



Uncertain growth model for the cumulative number of COVID-19 infections in China

Zhe Liu¹ 

Accepted: 7 September 2020 / Published online: 19 September 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

As a type of coronavirus, COVID-19 has quickly spread around the majority of countries worldwide, and seriously threatens human health and security. This paper aims to depict cumulative numbers of COVID-19 infections in China using the growth model chosen by cross validation. The residual plot does not look like a null plot, so we can not find a distribution function for the disturbance term that is close enough to the true frequency. Therefore, the disturbance term can not be characterized as random variables, and stochastic regression analysis is invalid in this case. To better describe this pandemic automatically, this paper first employs uncertain growth models with the help of uncertain hypothesis tests to detect and modify outliers in data. The forecast value and confidence interval for the cumulative number of COVID-19 infections in China are provided.

Keywords Uncertainty theory · Uncertain statistics · Uncertain regression analysis · Uncertain hypothesis test · COVID-19

1 Introduction

Regression analysis estimates relationships among variables. Although stochastic regression analysis has a long history of development, they are all considered under the framework of probability theory. However, the premise of probability theory, i.e., the estimated distribution being close enough to the true frequency, cannot be satisfied in many cases. Motivated by this, Liu (2007, 2009) founded uncertainty theory based on normality, duality, subadditivity, and product axioms to better address the inaccuracy of the human system. Currently, this theory has been successfully applied in uncertain statistics (Liu 2010). For example, Yang and Liu (2019) first presented uncertain time series analysis to predict future values based on imprecise observations. Following that,

✉ Zhe Liu
zliu18801296138@126.com

¹ School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China

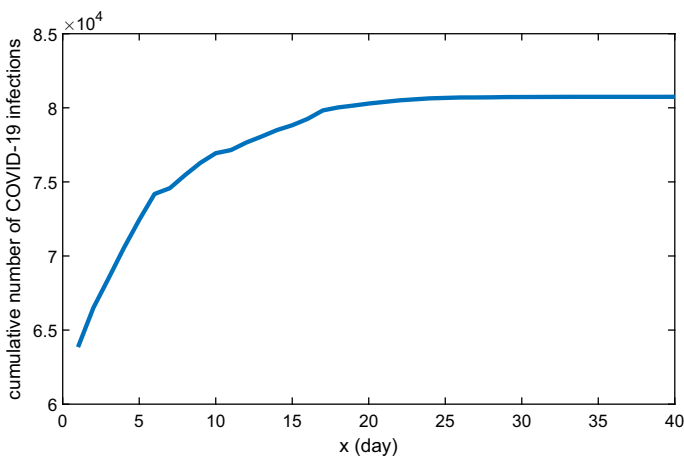
more research interests (Yang and Ni 2020; Liu and Yang 2020b; Tang 2020) were drawn to study and extend this topic. Especially, Ye and Yang (2020) applied uncertain time series to modelling cumulative numbers of COVID-19 infections. Uncertain regression analysis has been developed to address relationships between variables under the framework of uncertainty theory. Yao and Liu (2018) proposed least squares estimations for unknown parameters in uncertain multiple regression models. Motivated by this, researchers considered other estimations such as least absolute deviations estimations (Liu and Yang 2020a), Tukey's biweight estimations (Chen 2020), and maximum likelihood estimations (Lio and Liu 2020). In addition, Lio and Liu (2018) provided the confidence interval for the response variable. To evaluate different uncertain regression models, cross-validation methods (Liu and Jia 2020; Liu 2019) have attracted some scholars' attention. Later, estimations for the unknown parameters in uncertain multivariate regression models were derived (Ye and Liu 2020b; Zhang et al. 2020). The appropriateness of estimations for unknown parameters in uncertain regression analysis can be tested using the uncertain hypothesis test (Ye and Liu 2020a).

There is no doubt that emerging infectious diseases seriously threaten human health and security. Currently, as an emerging respiratory infectious disease, coronavirus disease 2019 (abbreviated "COVID-19") is caused by severe acute respiratory syndrome coronavirus, and is characterized as a "pandemic" by the World Health Organization. Reported illnesses of COVID-19 have ranged from very mildness to severity, including illness resulting in death. In a short time, the outbreak has rapidly worsened, and has received considerable global attention. After the irresolution in December 2019 and January 2020, the Chinese government has taken a series of multifaceted public health interventions, such as home confinement, traffic restrictions, and centralized quarantine, to effectively strengthen the control of the COVID-19 outbreak. In contrast, many other countries are in the acceleration phase of the pandemic. According to the World Health Organization (2020), the number of confirmed cases worldwide reached 332,930 by March 23, 2020. The relationship between the cumulative number of COVID-19 infections and time is a core issue reflecting the severity of COVID-19. For this purpose, growth models as a type of nonlinear regression model represent how a particular quantity increases over time, and have been used to describe previous epidemic growth patterns (Chowell et al. 2016).

