Uncertain regression model with autoregressive time series errors

Dan Chen

Accepted: 25 September 2021 / Published online: 21 October 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
Uncertain regression model is a powerful analytical tool for exploring the relationship between explanatory variables and response variables. It is assumed that the errors of regression equations are independent. However, in many cases, the error terms are highly positively autocorrelated. Assuming that the errors have an autoregressive structure, this paper first proposes an uncertain regression model with autoregressive time series errors. Then, the principle of least squares is used to estimate the unknown parameters in the model. Besides, this new methodology is used to analyze and predict the cumulative number of confirmed COVID-19 cases in China. Finally, this paper gives a comparative analysis of uncertain regression model, difference plus uncertain autoregressive model, and uncertain regression model with autoregressive time series errors. From the comparison, it is concluded that the uncertain regression model with autoregressive time series errors can improve the accuracy of predictions compared with the uncertain regression model.

Keywords: Uncertain regression analysis · Uncertain time series analysis · Uncertain hypothesis test · Uncertainty theory · COVID-19

1 Introduction
Uncertain statistics is a set of mathematical techniques for collecting, analyzing, and interpreting data by uncertainty theory (Liu 2007). Uncertain regression analysis as a branch of uncertain statistics is a set of statistical techniques that use uncertainty theory to explore the relationship between explanatory variables and response variables. The study of uncertain regression analysis was started by Yao and Liu (2018) by assuming that the disturbance term is an uncertain variable instead of a stochastic variable. To make point estimation for the unknown parameters in an uncertain regression model, they suggested the least squares estimation. Then, Liu and Yang (2020) considered the least absolute deviations estimation, and Lio and Liu (2020) proposed the maximum likelihood estimation. Lio and Liu (2018) further gave interval estimation of the predictive response variables based on the uncertain disturbance term. To evaluate the appropriateness of fitted regression model and estimated disturbance term, Ye and Liu (2021) introduced uncertain hypothesis test. In addition, uncertain regression analysis has been extended in various directions, such as multivariate regression analysis (Song and Fu 2018; Ye and Liu 2020; Zhang et al. 2020) and cross-validation (Liu and Jia 2020; Liu 2019). Uncertain time series analysis as another branch of uncertain statistics is a set of statistical techniques that use uncertainty theory to predict future values based on the previously observed values. The study of uncertain time series analysis was started by Yang and Liu (2019) by assuming that the disturbance term is an uncertain variable instead of a stochastic variable. To explore the relationship between these observations, they presented an uncertain autoregressive (UAR) model. Furthermore, they applied the principle of least squares to estimating the unknown autoregressive parameters. Then, Yang et al. (2020) considered the least absolute deviations estimation, and Chen and Yang (2021a, b) investigated the ridge estimation and the maximum likelihood estimation. To determine the optimum order of the UAR model, Liu and Yang (2020) gave cross-validation. In addition, some researchers studied other uncertain time series models, such as the 1-order uncertain moving average model (Yang and Ni 2021) and the uncertain vector autoregressive model (Tang 2020). To estimate the unknown parameters in an uncertain differential equation that fits...
the observed data as much as possible, several methods were proposed, for example, method of moments (Yao and Liu 2020), minimum cover estimation (Yang et al. 2020), least squares estimation (Sheng et al. 2020), generalized moment estimation (Liu 2021), and maximum likelihood estimation (Liu and Liu 2020). To obtain the unknown initial value of uncertain differential equation based on observed data, Lio and Liu (2021) proposed an estimation method.

In recent studies, these statistical techniques were utilized for COVID-19 prediction. For example, Liu (2021) applied the uncertain logistic growth model to fitting the cumulative number of confirmed COVID-19 cases in China. Ye and Yang (2021) applied the UAR model to analyzing the second difference of the cumulative number. Following that, Chen et al. (2021) presented an uncertain SIR model, starting spreading in China, Lio and Liu (2021) inferred the zero-day of COVID-19 spread using the initial value estimation.

This paper proposes a new methodology for the analysis and prediction of time series data. In 1949, Cochrane and Orcutt (1949) presented evidence showing that the error terms involved in most current formulations of economic relations are highly positively autocorrelated. They indicated that in many cases the assumption of random error terms is not a very good approximation to the truth. Assuming that the errors are an autoregressive process with finite order, Durbin (1960) proposed a two-stage procedure that yields asymptotically efficient estimates in linear model. Similarly, it is an oversimplification to assume that error terms are independent in time in uncertain regression analysis. To improve this situation, in this paper a different type of model is considered in which the errors in the model have an autoregressive structure, i.e., uncertain regression model with autoregressive time series errors.

