Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2021;7(4): 382-391

Original Article

Modeling the Number of COVID-19 Total Cases in Iran Using Gompertz and Logistic Growth Curves

Hossein Zamani

Department of Statistics, Faculty of Science, University of Hormozgan, Bandar Abbas, Iran.

ARTICLE INFO

ABSTRACT

Received 28.08.2021
Revised 14.09.2021
Accepted 27.10.2021
Published 15.12.2021

Key words:
Gompertz model;
Logistic model:
COVID-19;
Growth curve

Introduction: The growth curve are time dependece regression models which commonly are useful in describing the rapid growth of total cases or deaths in a pandemic situation.
Methods: The Gompertz and logistic functions are useful to describe the growth curve of a population or any time dependence variable such as metabolic rate, growth of tumors and total number of cases or deaths in a pervasive disease. The logistics family of growth curve including logistic, SSlogistic, generalized logistic and power logistic and Gompertz models were considered to describe the growth curve of total_cases_per_million (t_c_p_m) of COVID-19 in Iran during the 19-Feb-2020 to 28-May-2021. The models were fitted to data using nls function in R and the fitting accuracy was evaluated using the numerical and graphical approaches.
Results: The logistic family and Gompertz growth curve were applied to fit the total_cases_per_million of COVID-19 in Iran as the response versus the time in days as predictor variable. The training and testing RMSE criterions were considered as the numerical criterions to assess the model accuracy. The growth curve of fitted models was compared with the growth curve of observed data. Results indicated that the logistic and Gompertz models provided a better description of target variable than the alternatives.
Conclusion: As results shown, the logistic and Gompertz models provided a better description of response variable than the alternatives. Therefore, the logistic and Gompertz models are able to describe and forecast the COVID-19 variables (including total cases, death, recovered and so on) very well.

Introduction

The coronavirus was firstly observed in December 2019 in Wuhan city, Hubei province, China[1] and rapidly was spread all over the world. WHO officially named this fast infection as COVID-19 on Feb 2020. It is one of the most widespread pandemic disease in human history.[2] The global spread of Covid-19 (occurred in more than 150 countries), led to the announcement by the WHO in mid-March 2020 that COVID-19 was recognized as a global pandemic.[3] Up to July 2021, there were approximately 190 million worldwide cases

*.Corresponding Author: zamani.huni@hormozgan.ac.ir

of this disease with 4570000 deaths (WHO, 2021).

The coronavirus belongs to the Coronaviridae family, a large family of viruses that contains 3 genotypes of alpha, beta, and gamma and according to reports, 7 species of them affecting humans.[4,14,17,18] However, it is estimated 1/3 of the common cold to be caused by coronaviridae which almost all are unnamed types[26]

According to studies and reports the virus SARS-COV-2 which causes COVID-19 is transmitted directly and indirectly generally through contact, droplets and aspirated with a wide range of symptoms, mostly including fever, cough, shortness of breath and death.[19] The mean of incubation period of SARS-COV-2 has been reported 4.8 to 9 days.[20]Outbreaks of the disease have led most countries to apply quarantine conditions or to encourage social distancing and to impose restrictions on the size of the population to break the chain of infection.[21] Therefore, there is an urgent need for the scientific community to take action to assist governments in their efforts to control and prevent the transmission of the virus.[22] Studies of statistical patterns of the disease play an important role in descriptive epidemiology. Various studies have been conducted on modeling COVID-19 data and its related variables. In this regard, many authors have used mathematical and statistical models to describe variables related to COVID-19 and explore the behavior of this virus.[1] Moreover, several authors have applied time series models to forecast the confirmed and recovered cases and death rate of COVID-19,[5-7] COVID-19 pandemic[7] and transmission.[8] The copula function was also considered[2,17] to investigate dependence structure between COVID-19 variables and environment.