To guide prevention and mitigation plans, this paper aims to characterize the cumulative number of COVID-19 infections in China excluding imported cases (hereinafter referred to as cumulative numbers of COVID-19 infections in China) using growth models from February 13 to March 23, 2020. The data before February 13 are not used because they are not real-time data due to the limitation of testing ability. Section 2 will choose a best S-shaped growth model using cross validation among several growth models, and give the results using stochastic growth models. Section 3 is going to introduce some fundamental knowledge about the uncertain growth model. The uncertain hypothesis test and data modification will be introduced in Sect. 4. Section 5 will present an algorithm. Following that, Sect. 6 will analyze cumulative numbers of COVID-19 infections in China using uncertain growth models and uncertain hypothesis tests. Section 7 is going to compare results obtained from the stochastic growth model and uncertain growth model. Finally, Sect. 8 will conclude this paper.

Table 1 Cumulative numbers of COVID-19 infections in China excluding imported cases from January 20 to March 23, 2020 reported by http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml

291	440	571	830	1287	1975	2744	4515
5974	7711	9692	11,791	14,380	17,205	20,438	24,324
28,018	31,161	34,546	37,198	40,171	42,638	44,653	59,804
63,851	66,492	68,500	70,548	72,436	74,185	74,576	75,465
76,288	76,936	77,150	77,658	78,064	78,497	78,824	79,251
79,824	80,026	80,151	80,270	80,389	80,516	80,591	80,632
80,668	80,685	80,699	80,708	80,725	80,729	80,733	80,737
80,738	80,739	80,739	80,739	80,739	80,740	80,740	80,744

**Fig. 1** Cumulative numbers of COVID-19 infections in China excluding imported cases from February 13 ($x = 1$) to March 23, 2020 ($x = 40$) reported by National Health Commission of China

2 Stochastic growth model for the cumulative number of COVID-19 infections in China

Before February 13, cumulative numbers of COVID-19 infections in China are not real-time data due to the limitation of testing ability. Thus we focus on cumulative numbers of COVID-19 infections in China from February 13 to March 23, 2020 which can be seen in Table 1.

As Fig. 1 shows, in this period, the pandemic growth pattern in China coincides with an S-shaped growth curve due to the urgent and ambitious measures taken by the government. For example, since January 23, the government blocked all outbound transportation from Wuhan and banned public transit. A series of social distancing measures such as compulsory mask wearing and cancellation of social gatherings, were also implemented. On February 17, the government initiated a door-to-door symptom screening for all residents.

Table 2 Different S-shaped growth models and average testing error (ATE) with 5 fold cross validation

Model	Form	ATE
Logistic	$y = \beta_0 / (1 - \beta_1 \exp(-\beta_2 x)) + \epsilon$	1.6693×10^5
Gompertz	$y = \beta_0 \exp(-\beta_1 \exp(-\beta_2 x)) + \epsilon$	2.0507×10^7
Mitcherlich	$y = \beta_0 - \beta_1 \beta_2^x + \epsilon$	1.9857×10^7
Monomolecular	$y = \beta_0 (1 - \beta_1 \exp(-\beta_2 x)) + \epsilon$	2.0445×10^7
Negative exponential	$y = \beta_0 (1 - \exp(-\beta_1 x)) + \epsilon$	9.3902×10^7

To select a model with the best generalization ability among some S-shaped growth models, we apply v -fold cross validation. Data are partitioned into v almost equal-sized subsets, and data in each subset are treated as future values to calculate square testing error with parameters estimated using data in other subsets. Then, the model with the smallest average testing error is selected. We use different S-shaped growth models to investigate how the cumulative number of COVID-19 infections in China increase over time. Average testing errors for different S-shaped growth models are shown in Table 2.