The rest of this paper is organized as follows. Section 2 introduces an uncertain regression model with autoregressive time series errors in detail, including parameter estimation, residual analysis, forecast value, and confidence interval. In Sect. 3, the approach is applied to modeling the cumulative number of confirmed COVID-19 cases in China. A comparative study on uncertain regression model, difference plus UAR model, and uncertain regression model with autoregressive time series errors is analyzed in Sect. 4. Section 5 shows that stochastic regression model with autoregressive time series errors is not suitable. Finally, some conclusions are made in Sect. 6.

2 Uncertain regression model with autoregressive time series errors

The uncertain regression model with autoregressive time series errors has the form of

\[ Y_t = f(X_{t1}, X_{t2}, \cdots, X_{tp}) + Z_t \]  

where \( Y_t \) is a response series, \( (X_{t1}, X_{t2}, \cdots, X_{tp}) \) is a vector of explanatory series, \( \beta \) is an unknown vector of parameters, \( f(X_{t1}, X_{t2}, \cdots, X_{tp}) \) represents the effect of \( (X_{t1}, X_{t2}, \cdots, X_{tp}) \) on \( Y_t \), and \( Z_t \) is an error series for \( t = 1, 2, \cdots, n \). We assume that the errors of (1) follow a k-order uncertain autoregressive model, that is,

\[ Z_t = a_0 + \sum_{i=1}^{k} a_i Z_{t-i} + \epsilon_t \]

where the autoregressive coefficients \( a_0, a_1, \cdots, a_k \) are unknown, and \( \epsilon_t \) are uncertain disturbances (uncertain variables) for \( t = k+1, k+2, \cdots, n \).

**Remark 1** Uncertain regression (linear or nonlinear) and uncertain autoregressive models are special cases of (1).

**Remark 2** When \( (X_{t1}, X_{t2}, \cdots, X_{tp}) \) contains the time variable \( t \), (1) is used by economists to study the trend of \( Y_t \).

In the regression model with autoregressive time series errors (1), if \( f(X_{t1}, X_{t2}, \cdots, X_{tp}) \) is a linear function, i.e.,

\[
\begin{align*}
Y_t &= \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_p X_{tp} + Z_t \\
Z_t &= a_0 + \sum_{i=1}^{k} a_i Z_{t-i} + \epsilon_t,
\end{align*}
\]

then it is called a linear regression model with autoregressive time series errors. If \( f(X_{t1}, X_{t2}, \cdots, X_{tp}) \) is a logistic function, i.e.,

\[
\begin{align*}
Y_t &= \beta_0/(1 + \beta_1 \exp(-\beta_2 X_{t1})) + Z_t, \quad \beta_0, \beta_1, \beta_2 > 0 \\
Z_t &= a_0 + \sum_{i=1}^{k} a_i Z_{t-i} + \epsilon_t,
\end{align*}
\]

then it is called a logistic growth model with autoregressive time series errors.

2.1 Parameter estimation

Assume \( (x_{t1}, x_{t2}, \cdots, x_{tp}, y_t) \) are the observed data at times \( t \) for \( t = 1, 2, \cdots, n \), respectively. Based on the observed
data, the least squares estimate of $\beta$ in the uncertain regression model with autoregressive time series errors

$$
\begin{cases}
Y_t = f(X_{t1}, X_{t2}, \cdots, X_{tp}|\beta) + Z_t \\
Z_t = a_0 + \sum_{i=1}^{k} a_i Z_{t-i} + \varepsilon_t
\end{cases}
$$

(5)

is the solution, $\beta^*$, of the minimization problem,

$$
\min_{\beta} \sum_{t=1}^{n} (y_t - f(x_{t1}, x_{t2}, \cdots, x_{tp}|\beta))^2.
$$

(6)

Then, for each index $t$ ($t = 1, 2, \cdots, n$), the errors can be calculated as

$$
z_t = y_t - f(x_{t1}, x_{t2}, \cdots, x_{tp}|\beta^*).
$$

(7)

The errors $z_1, z_2, \cdots, z_n$ will be regarded as the samples of $Z_t$. The least squares estimate of $(a_0, a_1, \cdots, a_k)$ in the uncertain regression model with autoregressive time series errors is the solution, $(a_0^*, a_1^*, \cdots, a_k^*)$, of the minimization problem,

$$
\min_{a_0, a_1, \cdots, a_k} \sum_{t=k+1}^{n} \left( z_t - a_0 - \sum_{i=1}^{k} a_i z_{t-i} \right)^2.
$$

