Electronics, Communications and Networks
A.J. Tallón-Ballesteros et al. (Eds.)

doi:10.3233/FAIA231184

First-Order Linear Ordinary Differential Equation for Regression Modelling

Sie Long KEK [a,1], Chuei Yee CHEN [b] and Sze Qi CHAN [c]
[a] *Dept. of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia*
[b] *Dept. of Mathematics and Statistics, Universiti Putra Malaysia*
[c] *Business School, University of Southampton Malaysia*

Abstract. This paper discusses the data-driven regression modelling using first-order linear ordinary differential equation (ODE). First, we consider a set of actual data and calculate the numerical derivative. Then, a general equation for the first-order linear ODE is introduced. There are two parameters, namely the regression parameters, in the equation, and their value will be determined in regression modelling. After this, a loss function is defined as the sum of squared errors to minimize the differences between estimated and actual data. A set of necessary conditions is derived, and the regression parameters are analytically determined. Based on these optimal parameter estimates, the solution of the first-order linear ODE, which matches the actual data trend, shall be obtained. Finally, two financial examples, the sales volume of Proton cars and the housing index, are illustrated. Simulation results show that an appropriate first-order ODE model for these examples can be suggested. From our study, the practicality of using the first-order linear ODE for regression modelling is significantly demonstrated.

Keywords. first-order linear ODE, regression modelling, loss function, parameter estimates, numerical solution

1. Introduction

A regression model shows the linear or nonlinear relationship between a response variable and a sequence of explanatory variables [1, 2]. Regression analysis is a statistical technique used to study the changes in predictive variables based on available explanatory variables. So, raw data from experimental practices and activities in business and finance are mainly employed in forecasting and decision-making. Thus, applications of regression modelling have been widely investigated, for example, in healthcare [3], production [4] and finance [5]. In regression modelling, a mathematical model is constructed by observing the trend of the response variable. Most of these models are algebraic functions, including exponential, polynomial, and rational functions [6], but their derivative functions are not included. Using these functions to study the relationship and changes in the predictive variable concerning explanatory variables is straightforward. However, the input-output relationship involving estimating parameters in an algebraic function is a challenging part of regression.

[1]Corresponding Author: Sie Long Kek, Department of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia, Pagoh Campus, 84600 Panchor, Johor, Malaysia; E-mail: slkek@uthm.edu.my

The difficulty of regression modelling increases when using an ordinary differential equation (ODE) to construct an appropriate mathematical model for actual data. This is true since solving ODEs requires an efficient numerical solver for approximating the exact solution. Regression modelling with differential equations is not a new insight, and it can be backdated to the final report [7] in 1976 on regression with differential equation models. Since machine learning has been actively developed recently, using differential equations as a regression model meets machine learning's goal in learning. In the current works, Michele [8] established ODE models from a nonlinear regression in a polynomial form using raw data from biological systems. Khoo et al. [9] applied the neural ODEs to study the regression of macroeconomic data for the Green Solow model. In addition, García [10] studied the continuous time model of dynamical systems as solutions of differential equations from non-uniformly sampled or noisy observation through machine learning techniques. Quentin et al. [11] proposed a method for parameter estimation in nonlinear mixed effects models based on ODEs, aiming to regularize the estimation problem.

In our study, a first-order linear ODE used in regression modelling is explored. The novelty of our work is to propose an efficient computational algorithm for fitting actual data using a first-order linear ODE [12], in turn, to give an appropriate ODE model for actual data. For this purpose, the numerical derivative of actual data is calculated and stored for parameter estimation. A loss function is defined to minimize the differences between the estimates and actual data. By deriving the necessary conditions [13, 14], the regression parameters are analytically determined. With these optimal parameter estimates, the first-order linear ODE model for the actual data is formulated and the analytical solution of the ODE model [15] shall match the trend of actual data. For illustration, the actual data for the sales volume of Proton cars in Malaysia and Malaysia's housing index [16, 17] are studied. Simulation results show the effectiveness of using a first-order linear ODE model for regression modelling.