Therefore, we choose the logistic growth model

$$y = \beta_0 / (1 - \beta_1 \exp(-\beta_2 x)) + \epsilon$$

where the response variable y represents the cumulative number of COVID-19 infections in China, the explanatory variable x represents the day after February 12, 2020, and ϵ is a disturbance term. From January 20 to March 23, 2020, values of the explanatory variable x are integers from -23 to 40 , respectively. That is to say, the explanatory variable x equals to $-23, -22, \dots, 40$ on January 20, January 21, \dots , March 23, 2020, respectively. With cumulative numbers of COVID-19 infections in China from February 13 ($x = 1$) to March 23, 2020 ($x = 40$), the estimations $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ for unknown parameters $(\beta_0, \beta_1, \beta_2)$ using the function “lsqnonlin” in Version 8.6.0.267246 Matlab are

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (80786, -0.3088, 0.1841).$$

Thus the stochastic growth model is

$$y = 80786 / (1 + 0.3088 \exp(-0.1841x)) + \epsilon \quad (1)$$

where ϵ has an estimated expected value $\hat{\epsilon} = -0.3274$ and an estimated variance $\hat{\sigma}^2 = 256.24^2$.

To evaluate this model, we consider the coefficient of determination (R^2), which represents the proportion of the variance for a dependent variable that is explained by variables in a regression model. With data (x_i, y_i) , $i = 1, 2, \dots, n$, the total sum of squares is $SS_{tot} = \sum_i (y_i - \bar{y})^2$ where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, the sum of squares of residual is $SS_{res} = \sum_i (y_i - \hat{y}_i)^2$ where \hat{y}_i is the fitted value, and the coefficient of

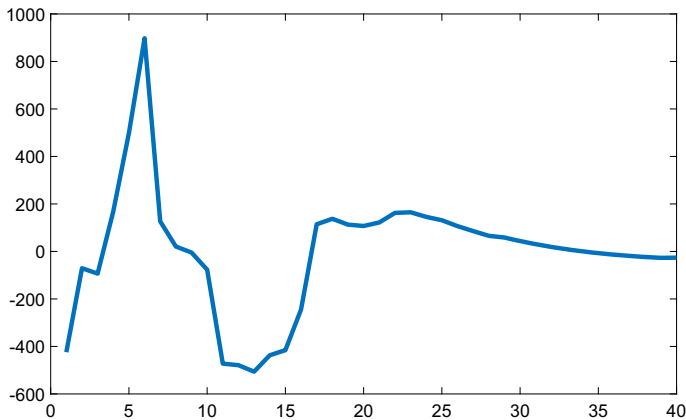


Fig. 2 Residual plot for the stochastic growth model (1)

determination R^2 is $R^2 = 1 - SS_{res}/SS_{tot}$. The R^2 for the stochastic growth model (1) is 0.9963, where in the best case $R^2 = 1$.

Stochastic regression analysis requires that a residual plot should look like a null plot which has constant mean, constant variance, and no separated points. However, it seems that the residual plot for the stochastic growth model (1) shown in Fig. 2 is not a null plot. Under this situation, if we use stochastic regression analysis, the distribution function we give is not close enough to the real frequency for the disturbance term. Therefore, the disturbance term can not be characterized as a random variable. As a result, stochastic regression analysis is invalid at this case, and that is the reason why we try to use uncertain regression analysis next.

3 Uncertain growth model for the cumulative number of COVID-19 infections in China

In this section, we introduce some fundamental knowledge about the uncertain growth model, including parameter estimation, residual analysis, and prediction.

The uncertain logistic growth model is

$$y = \beta_0 / (1 - \beta_1 \exp(-\beta_2 x)) + \epsilon \quad (2)$$

where y represents the cumulative number of COVID-19 infections in China, x represents the day after February 12, 2020, and ϵ is a disturbance term characterized as an uncertain variable. With data (x_i, y_i) , $i = 1, 2, \dots, n$ for the uncertain growth model (2), the least squares estimations Yao and Liu (2018) $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ for $(\beta_0, \beta_1, \beta_2)$ is the solution of the minimization problem

$$\min_{\beta_0 > 0, \beta_1 < 0, \beta_2 > 0} \sum_{i=1}^n (y_i - \beta_0 / (1 - \beta_1 \exp(-\beta_2 x_i)))^2, \quad (3)$$

and the fitted growth model is

$$y = \hat{\beta}_0 / (1 - \hat{\beta}_1 \exp(-\hat{\beta}_2 x)).$$

In the uncertain growth model (2), there exists a disturbance term ϵ that represents the difference between the predicted response variable's value and the true response variable's value. The estimated expected value and variance for ϵ Lio and Liu (2018) are suggested as

$$\hat{e} = \frac{1}{n} \sum_{i=1}^n \tilde{\epsilon}_i, \quad (4)$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (\tilde{\epsilon}_i - \hat{e})^2, \quad (5)$$

respectively, where $\tilde{\epsilon}_i$ are the i -th residuals defined as

$$\tilde{\epsilon}_i = y_i - \beta_0 / (1 - \beta_1 \exp(-\beta_2 x_i)), \quad (6)$$