(8)

Thus, the fitted regression model with autoregressive time series errors is determined by

$$
\begin{cases}
Y_t = f(X_{t1}, X_{t2}, \cdots, X_{tp}|\beta^*) + Z_t \\
Z_t = a_0^* + \sum_{i=1}^{k} a_i^* Z_{t-i}.
\end{cases}
$$

(9)

**Example 1** Let $(x_{t1}, x_{t2}, \cdots, x_{tp}, y_t)$ be the observed data at times $t$ for $t = 1, 2, \cdots, n$, respectively. The least squares estimates of $\beta_0, \beta_1, \cdots, \beta_p$ and $a_0, a_1, \cdots, a_k$ in the linear regression model with autoregressive time series errors are

$$
\begin{cases}
Y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_p x_{tp} + Z_t \\
Z_t = a_0 + \sum_{i=1}^{k} a_i Z_{t-i} + \varepsilon_t
\end{cases}
$$

(10)

solve the minimization problems,

$$
\min_{\beta_0, \beta_1, \cdots, \beta_p} \sum_{t=1}^{n} \left( y_t - \beta_0 - \sum_{j=1}^{p} \beta_j x_{tj} \right)^2
$$

(11)

and

$$
\min_{a_0, a_1, \cdots, a_k} \sum_{t=k+1}^{n} \left( z_t - a_0 - \sum_{i=1}^{k} a_i z_{t-i} \right)^2.
$$

(12)

respectively, where

$$
z_t = y_t - \beta_0^* - \beta_1^* x_{t1} - \beta_2^* x_{t2} - \cdots - \beta_p^* x_{tp}
$$

(13)

for $t = 1, 2, \cdots, n$.

**Example 2** Let $(x_t, y_t)$ be the observed data at times $t$ for $t = 1, 2, \cdots, n$, respectively. The least squares estimates of $\beta_0, \beta_1, \beta_2$ and $a_0, a_1, \cdots, a_k$ in the logistic growth model with autoregressive time series errors are

$$
\begin{cases}
Y_t = \beta_0/(1 + \beta_1 \exp(-\beta_2 x_{t1})) + Z_t, \quad \beta_0, \beta_1, \beta_2 > 0 \\
Z_t = a_0 + \sum_{i=1}^{k} a_i Z_{t-i} + \varepsilon_t
\end{cases}
$$

(14)

solve the minimization problems,

$$
\min_{\beta_0>0, \beta_1>0, \beta_2>0} \sum_{t=1}^{n} (y_t - \beta_0/(1 + \beta_1 \exp(-\beta_2 x_{t1})))^2
$$

(15)

and

$$
\min_{a_0, a_1, \cdots, a_k} \sum_{t=k+1}^{n} \left( z_t - a_0 - \sum_{i=1}^{k} a_i z_{t-i} \right)^2,
$$

(16)

respectively, where

$$
z_t = y_t - \beta_0^*/(1 + \beta_1^* \exp(-\beta_2^* x_{t1}))
$$

(17)

for $t = 1, 2, \cdots, n$.

### 2.2 Residual analysis

**Definition 1** Let $(x_{t1}, x_{t2}, \cdots, x_{tp}, y_t)$ be the observed data at times $t$ for $t = 1, 2, \cdots, n$, respectively, and let the fitted regression model with autoregressive time series errors be

$$
\begin{cases}
Y_t = f(X_{t1}, X_{t2}, \cdots, X_{tp}|\beta^*) + Z_t \\
Z_t = a_0^* + \sum_{i=1}^{k} a_i^* Z_{t-i}.
\end{cases}
$$

(18)

Then, for each index $t$ ($t = k + 1, k + 2, \cdots, n$), the term

$$
\hat{\varepsilon}_t = y_t - f(x_{t1}, x_{t2}, \cdots, x_{tp}|\beta^*) - a_0^*
$$

(19)

$$
- \sum_{i=1}^{k} a_i^* \left( y_{t-i} - f(x_{t-i,1}, x_{t-i,2}, \cdots, x_{t-i,p}|\beta^*) \right)
$$

is called the $t$-th residual.