The paper is organized as follows. In Section 2, a regression problem that involves a first-order linear ODE is described. In Section 3, the methodology of the estimation of parameters and finding the solution of the linear ODE are discussed. In Section 4, two illustrations of financial data for data-driven regression modelling are provided. Finally, a concluding remark is made.

2. Problem Description

Consider a set of actual data, given by

$$y = \{y_1, y_2, \cdots, y_n\}, \tag{1}$$

for time $t_i, i = 1, 2, \cdots, n$, where n is the number of data points, and the derivative of data is expressed by

$$y_i' = \frac{y_{i+1} - y_i}{t_{i+1} - t_i}, i = 1, 2, \cdots, n-1. \tag{2}$$

Define a loss function [6]

$$J_{se}(\alpha, \beta) = \sum_{i=1}^{n} (y_i' - \alpha - \beta y_i)^2 \tag{3}$$

with a first-order linear ODE [12]

$$\frac{dy}{dt} = \alpha + \beta y, \tag{4}$$

where α and β are the unknown parameters, and y is the solution set point for the first-order linear ODE in Eq. (4). Therefore, this problem is referred to as a regression problem with a first-order linear ODE, and the parameters α and β are known as the regression parameters to be determined later.

3. Solution Method

Now, consider the gradients of the loss function in Eq. (3) [14],

$$\frac{\partial J_{se}}{\partial \alpha} = -2\sum_{i=1}^{n-1}(y_i' - \alpha - \beta y_i), \tag{5}$$

$$\frac{\partial J_{se}}{\partial \beta} = -2\sum_{i=1}^{n-1}(y_i)(y_i' - \alpha - \beta y_i). \tag{6}$$

Setting these gradients to zero and doing some algebraic manipulations to gain

$$\hat{\alpha} = \bar{y}' - \hat{\beta}\bar{y}, \tag{7}$$

$$\hat{\beta} = \left(\sum_{i=0}^{n}(y_i)(y_i' - \bar{y}')\right) \times \left(\sum_{i=0}^{n}(y_i)(y_i - \bar{y})\right)^{-1}. \tag{8}$$

Substitute $\hat{\alpha}$ and $\hat{\beta}$ into Eq. (4), we have

$$\frac{dy}{dt} = \hat{\alpha} + \hat{\beta} y. \tag{9}$$

Notice that the analytical solution [12] for Eq. (9) is presented by

$$\hat{y}(t) = y(\tau)e^{\hat{\beta}(t-\tau)} - \frac{\hat{\alpha}}{\hat{\beta}}(1 - e^{\hat{\beta}(t-\tau)}), \tag{10}$$

for $t > \tau$, where $\hat{y}$ is the estimated point to the actual data point in Eq. (1). Thus, Eq. (10) represents the regression model for the actual data in Eq. (1), which closely tracks the actual data trend.

4. Simulation Examples

For illustration, we consider two sets of actual data, which are sales volume of Proton cars in Malaysia and Malaysia's housing index [16, 17]. The first-order linear ODE model for these actual data is constructed using the computational method discussed in Section 3. The GNU Octave scientific computing tool is applied to obtain simulation results.

Table 1. Sales volume of Proton cars in Malaysia from 2013 to 2022 (in 1,000 units).

Year	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Vol.	138.75	115.78	102.18	72.29	70.99	64.74	100.18	108.52	114.71	141.43

4.1. Sales Volume of Proton Cars in Malaysia

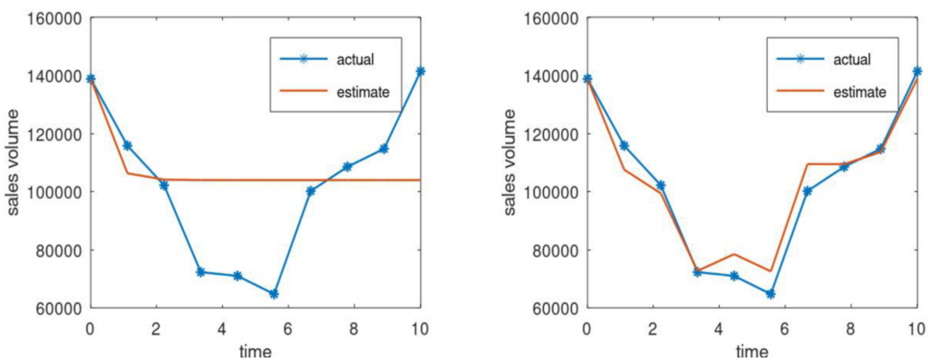
Consider the actual data for sales volume of Proton cars in Malaysia from 2013 to 2022 (in 1,000 units) as presented in Table 1 [16]. From the simulation results, the first-order linear ODE model for the sales volume of Proton cars is