$i = 1, 2, \dots, n$, respectively. If we further assume the disturbance term ϵ is a normal uncertain variable, then the uncertainty distribution for ϵ can be suggested as

$$\Phi(x) = \left(1 + \exp\left(\frac{\pi(\hat{e} - x)}{\sqrt{3}\hat{\sigma}}\right) \right)^{-1}.$$

For a new explanatory variable x , the forecast uncertain variable Lio and Liu (2018) of y is

$$\hat{y} = \hat{\beta}_0 / (1 - \hat{\beta}_1 \exp(-\hat{\beta}_2 x)) + \epsilon, \quad (7)$$

and the forecast value of y is

$$\mu = \hat{\beta}_0 / (1 - \hat{\beta}_1 \exp(-\hat{\beta}_2 x)) + \hat{e}. \quad (8)$$

According to the operational law for calculating the inverse uncertainty distributions Liu (2007), the inverse uncertainty distribution of \hat{y} equals to

$$\hat{\Psi}^{-1}(\alpha) = \hat{\beta}_0 / (1 - \hat{\beta}_1 \exp(-\hat{\beta}_2 x)) + \Phi^{-1}(\alpha) \quad (9)$$

with

$$\Phi^{-1}(\alpha) = \hat{e} + \frac{\sqrt{3}\hat{\sigma}}{\pi} \ln \frac{\alpha}{1-\alpha}.$$

Then the β ($0 < \beta < 1$) confidence interval of y Lio and Liu (2018) is

$$[\mu - b, \mu + b] \quad (10)$$

where b is the minimum value such that

$$\hat{\Psi}(\mu + b) - \hat{\Psi}(\mu - b) \geq \beta.$$

4 Uncertain hypothesis test and data modification

In this section, we apply the uncertain hypothesis test Ye and Liu (2020a) to testing the appropriateness of the estimated expected value \hat{e} and the estimated variance $\hat{\sigma}^2$ for the disturbance term ϵ in the uncertain growth model (2), and modify outliers.

Suppose that the disturbance term ϵ in the uncertain growth model (2) is a normal uncertain variable with expected value e and variance σ^2 , respectively, i.e., $\epsilon \sim \mathcal{N}(e, \sigma)$. Then for the two-side hypothesis

$$H_0 : e = \hat{e} \text{ and } \sigma = \hat{\sigma} \text{ versus } H_1 : e \neq \hat{e} \text{ or } \sigma \neq \hat{\sigma},$$

the uncertain hypothesis test is

$$W = \left\{ (\tilde{\epsilon}_1, \tilde{\epsilon}_2, \dots, \tilde{\epsilon}_n) \mid \Phi^{-1}(\alpha) \leq \tilde{\epsilon}_i \leq \Phi^{-1}(1 - \alpha), i = 1, 2, \dots, n \right\}^c \quad (11)$$

where $\tilde{\epsilon}_i$ are defined in (6), and

$$\Phi^{-1}(\alpha) = \hat{e} + \frac{\sqrt{3}\hat{\sigma}}{\pi} \ln \frac{\alpha}{1 - \alpha}.$$

The estimated expected value \hat{e} and estimated variance $\hat{\sigma}^2$ for the disturbance term ϵ in the uncertain growth model (2) pass the test if and only if $(\tilde{\epsilon}_1, \tilde{\epsilon}_2, \dots, \tilde{\epsilon}_n) \notin W$.

For each i ($i = 1, 2, \dots, n$), if

$$\tilde{\epsilon}_i < \hat{e} + \frac{\sqrt{3}\hat{\sigma}}{\pi} \ln \frac{\alpha}{1 - \alpha} \text{ or } \tilde{\epsilon}_i > \hat{e} + \frac{\sqrt{3}\hat{\sigma}}{\pi} \ln \frac{1 - \alpha}{\alpha},$$

then (x_i, y_i) is called an outlier. We modify y_i as

$$y_i = \hat{\beta}_0 / (1 - \hat{\beta}_1 \exp(-\hat{\beta}_2 x_i)),$$

and get the corresponding modified data (x_i, y_i) .

5 Algorithm

In this section, we present an algorithm to summarize the previous section.