The residuals $\hat{\varepsilon}_{k+1}, \hat{\varepsilon}_{k+2}, \cdots, \hat{\varepsilon}_n$ will be regarded as the samples of the uncertain disturbance terms $\varepsilon_t$ in the uncertain
regression model with autoregressive time series errors

\[
\begin{align*}
Y_t &= f(X_{t1}, X_{t2}, \ldots, X_{tp}) + Z_t \\
Z_t &= a_0 + \sum_{i=1}^{k} a_i Z_{t-i} + \epsilon_t.
\end{align*}
\]  

(20)

Thus, the expected value of \( \epsilon_t \) can be estimated as the average of residuals, i.e.,

\[
\hat{\epsilon} = \frac{1}{n-k} \sum_{t=k+1}^{n} \hat{\epsilon}_t,
\]

(21)

and the variance can be estimated as

\[
\hat{\sigma}^2 = \frac{1}{n-k} \sum_{t=k+1}^{n} (\hat{\epsilon}_t - \hat{\epsilon})^2.
\]

(22)

Therefore, we may assume the estimated disturbance terms \( \hat{\epsilon}_t \) follow the normal uncertainty distribution \( \mathcal{N}(\hat{\epsilon}, \hat{\sigma}) \).

**Example 3** Let \((x_{1t}, x_{2t}, \ldots, x_{pt}, y_t)\) be the observed data at times \( t \) for \( t = 1, 2, \ldots, n \), respectively, and let the fitted linear regression model with autoregressive time series errors be

\[
\begin{align*}
Y_t &= \beta_0^* + \beta_1^* x_{t1} + \beta_2^* x_{t2} + \cdots + \beta_p^* x_{tp} + Z_t \\
Z_t &= a_0^* + \sum_{i=1}^{k} a_i^* Z_{t-i}.
\end{align*}
\]

(23)

The expected value of estimated disturbance terms \( \hat{\epsilon}_t \) is

\[
\hat{\epsilon} = \frac{1}{n-k} \sum_{t=k+1}^{n} \left( y_t - \beta_0^* - \sum_{j=1}^{p} \beta_j^* x_{tj} - a_0^* - \sum_{j=1}^{p} \sum_{i=1}^{k} a_i^* (y_{t-i} - \beta_0^* - \sum_{j=1}^{p} \beta_j^* x_{t-j,i}) \right),
\]

(24)

and the variance is

\[
\hat{\sigma}^2 = \frac{1}{n-k} \sum_{t=k+1}^{n} \left( y_t - \beta_0^* - \sum_{j=1}^{p} \beta_j^* x_{tj} - a_0^* - \sum_{j=1}^{p} \sum_{i=1}^{k} a_i^* (y_{t-i} - \beta_0^* - \sum_{j=1}^{p} \beta_j^* x_{t-j,i}) - \hat{\epsilon} \right)^2.
\]

(25)

**Example 4** Let \((x_t, y_t)\) be the observed data at times \( t \) for \( t = 1, 2, \ldots, n \), respectively, and let the fitted logistic growth model with autoregressive time series errors be

\[
\begin{align*}
Y_t &= \beta_0^*/(1 + \beta_1^* \exp(-\beta_2^* X_t)) + Z_t, \quad \beta_0^*, \beta_1^*, \beta_2^* > 0 \\
Z_t &= a_0^* + \sum_{i=1}^{k} a_i^* Z_{t-i}.
\end{align*}
\]

(26)

The expected value of estimated disturbance terms \( \hat{\epsilon}_t \) is

\[
\hat{\epsilon} = \frac{1}{n-k} \sum_{t=k+1}^{n} \left( y_t - \beta_0^*/(1 + \beta_1^* \exp(-\beta_2^* x_t)) - a_0^* - \sum_{i=1}^{k} a_i^* (y_{t-i} - \beta_0^*/(1 + \beta_1^* \exp(-\beta_2^* x_{t-i}))) \right),
\]

(27)

and the variance is

\[
\hat{\sigma}^2 = \frac{1}{n-k} \sum_{t=k+1}^{n} \left( y_t - \beta_0^*/(1 + \beta_1^* \exp(-\beta_2^* x_t)) - a_0^* - \sum_{i=1}^{k} a_i^* (y_{t-i} - \beta_0^*/(1 + \beta_1^* \exp(-\beta_2^* x_{t-i}))) - \hat{\epsilon} \right)^2.
\]

(28)

**Remark 3** After that, we apply uncertain hypothesis test (Ye and Liu 2021) to evaluating the appropriateness of fitted regression model with autoregressive time series errors and estimated disturbance terms.