Further analysis has been regarded by using the binary logistic model[10-13] to predict and classify the COVID-19 data set. In addition to this, the Gompertz model was used to explore the growth curve of population variables related to this pandemic.[14-16]

The current study considers the various growth curve including logistic family (logistic, SSlogistic, generalized logistic power logistic) and Gompertz models to explore the pattern of the total_cases_per_million (t_c_p_m) variable of COVID-19 in Iran and then forecasts the future values. Total_cases_per_million is defined as the number of confirmed cases per million populations. To avoid large numerical values and coefficients, the scaled version of variable was considered which was obtained by dividing the t_c_p_m by 1000.

Methods

In present study we have used the growth curve models to explore and forecast the pattern of the t_c_p_m variable in Iran based on the dataset which daily is collected by the Ministry of Health and Medical Education of Iran and reported by Our World in Data. The data contains 465 daily reports of COVID-19 variables during the 19-Feb-2020 to 28-May-2021. Due to the importance of total number of confirmed cases of COVID-19 we have considered this variable to model its growth curve and forecasts its future values using several growth models.

The logistic family of growth curve including, logistic, SSlogistic, generalized logistic and power logistic and also the Gompertz model were applied to fit the t_c_p_m as the response variable versus the time in days as the predictor variable.

The logistics family of growth curve, are widely used in modelling COVID-19 infection.[23] These

models are useful for population studies which their growth curve is generally S-shape. Theses curves are used to describe various biological processes but recently have been applied in epidemiology for real-time prediction of diseases.[23] Different parameterization and functional forms of logistic models have been used by authors in literature. Through current study we apply various form of these models.
Growth models review

Logistic

The logistic growth curve is defined as[23]

$$f(t;\theta_1,\theta_2,\theta_3)=\frac{\theta_1}{1+e^{\theta_2+\theta_3 t}}, \quad (1)$$

where Θ_1, Θ_2 and Θ_3 are real values and t is the predictor variable. In current study, the predictor variable t is a counting variable that indicates the time in days which commonly starts at 1 continues until the end of the period. However, in this work to obtain the initial values of parameters easier, we cover the range of 0 to 464 days.

Gompertz

The Gompertz model was introduced in 1825 by the actuary Benjamin Gompertz for actuarial purposes to examine the patterns of death across the life course not only in humans but also in a wide range of other organisms. However, recently, it was considered to model infectious diseases by authors.[27] Two different types of Gompertz model were defined according to the type of location parameter.[24] In type I, a single parameter controls the time while in type II a single parameter controls the starting value of the curve. The type I Gompertz is a special case of generalized logistic model while the type II is obtained from a different re-parameterization. In present study, to avoid similarity and using various models, we have used a special case of type II Gompertz model with the functional form[24]

$$f(t;\theta_1,\theta_2,\theta_3)=\theta_1 e^{-\theta_2 e^{-\theta_3 t}}, \quad (2)$$

where Θ_1, Θ_2 and Θ_3 are real values and t is the independent variable.

SSLogistic

A special case of three parameters logistic model which is known as SSlogistic was defined and applied for modeling growth curve[25] with the following functional form

$$f(t;\theta_1,\theta_2,\theta_3)=\frac{\theta_1}{1+e^{(\theta_2-t)/\theta_3}}, \quad (3)$$

where Θ_1, Θ_2 and Θ_3 are real values and t is the predictor variable.

Generalized logistic

Generalized logistic is a four-parameters logistic that is widely used as a flexible function in modeling S-shape growth curves. This model has the functional form as:[23]

$$f(t;\xi,\theta_1,\theta_2,\theta_3,\xi)=\frac{\theta_1}{\left(1+\xi e^{(\theta_2-t)/\theta_3}\right)^{1/\xi}}, \quad (4)$$

where Θ_1, Θ_2 and Θ_3 are real values and ξ is positive. The generalized logistic model reduces to SSlogistic when $\xi=1$ and converges to the Gopmertz type I when $\xi \to 0$.
In epidemiological application Θ_1, Θ_2 and Θ_3 represent the final epidemic size, infection rate, and lag phase, respectively.[23] The parameter ξ

helps the function to be more flexible in fitting complex curve, however, extra parameters generally, results in overfitting and hence larger error rate.