$$\frac{dy}{dt} = 252{,}702.6501 - 2.431y. \tag{11}$$

The solution for this linear ODE model in Eq. (11) is

$$\hat{y}(t) = y(\tau)e^{-2.431(t-\tau)} + \frac{252{,}702.6501}{2.431}(1 - e^{-2.431(t-\tau)}), \quad t > \tau. \tag{12}$$

Figure 1 shows the sales volume and the solution of the first-order linear ODE model. Their values are apparently different before using the regression procedure. The blue line with asterisks (*) is the actual data curve, and the red is the solution curve for the first-order linear ODE model, which is labeled as the estimate curve. Figure 1 (b) demonstrates a satisfactory curve-fitting result, where the estimated solution follows the actual data trend. The model has a MSE value of 1.2738×10^{10} and the output solution gives a MSE value of 1.4917×10^{7}. These MSE values are large because the sales volume of Proton cars are in the large amount.



(a) Before regression procedure. (b) After regression procedure.

Figure 1. Estimated and actual data for sales volume of Proton cars.

4.2. Malaysia Housing Index

Consider the Malaysia's housing index from 2021 Quarter 1 (Q1) to 2023 Quarter 3 (Q3) as given in Table 2 [17]. Simulation results show that the first-order linear ODE model for the Malaysia's housing index is

Table 2. Malaysia housing index from 2021 to 2023.

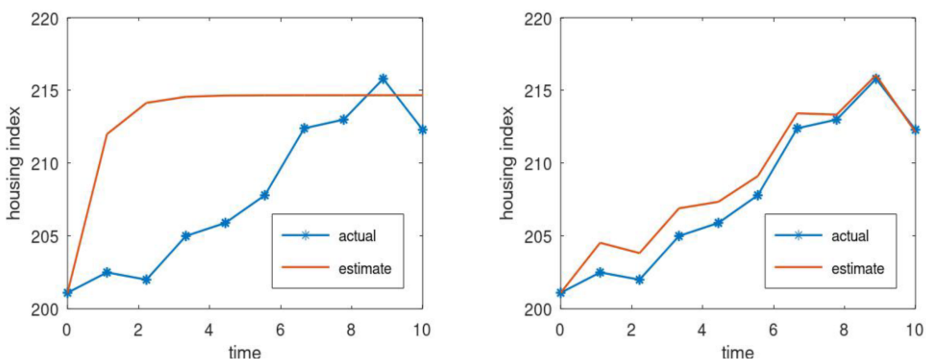
Year	2021				2022				2023	
Quarter	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2
Index	201.1	202.5	202.0	205.0	205.9	207.8	212.4	213.0	215.8	212.3

$$\frac{dy}{dt} = 314.012 - 1.463y, \tag{13}$$

and the solution for the model is

$$\hat{y}(t) = y(\tau)e^{-1.463(t-\tau)} + \frac{314.012}{1.463}(1 - e^{-1.463(t-\tau)}). \tag{14}$$

Figure 2 shows the actual data and estimated solution before and after using the regression procedure. The blue line with asterisks (*) is the actual data curve, and the red is the exponential curve for the estimated solution. Figure 2(b) shows the regression result for the linear ODE model used in fitting the actual data curve, where the estimated solutions provide a best-fit outcome to track the actual data trend. The MSE for the linear ODE model is 153.62, and the MSE for the output model is 0.6089.



(a) Before regression procedure. (b) After regression procedure.