Step 1 (Parameter estimation). With the data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, compute least square estimations $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ for $(\beta_0, \beta_1, \beta_2)$ in the uncertain logistic growth model $y = \beta_0 / (1 - \beta_1 \exp(-\beta_2 x)) + \epsilon$, i -th residuals $\tilde{\epsilon}_i$ ($i = 1, 2, \dots, n$), the estimated expected value \hat{e} , and the estimated variance $\hat{\sigma}^2$ according to Eqs. (3), (6), (4), and (5), respectively.

Step 2 (Uncertain hypothesis test). For a given significance level α (e.g. $\alpha = 0.01$), construct the uncertain hypothesis test W defined in (11).

Step 3 (Data modification) If $(\tilde{\epsilon}_1, \tilde{\epsilon}_2, \dots, \tilde{\epsilon}_n) \notin W$, go to Step 4. Otherwise, for each i ($i = 1, 2, \dots, n$), if

$$\tilde{\epsilon}_i < \hat{e} + \frac{\sqrt{3}\hat{\sigma}}{\pi} \ln \frac{\alpha}{1-\alpha} \text{ or } \tilde{\epsilon}_i > \hat{e} + \frac{\sqrt{3}\hat{\sigma}}{\pi} \ln \frac{1-\alpha}{\alpha},$$

then set $y_i = \hat{\beta}_0 / (1 - \hat{\beta}_1 \exp(-\hat{\beta}_2 x_i))$. Go to Step 1.

Step 4 (Forecast) For a new explanatory variable x , calculate the forecast uncertain variable, the forecast value, and the β (e.g. $\beta = 0.95$) confidence interval of y suggested as Eqs. (7), (8), and (10), respectively.

6 Real data analysis using the cumulative number of COVID-19 infections in China

6.1 The first iteration

With cumulative numbers of COVID-19 infections in China from February 13 to May 2, 2020 shown in Table 1, the least squares estimations $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ for $(\beta_0, \beta_1, \beta_2)$ are

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (80786, -0.3088, 0.1841)$$

according to Eq. (3), and the fitted growth model is

$$y = 80786 / (1 + 0.3088 \exp(-0.1841x)). \quad (12)$$

It follows from Eq. (6) that for each i ($i = 1, 2, \dots, 40$), the i -th residual can be calculated as

$$\tilde{\epsilon}_i = y_i - 80786 / (1 + 0.3088 \exp(-0.1841x_i)),$$

and the estimated expected value and estimated variance of the disturbance term ϵ are

$$\hat{e} = \frac{1}{40} \sum_{i=1}^{40} \tilde{\epsilon}_i = -0.3274,$$

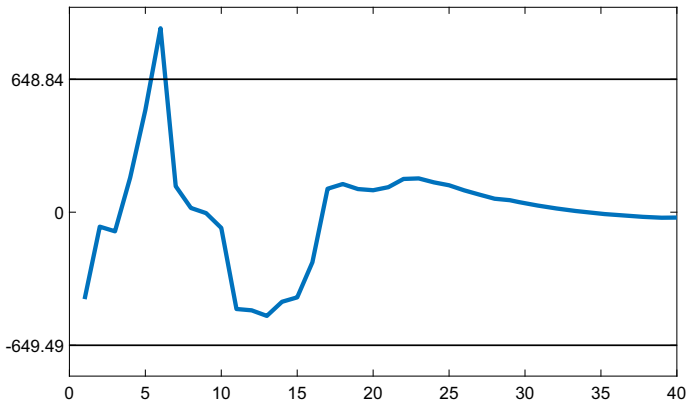


Fig. 3 Residuals for uncertain growth model (12)

and

$$\hat{\sigma}^2 = \frac{1}{40} \sum_{i=1}^{40} (\tilde{\epsilon}_i - \hat{\epsilon})^2 = 256.24^2,$$

respectively, according to Eqs. (4) and (5). Assuming that the disturbance term for uncertain growth model (12) is a normal uncertain variable with expected value -0.3274 and variance 256.24^2 , i.e., $\epsilon \sim \mathcal{N}(-0.3274, 256.24)$, the uncertain hypothesis test (11) for model (12) at a level of significance $\alpha = 0.01$ is

$$W = \left\{ (\tilde{\epsilon}_1, \tilde{\epsilon}_2, \dots, \tilde{\epsilon}_{40}) \mid -649.49 \leq \tilde{\epsilon}_i \leq 648.84, i = 1, 2, \dots, 40 \right\}^c.$$

As shown in Fig. 3, data on February 18 does not pass the test.