Power Logistic

The power logistic is another form of generalized logistic which is obtained by using a new parameterization. The functional form of power logistic is defined as

$$f\,(t;\theta_1,\theta_2,\theta_3,\xi)=\frac{\theta_1}{\left(1+e^{\theta_2+\theta_3 t}\right)^{\xi}}, \quad (5)$$

where Θ_1, Θ_2 and Θ_3 are real and ξ is positive The power logistic is reduces to logistic model when $\xi=1$ Generally, additional parameter causes the model to be more flexible in fitting curves.

In order to explore the role of each parameter in the models, the growth curve of mentioned models was provided for different values of parameters in Figure 1.

Model accuracy

In statistics, several criterions such as mean square error (MSE), root of mean square error (RMSE), mean absolute error (MAE) and Akaike information criterion (AIC) have

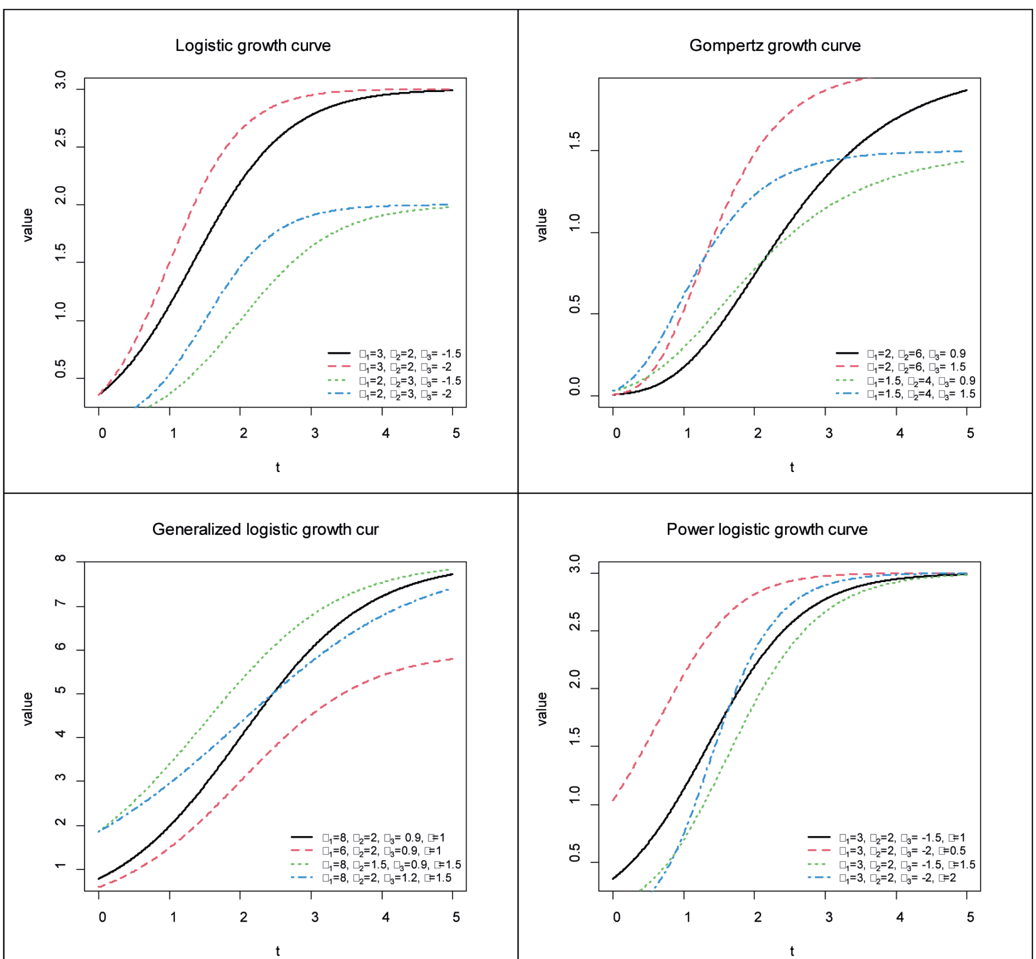

Figure 1. Models growth curve based on different values of parameters.