**2.3 Forecast value**

Let \((x_{n+1,1}, x_{n+1,2}, \ldots, x_{n+1,p})\) be an explanatory vector at time \( n + 1 \). Assume (i) the fitted regression model with autoregressive time series errors is

\[
\begin{align*}
Y_t &= f(X_{t1}, X_{t2}, \ldots, X_{tp}) + Z_t \\
Z_t &= a_0^* + \sum_{i=1}^{k} a_i^* Z_{t-i},
\end{align*}
\]

(29)

and (ii) the estimated disturbance terms \( \hat{\epsilon}_t \) follow the normal uncertainty distribution with expected value \( \hat{\epsilon} \) determined by (21) and variance \( \hat{\sigma}^2 \) determined by (22). The forecast uncertain variable of \( Y_{n+1} \) with respect to \((x_{n+1,1}, x_{n+1,2}, \ldots, x_{n+1,p})\) is determined by

\[
\hat{Y}_{n+1} = f(x_{n+1,1}, x_{n+1,2}, \ldots, x_{n+1,p}) + a_0^* + \sum_{i=1}^{k} a_i^* (y_{n+1-i} - f(x_{n+1-i,1}, x_{n+1-i,2}, \ldots, x_{n+1-i,p}) | \beta^*) + \hat{\epsilon} + \hat{\epsilon}_{n+1},
\]

(30)

and the forecast value is defined as the expected value of the forecast uncertain variable \( \hat{Y}_{n+1} \), i.e.,

\[
\hat{y}_{n+1} = f(x_{n+1,1}, x_{n+1,2}, \ldots, x_{n+1,p}) + a_0^* + \sum_{i=1}^{k} a_i^* (y_{n+1-i} - f(x_{n+1-i,1}, x_{n+1-i,2}, \ldots, x_{n+1-i,p}) | \beta^*) + \hat{\epsilon}.
\]

(31)
**Example 5**  Let \((x_{n+1,1}, x_{n+1,2}, \cdots, x_{n+1,p})\) be an explanatory vector at time \(n+1\). Assume (i) the fitted linear regression model with autoregressive time series errors is
\[
\begin{align*}
Y_t &= \beta_0^* + \sum_{j=1}^{p} \beta_j^* x_{t,j} + Z_t, \\
Z_t &= a_0^* + \sum_{i=1}^{k} a_i^* Z_{t-i},
\end{align*}
\] (32)

and (ii) the estimated disturbance terms \(\tilde{e}_t\) follow the normal uncertainty distribution with expected value \(\hat{\epsilon}\) and variance \(\hat{\sigma}^2\). The forecast uncertain variable of \(Y_{n+1}\) with respect to \((x_{n+1,1}, x_{n+1,2}, \cdots, x_{n+1,p})\) is
\[
\hat{Y}_{n+1} = \beta_0^* + \sum_{j=1}^{p} \beta_j^* x_{n+1,j} + a_0^*
\] (33)

\[+ \sum_{i=1}^{k} a_i^* \left( y_{n+1-i} - \beta_0^* - \sum_{j=1}^{p} \beta_j^* x_{n+1-i,j} \right) + \tilde{e}_{n+1}, \]

and the forecast value of \(Y_{n+1}\) is
\[
\hat{y}_{n+1} = \beta_0^* + \sum_{j=1}^{p} \beta_j^* x_{n+1,j} + a_0^*
\] (34)

\[+ \sum_{i=1}^{k} a_i^* \left( y_{n+1-i} - \beta_0^* - \sum_{j=1}^{p} \beta_j^* x_{n+1-i,j} \right) + \hat{\epsilon}.
\]

**Example 6**  Let \(x_{n+1}\) be an explanatory variable at time \(n+1\). Assume (i) the fitted logistic growth model with autoregressive time series errors is
\[
\begin{align*}
Y_t &= \beta_0^*/(1 + \beta_1^* \exp(-\beta_2^* x_{t,1})) + Z_t, \\
Z_t &= a_0^* + \sum_{i=1}^{k} a_i^* Z_{t-i},
\end{align*}
\] (35)

and (ii) the estimated disturbance terms \(\tilde{e}_t\) follow the normal uncertainty distribution with expected value \(\hat{\epsilon}\) and variance \(\hat{\sigma}^2\). The forecast uncertain variable of \(Y_{n+1}\) with respect to \(x_{n+1}\) is
\[
\hat{Y}_{n+1} = \beta_0^*/(1 + \beta_1^* \exp(-\beta_2^* x_{n+1})) + a_0^*
\] (36)

\[+ \sum_{i=1}^{k} a_i^* \left( y_{n+1-i} - \beta_0^*/(1 + \beta_1^* \exp(-\beta_2^* x_{n+1-i})) \right) + \tilde{e}_{n+1},
\]

and the forecast value of \(Y_{n+1}\) is
\[
\hat{y}_{n+1} = \beta_0^*/(1 + \beta_1^* \exp(-\beta_2^* x_{n+1})) + a_0^*
\] (37)

\[+ \sum_{i=1}^{k} a_i^* \left( y_{n+1-i} - \beta_0^*/(1 + \beta_1^* \exp(-\beta_2^* x_{n+1-i})) \right) + \hat{\epsilon}.
\]