Figure 2. Estimated and actual data for housing index.

5. Concluding Remark

Applying the regression procedure in the first-order linear ODE to fit the actual data points was discussed in this paper. The aim is to propose an appropriate first-order linear ODE model for the actual data. By minimizing the differences between the derivative of the actual data and the linear ODE model, parameters in the linear ODE model are optimally estimated. Hence, solving the linear ODE model with these optimal parameter estimates can give a best-fit model to the actual data points. For illustration, the sales volume of Proton cars and Malaysia's housing index were studied. Simulation results showed the accuracy and effectiveness of the computational algorithm for regression purposes. In conclusion, using the first-order linear ODE in fitting the actual data points is verified and provides a new idea for curve fitting. For future research, it is recommended to apply the nonlinear ODE model for handling the regression problem, which will point out the new insight to the regression community.

6. Acknowledgment

This research was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through Research Grant Tier 1 2022 (vot Q121).

References

[1] Christer T. Applied Regression analysis: Doing, interpretating and reporting. 1st Ed. Routledge. 2019.
[2] Rudolf JF, William JW and Ping S. Regression Analysis. 2nd Ed. Academic Press. 2006.
[3] Bzovsky S, Phillips MR, Guymer RH et al. The clinician's guide to interpreting a regression analysis. Eye. 2022; 36: 1715–1717.
[4] Phan DT, Nguyen LM, Murali P, Pham NH, Liu H and Kalagnanam JR. Regression optimization for system-level production control. American Control Conference (ACC), New Orleans, LA, USA, 2021; pp. 5023-5028.
[5] Pervez A and Ali I. Robust regression analysis in analyzing financial performance of public sector banks: A case study of India. Annals of Data Science. 2022.
[6] Anusha I. Introduction to Regression Analysis (Easy Statistics). anushabooks.com. 2020.
[7] Hunter, CD. Regression with Differential Equation Models. Final Report. DARCOM Intern Training Center Texarkana. 1976.
[8] Michele N. Forming the differential equations from simple polynomial regression models. SM Journal of Biometrics & Biostatistics, 2020, 5(4), 1-4.
[9] Khoo, ZY., Lee, K.H., Huang, Z., Bressan, S. (2021). Neural Ordinary Differential Equations for the Regression of Macroeconomics Data Under the Green Solow Model. In: Strauss, C., Kotsis, G., Tjoa, A.M., Khalil, I. (eds) Database and Expert Systems Applications. DEXA 2021. Lecture Notes in Computer Science(), vol 12923. Springer, Cham.
[10] García P. Modeling systems with machine learning based differential equations. Chaos, Solitons & Fractals, 2022; 165, 112872.
[11] Quentin C, Chloé P, Irene B, Rodolphe T, and Mélanie P. Parameter Estimation in Nonlinear Mixed Effect Models based on Ordinary Differential Equations: An Optimal Control Approach. Computational Statistics, 2023. hal-03335826v2f.
[12] William EB, Richard CD and Douglas BM. Elementary differential equations and boundary value problems, 12th Ed. Wiley. 2021.
[13] Wu Y, Wang W and Wang X. Sufficient and necessary conditions for the strong consistency of LS estimators in simple linear EV regression models. Journal of Statistical Computation and Simulation. 2023; 93(8), 1244-1262.
[14] Wang X, Yan L and Zhang Q. Research on the application of gradient descent algorithm in machine learning. International Conference on Computer Network, Electronic and Automation (ICCNEA), Xi'an, China, 2021; pp. 11-15.
[15] Al Ghafli AA, Nawaz Y, Al Salman HJ and Mansoor M. Extended Runge-Kutta scheme and neural network approach for SEIR epidemic model with convex incidence rate. Processes. 2023; 11(9):2518.
[16] Statista. Proton cars sales volume in Malaysia from 2013-2022. Retrieved from https://www.statista.com/statistics/869527/malaysia-sales-volume-proton-cars/.
[17] Trading Economics. Malaysia housing price index. Retrieved from https://tradingeconomics.com/malaysia/housing-index.