been defined to evaluate model accuracy. However, due to the analytical measures the MSE and RMSE are preferred. In present study, in addition to the graphical approach, we applied the training RMSE and testing RMSE as well as the AIC as the numerical criterion to evaluate model accuracy and choose the model that has the best performance in fitting t_c_p_m variable. The training data is a part of data which is used for model construction and estimating parameters and the testing data is used for assessing the fitted models. The RMSE and AIC are defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$
$$AIC = 2k + n\log(RSS) \qquad (6)$$

Where $MSE = RSS / n$ A lower RMSE and AIC indicates that the model has a better fitting and lower error rate.

Results

Several growth models of logistic family and Gompertz were applied to fit the growth curve of t_c_p_m variable of COVID-19 in Iran. To model this data, the variable t_c_p_m was considered as the response and t, the time in days, was considered as the predictor variable. Since the data was collected over a period of 465 days, commonly the value of t ranges from 1 to 465. But to obtain more easily the initial values of the parameters required for the convergence of the regression models, the value of the independent variable t was considered t = 0 for the first day (here Feb, 19, 2020) and t=464 for the last day. This is because the initial values of the parameters can be easily calculated using t = 0 and t = 1 values and therefore considering the independent variable starts from t = 0 can be helpful.

In order to converge the regression function faster and also to obtain a more accurate estimate of the parameters, it is necessary to consider appropriate initial values for the parameters because the optimization functions in R are sensitive to the initial values. Therefore, to make the initial values easier to obtain, we used a re-scaling on the independent variable t such that the value of t = 0 was considered for the first day (here Feb, 19, 2020) and t=464 for the last day. This is because initial values based on t = 0 and t = 1 will be easily obtained.

Figure 2 illustrates the growth curve of the t_c_p_m per 1000 as the response variable in Iran over the study period. As it turns out, the growth pattern is a sigmoidal (S-shaped) curve, so that non-linear regression models with sigmoidal curve such as logistics and Gompertz can be appropriate candidates for fitting this growth curves.

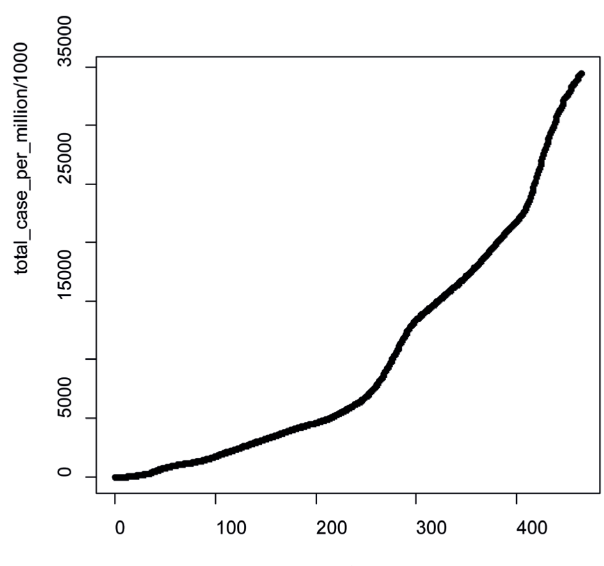

Figure 2. Growth curve of t_c_p_m variable in Iran during the study period

For fitting this growth curve, five regression models including logistic, Gompertz, SSlogistic, generalized logistics and power logistic have been considered.