### 2.4 Confidence interval

Let \((x_{n+1,1}, x_{n+1,2}, \cdots, x_{n+1,p})\) be an explanatory vector at time \(n+1\). Assume the forecast uncertain variable of \(Y_{n+1}\) with respect to \((x_{n+1,1}, x_{n+1,2}, \cdots, x_{n+1,p})\) is
\[
\hat{Y}_{n+1} = f(x_{n+1,1}, x_{n+1,2}, \cdots, x_{n+1,p}|\beta^*) + a_0^* + \sum_{i=1}^{k} a_i^* \left( y_{n+1-i} - f(x_{n+1-i,1}, x_{n+1-i,2}, \cdots, x_{n+1-i,p}|\beta^*) \right) + \tilde{e}_{n+1}, \] (38)

\[\tilde{e}_{n+1} \sim N(\hat{\epsilon}, \hat{\sigma}).\]

Then, the forecast value of \(Y_{n+1}\) is
\[
\hat{y}_{n+1} = f(x_{n+1,1}, x_{n+1,2}, \cdots, x_{n+1,p}|\beta^*) + a_0^* + \sum_{i=1}^{k} a_i^* \left( y_{n+1-i} - f(x_{n+1-i,1}, x_{n+1-i,2}, \cdots, x_{n+1-i,p}|\beta^*) \right) + \hat{\epsilon}.
\] (39)

It follows from the operational law that \(\hat{Y}_{n+1}\) has a normal uncertainty distribution \(N(\hat{y}_{n+1}, \hat{\sigma})\), i.e.,
\[
\hat{\Phi}_{n+1}(x) = \left( 1 + \exp \left( \pi (\hat{y}_{n+1} - x) / \sqrt{3 \hat{\sigma}} \right) \right)^{-1}.
\] (40)

Taking \(\alpha\) as the confidence level, it is easy to verify that
\[
\hat{b} = \hat{\sigma} \sqrt{3 / \pi} \ln \frac{1 + \alpha}{1 - \alpha}
\] (41)

is the minimum value \(b\) such that
\[
\hat{\Phi}_{n+1}(\hat{y}_{n+1} + b) - \hat{\Phi}_{n+1}(\hat{y}_{n+1} - b) \geq \alpha.
\] (42)

Since \(M(\hat{y}_{n+1} - \hat{b} \leq \hat{Y}_{n+1} \leq \hat{y}_{n+1} + \hat{b}) \geq \hat{\Phi}_{n+1}(\hat{y}_{n+1} + b) - \hat{\Phi}_{n+1}(\hat{y}_{n+1} - b) = \alpha\), the \(\alpha\) confidence interval of \(Y_{n+1}\) is
\[
\hat{y}_{n+1} \pm \hat{\sigma} \sqrt{3 / \pi} \ln \frac{1 + \alpha}{1 - \alpha}.
\] (43)

### 3 Modeling the cumulative number of confirmed COVID-19 cases

In this section, the uncertain regression model with autoregressive time series errors is applied to analyzing the cumulative number of confirmed COVID-19 cases by local transmission in China. We use the same data for comparison with uncertain regression model (Liu 2021) and difference plus uncertain autoregressive model (Ye and Yang 2021), that is, the cumulative number of confirmed COVID-19 cases.
excluding imported cases from February 13 to March 23, 2020 (see Table 1). The data are plotted in Fig. 1.

Let \( t = 1, 2, \ldots, 40 \) represent the dates (\( t \)) from February 13 to March 23. For example, \( t = 1 \) and \( t = 40 \) represent February 13 and March 23, respectively. In order to determine the functional relationship between \( t \) (the date) and \( Y_t \) (the cumulative number of confirmed COVID-19 cases on date \( t \)), we may use the observed data

\[
(t, y_t), \quad t = 1, 2, \ldots, 40
\]

where \( y_t \) are the cumulative numbers shown in Table 1 on days \( t, \) \( t = 1, 2, \ldots, 40, \) respectively. For example,

\[
y_1 = 63851, \quad y_{40} = 80744.
\]