6.2 The second iteration

Modify data on February 18, and reestimate unknown parameters in the logistic growth model. It follows from Eq. (3) that uncertain least squares estimations are

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (80809, -0.3090, 0.1814),$$

and the fitted growth model is

$$y = 80809 / (1 + 0.3090 \exp(-0.1814x)). \quad (13)$$

Similarly, the estimated expected value and estimated variance of the disturbance term in uncertain growth model (13) are $\hat{\epsilon} = 0.1196$ and $\hat{\sigma}^2 = 208.28^2$, respectively, according to Eqs. (4) and (5). Assume that the disturbance term is a normal uncertain variable with expected value 0.1196 and variance 208.28^2 , i.e., $\epsilon \sim$

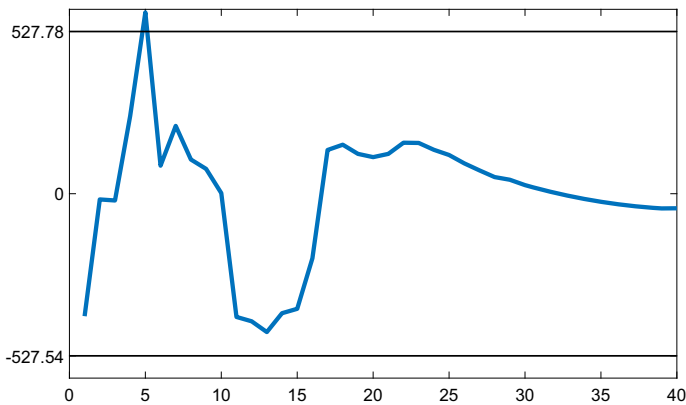


Fig. 4 Residuals for uncertain growth model (13)

$\mathcal{N}(0.1196, 208.28)$. Then the uncertain hypothesis test (11) at a level of significance $\alpha = 0.01$ is

$$W = \left\{ (\tilde{\epsilon}_1, \tilde{\epsilon}_2, \dots, \tilde{\epsilon}_{40}) \mid -527.54 \leq \tilde{\epsilon}_i \leq 527.78, i = 1, 2, \dots, 40 \right\}^c.$$

As shown in Fig. 4, data on February 17 does not pass the test.

6.3 The third iteration

Modify data on February 17, and reestimate unknown parameters in the logistic growth model. According to Eq. (3), uncertain least squares estimations are

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (80822, -0.3100, 0.1802),$$

and the fitted growth model is

$$y = 80822 / (1 + 0.3100 \exp(-0.1802x)). \quad (14)$$

The estimated expected value and estimated variance of the disturbance term in uncertain growth model (14) are $\hat{e} = 0.5998$ and $\hat{\sigma} = 183.82^2$, respectively, according to Eqs. (4) and (5). Assume that the disturbance term is a normal uncertain variable with expected value 0.5998 and variance 183.82^2 , i.e., $\epsilon \sim \mathcal{N}(0.5998, 183.82)$. Then the uncertain hypothesis test (11) at a level of significance $\alpha = 0.01$ is

$$W = \left\{ (\tilde{\epsilon}_1, \tilde{\epsilon}_2, \dots, \tilde{\epsilon}_{40}) \mid -465.09 \leq \tilde{\epsilon}_i \leq 466.29, i = 1, 2, \dots, 40 \right\}^c.$$

As shown in Fig. 5, all data pass the test.

Fitted uncertain regression growth model (14) and cumulative numbers of COVID-19 infections in China from January 20 to March 23, 2020, are shown in Fig. 6. Forecast

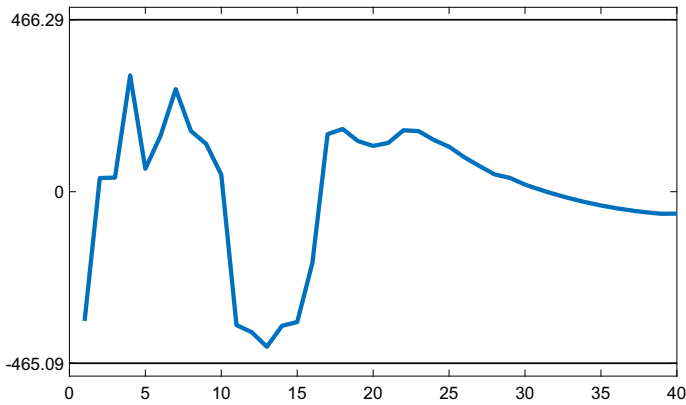


Fig. 5 Residuals for uncertain growth model (14)