The nls function in R package was used to fit each of these models (see Appendix A).

However, in order to converge the models and to estimate parameters correctly, the nls function requires appropriate initial values for the parameters. The nls function is very sensitive to the initial value of parameters and a model does not converge if the appropriate initial values of parameters are not considered. The appropriate initial values for the logistic parameters can be obtained by $\theta_2^{(0)} = \log\left(\theta_1^{(0)} / y(0) - 1\right)$ $\theta_1^{(0)} = y(\max)$ and $\theta_3^{(0)} = \log\left(\theta_1^{(0)} / y(1) - 1\right) - \theta_2^{(0)}$ where $y(\max)$ the maximum of the response variable and $y(0)$ and $y(1)$ are values of the response variable when t=0 and t=1 respectively. These initial values can also be used for the Gompertz model with some modifications. Once the models were fitted, the RMSE criterion was used as the numerical criterion to assess model accuracy. Table 1 displays the training RMSE and AIC values for each of the fitted models. As shown, the Gompertz, logistic and SSlogistic models have lower training RMSE than the alternatives in describing t_c_p_m growth curve.

Table 1. Training RMSE and AIC values of fitted models

Model	Training RMSE	Training AIC
Logistic	0.85	2710.905
Gompertz	0.76	2606.82
SSLogistic	0.86	2721.78
Generalized Logistic	17.28	5512.12
Power Logistic	21.92	5733.32

For further analysis, the growth curve of fitted models was provided versus the growth curve of observed data and illustrated in Figure 3. As shown, the logistic, SSlogistic and Gompertz models have better description of target variable than the alternatives. Despite different functional form of logistic and SSlogistc, they performed completely the same. However, for further analysis we considered the logistic model among these two due to the simpler functional form and interpretation.

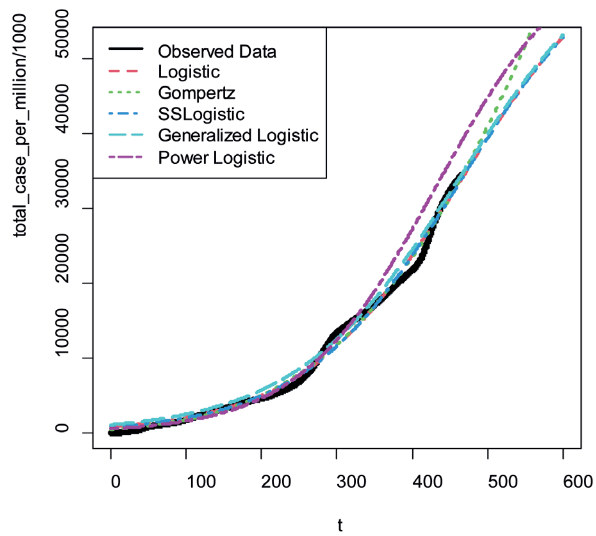

Figure 3. Graphical illustration of fitted models versus observed growth curve

Discussion

Table 1 indicates that the Gompertz model has lower training RMSE than the logistic, therefore, based on this criterion it can be concluded that the Gompertz model has better performance in fitting t_c_p_m response.

Further, by comparing the logistic and Gompertz growth curve provided in Figure 3, it can be concluded that both curves are in good agreement with the observation curve. However, the logistics curve tends to continue in the horizontal direction in forecasting future values while the Gompertz model continues to show an incremental model perpendicular to the x-axis. To provide a better description of the situation, the estimated curve and confidence interval of the logistic and Gompertz regression models are displayed in Figure 4.

As shown in Figure 4, the logistic curve grows slowly and tend to construct a sigmoidal curve while the Gompertz model grows sharper than the logistic curve. So, the logistic model has a potential to under-estimate future values while the Gompertz model has a potential to over-estimate the future values. To explore this, the testing criterions (error, RMSE) can be helpful.