In order to fit the above observed data, we employ the uncertain logistic growth model with autoregressive time series errors,

\[
\begin{aligned}
Y_t &= \beta_0/(1 + \beta_1 \exp(-\beta_2 t)) + Z_t, \quad \beta_0, \beta_1, \beta_2 > 0 \\
Z_t &= a_0 + \sum_{i=1}^{k} a_i Z_{t-i} + \epsilon_t
\end{aligned}
\]

where \( Z_t \) is an error series for \( t = 1, 2, \ldots, 40, \) and \( \epsilon_t \) are uncertain disturbances for \( t = k + 1, k + 2, \ldots, 40. \)

Based on the observed data \((t, y_t), \) \( t = 1, 2, \ldots, 40, \) Liu (2021) obtained the fitted logistic growth component

\[
Y_t = 80822/(1 + 0.3100 \exp(-0.1802t)).
\]

Then, the observed data of the error series \( Z_t \) are

\[
z_t = y_t - 80822/(1 + 0.3100 \exp(-0.1802t))
\]

for \( t = 1, 2, \ldots, 40 \) (see Table 2). Next, we apply the UAR\( (k) \) model to modeling \( z_1, z_2, \ldots, z_{40}. \)

First we determine the value of the order \( k \) by rolling origin cross validation (Liu and Yang 2020). Assume that \( T = 37, \) and the average testing error \( ATE(k) \) is

\[
ATE(k) = \frac{1}{3-m} \sum_{m=0}^{40} \left( z_t - a_0^m - \sum_{i=1}^{k} a_i^m z_{t-i} \right)^2
\]

where \( (a_0^m, a_1^m, \ldots, a_k^m)^m \) are the least squares estimations using the observation data in the training sets \( \{z_1, z_2, \ldots, z_{3m} \} \) for \( m = 0, 1, 2, \) respectively. Table 3 provides a quick summary of the value of \( ATE(k) \) with \( k \in \{1, 2, 3, 4, 5 \}. \) When \( k = 4, \) we get the minimum value of \( ATE(k) \). Thus the UAR component is

\[
Z_t = a_0 + \sum_{i=1}^{4} a_i Z_{t-i} + \epsilon_t.
\]
In order to test whether it is appropriate, given a significance level $\alpha = 0.01$, the uncertain hypothesis test for the hypotheses

$H_0 : e = 0.0000$ and $\sigma = 96.0254$ versus

$H_1 : e \neq 0.0000$ or $\sigma \neq 96.0254$

is

$$W = \{(w_5, w_6, \ldots, w_{40}) \mid -243.2729 \leq w_t \leq 243.2729, t = 5, 6, \ldots, 40\}$$

Using the error data $z_1, z_2, \ldots, z_{40}$ and solving the minimization problem

$$\min_{a_0, a_1, \ldots, a_4} \sum_{t=5}^{40} \left( Z_t - a_0 - \sum_{i=1}^{4} a_i Z_{t-i} \right)^2,$$ (51)

we obtain a fitted UAR(4) component

$$Z_t = -7.4091 + 0.8058 Z_{t-1} + 0.0642 Z_{t-2} - 0.0606 Z_{t-3} - 0.1655 Z_{t-4}.$$ (52)

From

$$\hat{e}_t = z_t + 7.4091 - 0.8058 Z_{t-1} - 0.0642 Z_{t-2} + 0.0606 Z_{t-3} + 0.1655 Z_{t-4},$$ (53)

we obtain 36 residuals $\hat{e}_5, \hat{e}_6, \ldots, \hat{e}_{40}$ shown in Fig. 2. Thus, the expected value of estimated disturbance terms is

$$\hat{e} = \frac{1}{36} \sum_{t=5}^{40} \hat{e}_t = 0.0000,$$ (54)

and the variance is

$$\sigma^2 = \frac{1}{36} \sum_{t=5}^{40} (\hat{e}_t - \hat{e})^2 = 96.0254^2.$$ (55)

Assume the estimated disturbance terms follow the normal uncertainty distribution

$$\mathcal{N}(0.0000, 96.0254).$$ (56)

In order to test whether it is appropriate, given a significance level $\alpha = 0.01$, the uncertain hypothesis test for the hypotheses

$H_0 : e = 0.0000$ and $\sigma = 96.0254$ versus

$H_1 : e \neq 0.0000$ or $\sigma \neq 96.0254$

is

$$W = \{(w_5, w_6, \ldots, w_{40}) \mid -243.2729 \leq w_t \leq 243.2729,$$ (57)

and solving the minimization problem

we obtain a new fitted UAR(4) component

$$Z_t = -6.9372 + 0.8836 Z_{t-1} + 0.0762 Z_{t-2} - 0.1614 Z_{t-3} - 0.1083 Z_{t-4}.$$ (62)