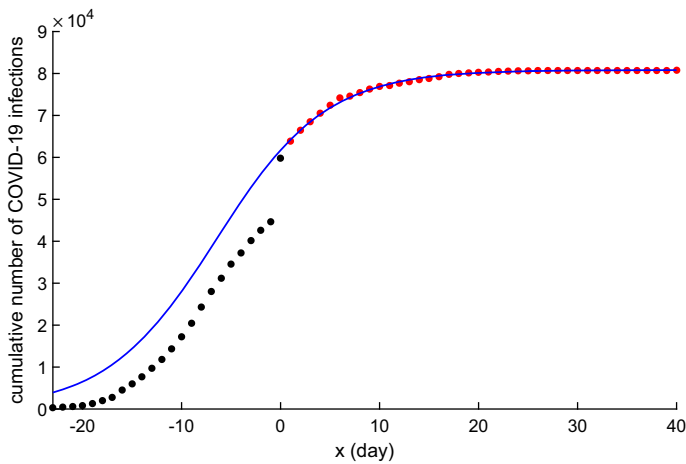


Fig. 6 Fitted growth model (14) (solid line), data (black dotted line) from January 20 ($x = -23$) to February 12 ($x = 0$), and data (red dotted line) from February 13 ($x = 1$) to March 23, 2020 ($x = 40$) for cumulative numbers of COVID-19 infections in China

values given by fitted uncertain regression growth model (14) are larger than data in Table 1 from January 20 to February 12. This result confirms that data before February 13 are not real-time data due to the limitation of testing ability. Next, we predict the cumulative number of COVID-19 infections in China on March 24, 2020 with the actual value 80744. According to Eqs. (7) and (8), with uncertain growth model (14) the forecast uncertain variable of the cumulative number of COVID-19 infections on March 24, 2020 in China is

$$\hat{y} = 80822 / (1 + 0.31 \exp(-0.1802 \times 41)) + \epsilon$$

where ϵ is a normal uncertain variable $\mathcal{N}(0.5998, 183.82)$, and the forecast value for the cumulative number of COVID-19 infections in China on March 24, 2020 is

$$\mu = 80822/(1 + 0.31 \exp(-0.1802 \times 41)) + 0.5998 = 80807.$$

It follows from Eq. (9) that the inverse uncertainty distribution of \hat{y} is

$$\hat{\Psi}^{-1}(\alpha) = 80822/(1 + 0.31 \exp(-0.1802 \times 41)) + \Phi^{-1}(\alpha)$$

where

$$\Phi^{-1}(\alpha) = 0.5998 + \frac{183.82\sqrt{3}}{\pi} \ln \frac{\alpha}{1 - \alpha}.$$

Then the 95% confidence interval for the cumulative number of COVID-19 infections in China on March 24, 2020, is determined as [80,439, 81,175] according to Eq. (10). From the uncertain growth model (14) we find that the cumulative number of COVID-19 infections in China tends to reach a saturation value by March 23, 2020. Thus, the containment measures implemented by China, such as the extreme lockdown of Wuhan and centralized quarantine, effectively strengthened the control of the COVID-19 outbreak.

7 Comparison results

In this section, we compare the stochastic growth model (1) and the uncertain growth model (14). Because the residual plot for the stochastic growth model (1) shown in Fig. 2 does not look like a null plot, we can not find a distribution function for the disturbance term that is close enough to the true frequency. Therefore, the disturbance term can not be characterized as a random variable, and stochastic regression analysis is invalid. What's more, as we can see from Table 3 the variance for the disturbance term in the stochastic growth model (1), i.e., $\hat{\sigma}^2 = 256.24^2$, is larger than that in the final uncertain growth model (14), i.e., $\hat{\sigma}^2 = 183.82^2$. With uncertain hypothesis tests, we detected and modified outliers in the cumulative number of COVID-19 infections in China. As a result, the disturbance term ϵ in the final uncertain growth model (14) has a smaller estimated variance. This fact shows that uncertain logistic growth model with the help of uncertain hypothesis tests is more suitable for handling the cumulative number of COVID-19 infections in China.

8 Conclusion

The emerging infectious disease COVID-19 seriously threatens human health and security worldwide. In this paper, the logistic growth model chosen by cross validation was implemented to depict the cumulative number of COVID-19 infections in China from February 13 to March 23, 2020. It seemed that the residual plot is not a null plot.