Generally, both of the numerical (RMSE and AIC) and graphical (Figure 4) evaluation criteria indicate that the Gompertz and logistic growth curves described the t_c_p_m variable in Iran very well. Therefore, these models can be good candidates for modeling any pandemic infection in a regional or international scale.

The regression coefficients and p-values of significant test of the parameters were provided in Table 2. As shown, all parameters are significant at the significance level of 95% so that both models are significant and capable to predict the growth curve of observed data.

The forecasting values and RMSE of testing data (the scaled t_c_p_m for the next 10 days) are provided in Table 3 for the logistics and Gompertz models. It can be seen that the forecasted values of logistic model grow slowly compared to the Gompertz model. However, the Gompertz model estimated the observed

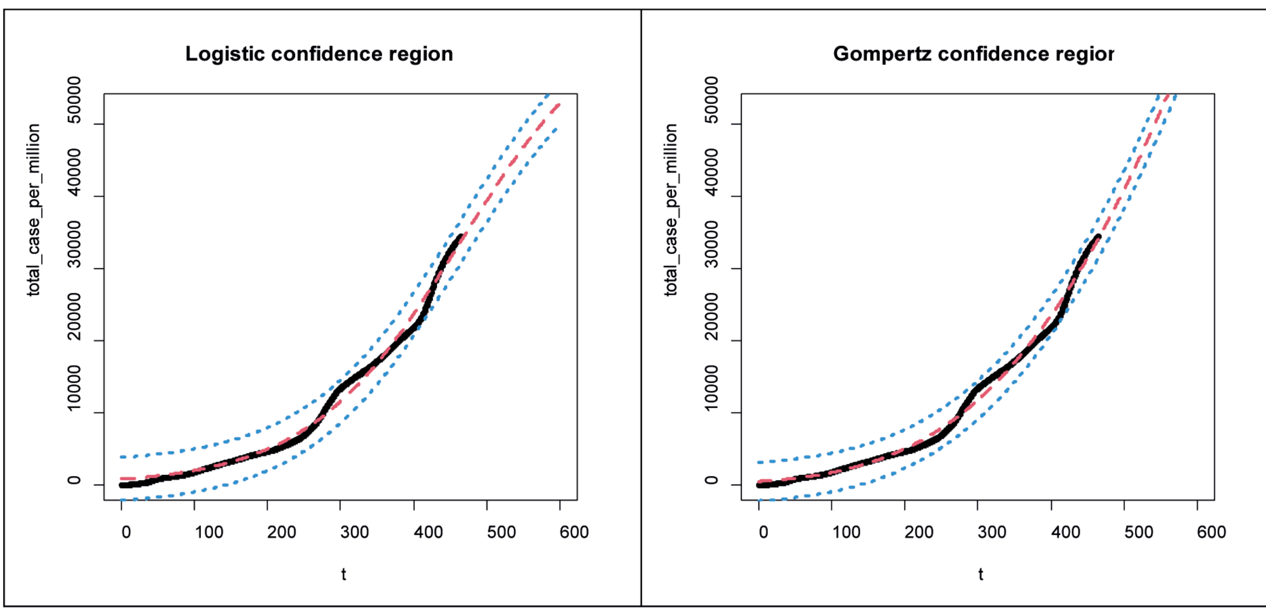

Figure 4. Growth curve and confidence interval of logistic and Gompertz models

Table 2. Regression coefficients of the logistic and Gompertz models

Parameters	Estimates	Std.Error	t-value	p-value
		Logistic Model		
θ_1	66.9603411	2.8777	23.27	2e-16 ***
θ_2	4.46266600	0.0247	180.24	2e-16 ***
θ_3	-0.0096369	0.00015	-61.60	2e-16 ***
		Gompertz Model		
θ_1	395.8	60.90	6.499	2e-16 ***
θ_2	6.797	0.09542	71.230	2e-16 ***
θ_3	0.00219	0.0001042	21.045	2e-16 ***

