analysis*. https://cran.r project.org/web/packages/robust/robust.pdf.