Table 3 Stochastic growth model versus uncertain growth model where μ and $\hat{\sigma}^2$ represent the forecast value for the cumulative number of COVID-19 infections in China on March 24, 2020 (with the actual value 80744) and the estimated variance for the disturbance term, respectively

	Stochastic growth model	Uncertain growth model
Forecast value (μ)	80,773	80,807
Variance ($\hat{\sigma}^2$)	256.24 ²	183.82 ²

So we can not find a distribution function for the disturbance term that is close enough to the true frequency. As a result, the disturbance term can not be characterized as a random variable, and stochastic regression analysis is invalid in this case. This paper first employed the uncertain growth model with the help of uncertain hypothesis tests to detect and modify outliers in data, and produced a better result automatically. The forecast value and confidence interval for the cumulative number of COVID-19 on March 24, 2020 were given. In the future work, better ways to handle outliers and better fitting models may be considered.

Acknowledgements This work was supported by National Natural Science Foundation of China (No. 11771241).

References

- Chen, D. (2020). Tukey's biweight estimation for uncertain regression model with imprecise observations. *Soft Computing*, <https://doi.org/10.1007/s00500-020-04973-x>.
- Chowell, G., Sattenspiel, L., Bansal, S., & Viboud, C. (2016). Mathematical models to characterize early epidemic growth: A review. *Physics of Life Reviews*, 18, 66–97.
- Lio, W., & Liu, B. (2018). Residual and confidence interval for uncertain regression model with imprecise observations. *Journal of Intelligent & Fuzzy Systems*, 35, 2573–2583.
- Lio, W., & Liu, B. (2020). Uncertain maximum likelihood estimation with application to uncertain regression analysis. *Soft Computing*, 24, 9351–9360.
- Liu, B. (2007). *Uncertainty theory* (2nd ed.). Berlin: Springer.
- Liu, B. (2009). Some research problems in uncertainty theory. *Journal of Uncertain Systems*, 3, 3–10.
- Liu, B. (2010). *Uncertainty theory: A branch of mathematics for modeling human uncertainty*. Berlin: Springer.
- Liu, S. (2019). Leave-p-out cross validation test for uncertain Verhulst–Pearl model with imprecise observations. *IEEE Access*, 7, 131705–131709.
- Liu, Z., & Jia, L. (2020). Cross-validation for the uncertain Chapman-Richards growth model with imprecise observations. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 28, 769–783.
- Liu, Z., & Yang, Y. (2020a). Least absolute deviations estimation for uncertain regression with imprecise observations. *Fuzzy Optimization and Decision Making*, 19, 33–52.
- Liu, Z., & Yang, X. (2020b). Cross validation for uncertain autoregressive model. *Communications in Statistics-Simulation and Computation*. <https://doi.org/10.1080/03610918.2020.1747077>.
- National Health Commission of the People's Republic of China. http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml.
- Tang, H. (2020). Uncertain vector autoregressive model with imprecise observations. *Soft Computing*, <https://doi.org/10.1007/s00500-020-04991-9>.
- World Health Organization. (2020). Coronavirus disease 2019 (COVID-19) situation report-99. Accessed April 28, 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200502-covid-19-sitrep-103.pdf?sfvrsn=d95e76d8_6.

- Yang, X., & Liu, B. (2019). Uncertain time series analysis with imprecise observations. *Fuzzy Optimization and Decision Making*, 18, 263–278.
- Yang, X., & Ni, Y. (2020). Least-squares estimation for uncertain moving average model. *Communications in Statistics-Theory and Methods*,. <https://doi.org/10.1080/03610926.2020.1713373>.
- Yao, K., & Liu, B. (2018). Uncertain regression analysis: An approach for imprecise observations. *Soft Computing*, 22, 5579–5582.
- Ye, T., & Liu, B. (2020a). Uncertain hypothesis test with application to uncertain regression analysis. Technical Report.
- Ye, T., & Liu, Y. (2020b). Multivariate uncertain regression model with imprecise observations. *Journal of Ambient Intelligence and Humanized Computing*. <https://doi.org/10.1007/s12652-020-01763-z>.
- Ye, T., & Yang, X. (2020). Analysis and prediction of confirmed cases of COVID-19 in China by uncertain time series model. *Fuzzy Optimization and Decision Making*. <https://doi.org/10.1007/s10700-020-09339-4>.
- Zhang, C., Liu, Z., & Liu, J. (2020). Least absolute deviations for uncertain multivariate regression model. *International Journal of General Systems*, 49, 449–465.

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