Table 3. Forecasting future values and testing RMSE and AIC using logistic and Gompertz models

Date	Observed	Logistic	Gompertz
2021-05-29	34.55163	33.95073	34.31333
2021-05-30	34.68310	34.11201	34.49770
2021-05-31	34.81033	34.27326	34.68265
2021-06-01	34.94868	34.43447	34.86819
2021-06-02	35.06365	34.59564	35.05431
2021-06-03	35.17329	34.75675	35.31333
2021-06-03	35.24999	34.91781	35.42829
2021-06-04	35.31680	35.07880	35.61615
2021-06-05	35.37523	35.23972	35.80460
2021-06-06	35.48054	35.40055	35.99362
2021-06-07	35.60673	35.56130	36.18323
Testing RMSE	-	1.48	0.56
Testing AIC	-	41.00	19.62

values very well with the lower testing RMSE. Therefore, the testing criteria indicates that the Gompertz model has a better description of future values than the logistic model.

Conclusion

The current paper considered the logistic family of growth curve including logistic, SSlogistic, generalized logistic and power logistic and Gompertz models to describe the scaled total_cases_per_million of COVID-19 in Iran during the 19-Feb-2020 to 28-May-2021. As the results shown, the Gompertz and logistic models provided a better description of response variable than the alternatives. Therefore, the logistic and Gompertz models are able to describe and forecast the COVID-19 variables (including total cases, death, recovered and so on) very well. However, the Gompertz model has better description of t_c_p_m variable in Iran than the logistic model.

Conflict of Interest

Authors have no conflict of interests.

References

1. Jamshidi, B., Rezaei, M., Zargaran, S. J., & Najafi, F. (2020). Mathematical modeling the epicenters of coronavirus disease-2019 (COVID-19) pandemic. Epidemiologic Methods, 9(s1).

2. Jamshidi, B., Bekrizadeh, H., Zargaran, S. J., & Rezaei, M. (2021). Analysis of the Number of Tests, the Positivity Rate, and Their Dependency Structure during COVID-19 Pandemic. medRxiv.

3. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet, 395(10223), 497-506.

4. Cauchemez S, Van Kerkhove M, Riley S, Donnelly C, Fraser C, Ferguson N. Transmission scenarios for Middle East Respiratory Syndrome Coronavirus (MERSCoV) and how to tell them apart. Euro Surveill. Bull. Eur. Sur Les Mal. Transm. Eur.

Commun. Dis. Bull. 2013;18:20503.

5. Ahmadi, A., Fadaei, Y., Shirani, M., & Rahmani, F. (2020). Modeling and forecasting trend of COVID-19 epidemic in Iran until May 13, 2020. Medical Journal of the Islamic Republic of Iran, 34, 27.

6. Anderson, R. M., Fraser, C., Ghani, A. C., Donnelly, C. A., Riley, S., Ferguson, N. M., ... & Hedley, A. J. (2004). Epidemiology, transmission dynamics and control of SARS: the 2002–2003 epidemic. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 359(1447), 1091-1105.

7. Peeri, N. C., Shrestha, N., Rahman, M. S., Zaki, R., Tan, Z., Bibi, S., ... & Haque, U. (2020). The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? International journal of epidemiology, 49(3), 717-726.

8. Mesel-Lemoine, M., Millet, J., Vidalain, P. O., Law, H., Vabret, A., Lorin, V., ... & Tangy, F. (2012). A human coronavirus responsible for the common cold massively kills dendritic cells but not monocytes. Journal of virology, 86(14), 7577-7587.

9. Liu, Y., Yan, L. M., & Wan, L. Viral dynamics in mild and severe cases of COVID-19 The Lancet Infect Dis. 19, 2020. doi: 10.1016. S1473-3099 (20), 30232-2.

10. Alene, M., Yismaw, L., Assemie, M. A., Ketema, D. B., Gietaneh, W., & Birhan, T. Y. (2021). Serial interval and incubation period of COVID-19: a systematic review and meta-analysis. BMC Infectious Diseases, 21(1), 1-9.

11. McCloskey, B., Zumla, A., Ippolito, G., Blumberg, L., Arbon, P., Cicero, A., ... & Borodina, M. (2020). Mass gathering events and reducing further global spread of COVID-19: a political and public health dilemma. The Lancet, 395(10230), 1096-1099.

12. Zhou, C., Su, F., Pei, T., Zhang, A., Du, Y., Luo, B., ... & Xiao, H. (2020). COVID-19: challenges to GIS with big data. Geography and sustainability, 1(1), 77-87.

13. Maleki, M., Mahmoudi, M. R., Wraith, D., & Pho, K. H. (2020). Time series modelling to forecast the confirmed and recovered cases of COVID-19. Travel medicine and infectious disease, 37, 101742.

14. Maleki, M., Mahmoudi, M. R., Heydari, M. H., & Pho, K. H. (2020). Modeling and forecasting the spread and death rate of coronavirus (COVID-19) in the world using time series models. Chaos, Solitons & Fractals, 140, 110151.

15. Salgotra, R., Gandomi, M., & Gandomi, A. H. (2020). Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming. Chaos, Solitons & Fractals, 138, 109945.

16. Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons & Fractals, 135, 109864.

17. Mohammadi, F., Pourzamani, H., Karimi, H., Mohammadi, M., Mohammadi, M., Ardalan, N., ... & Jannati, M. (2021). Artificial neural network and logistic regression modelling to characterize COVID-19 infected patients in local areas of Iran. Biomedical journal.

18. Bhandari, S., Shaktawat, A. S., Tak, A., Patel, B., Shukla, J., Singhal, S., ... & Dube, A. (2020). Logistic regression analysis to predict mortality risk in COVID-19 patients from routine hematologic parameters. Ibnosina Journal of Medicine and Biomedical Sciences, 12(2), 123.

19. Mehrolia, S., Alagarsamy, S., & Solaikutty, V. M. (2021). Customers response to online food delivery services during COVID-19 outbreak using binary logistic regression. International journal of consumer studies, 45(3), 396-408.

20. Almeshal, A. M., Almazrouee, A. I., Alenizi, M. R., & Alhajeri, S. N. (2020). Forecasting the spread of COVID-19 in Kuwait using compartmental and logistic regression models. Applied Sciences, 10(10), 3402.

21. Torrealba-Rodriguez, O., Conde-Gutiérrez, R. A., & Hernández-Javier, A. L. (2020). Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models. Chaos, Solitons & Fractals, 138, 109946.

22. Berihuete, Á., Sánchez-Sánchez, M., & Suárez-Llorens, A. (2021). A Bayesian Model of COVID-19 Cases Based on the Gompertz Curve. Mathematics, 9(3), 228.

23. Lee, S. Y., Lei, B., & Mallick, B. (2020). Estimation of COVID-19 spread curves integrating global data and borrowing information. PloS one, 15(7), e0236860.

24. Liu, W., Tang, S., & Xiao, Y. (2015). Model selection and evaluation based on emerging infectious disease data sets including A/H1N1 and Ebola. Computational and mathematical methods in medicine, 2015.

25. Tjørve, K. M., & Tjørve, E. (2017). The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family. PloS one, 12(6), e0178691.

26. Pinheiro, J. C., & Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. Mixed-effects models in S and S-Plus, 3-56.

Appendix A:

```
> b10=max(y)

> b20=log(b10/y[1]-1)

>b30=log(b10/y[2]-1)-b20

> flogis=nls(y~b1/(1+exp(b2+b3*t)),start=list(b1=b10,b2=b20,b3=b30),

trace=FALSE)

> summary(flogis)
```