From

$$\hat{e}_t = z_t + 6.9372 - 0.8836 Z_{t-1} - 0.0762 Z_{t-2} + 0.1614 Z_{t-3} + 0.1083 Z_{t-4},$$ (63)

we obtain 36 residuals $\hat{e}_5, \hat{e}_6, \ldots, \hat{e}_{40}$ shown in Fig. 3. Thus, the expected value of estimated disturbance terms is

$$\hat{e} = \frac{1}{36} \sum_{t=5}^{40} \hat{e}_t = 0.0002.$$ (64)
and the variance is

\[ \hat{\sigma}^2 = \frac{1}{36} \sum_{t=5}^{40} (\hat{\varepsilon}_t - \hat{\varepsilon})^2 = 78.4578^2. \] (65)

Assume the estimated disturbance terms follow the normal uncertainty distribution

\[ N(0.0002, 78.4578). \] (66)

In order to test whether it is appropriate, given a significance level \( \alpha = 0.01 \), the uncertain hypothesis test for the hypotheses

\[ H_0 : \varepsilon = 0.0002 \text{ and } \sigma = 78.4578 \text{ versus } H_1 : \varepsilon \neq 0.0002 \text{ or } \sigma \neq 78.4578 \]

is

\[ W = \left\{ (w_5, w_6, \ldots, w_{40}) \left| -198.7665 \leq w_t \leq 198.7669, t = 5, 6, \ldots, 40 \right. \right\}. \] (67)

Since \((\hat{\varepsilon}_5, \hat{\varepsilon}_6, \ldots, \hat{\varepsilon}_{40}) \in W\), we reject \( H_0 \). It follows from

\[ \hat{\varepsilon}_5 = -244.6483 < -198.7665 \] (68)

that \( z_5 \) is an outlier, and is replaced with

\[
\begin{align*}
z_5 &= z_4 + \frac{z_6 - z_4}{2} \\
&= 313.1674 + \frac{152.8045 - 313.1674}{2} \\
&= 232.9860.
\end{align*}
\] (69)

After repeating the iterative procedure 5 times, we obtain a new fitted UAR(4) component

\[ Z_t = -4.6259 + 1.2608Z_{t-1} - 0.2837Z_{t-2} \\
-0.2820Z_{t-3} + 0.0939Z_{t-4}. \] (70)

From

\[ \hat{\varepsilon}_t = z_t + 4.6259 - 1.2608z_{t-1} + 0.2837z_{t-2} \\
+0.2820z_{t-3} - 0.0939z_{t-4}, \] (71)

we obtain 36 residuals \( \hat{\varepsilon}_5, \hat{\varepsilon}_6, \ldots, \hat{\varepsilon}_{40} \) shown in Fig. 4. Thus, the expected value of estimated disturbance terms is

\[ \hat{\varepsilon} = \frac{1}{36} \sum_{t=5}^{40} \hat{\varepsilon}_t = 0.0000, \] (72)

and the variance is

\[ \hat{\sigma}^2 = \frac{1}{36} \sum_{t=5}^{40} (\hat{\varepsilon}_t - \hat{\varepsilon})^2 = 53.4133^2. \] (73)

Assume the estimated disturbance terms follow the normal uncertainty distribution

\[ N(0.0000, 53.4133). \] (74)

In order to test whether it is appropriate, given a significance level \( \alpha = 0.01 \), the uncertain hypothesis test for the hypotheses

\[ H_0 : \varepsilon = 0.0000 \text{ and } \sigma = 53.4133 \text{ versus } H_1 : \varepsilon \neq 0.0000 \text{ or } \sigma \neq 53.4133 \]
Then, the forecast value on day 41 (March 24, 2020) is

\[ \hat{Y}_{41} = \frac{80822}{1 + 0.3100 \exp(-0.1802 \times 41)} - 4.6259 \\
+1.2608z_{40} - 0.2837z_{39} - 0.2820z_{38} + 0.0939z_{37} \\
+\varepsilon \in N(0.0000, 53.4133). \]  \tag{76}

Then, the forecast value on day 41 (March 24, 2020) is

\[ \hat{Y}_{41} = \frac{80822}{1 + 0.3100 \exp(-0.1802 \times 41)} - 4.6259 \\
+1.2608z_{40} - 0.2837z_{39} - 0.2820z_{38} + 0.0939z_{37} \\
+0.0000 \\
= 80755, \]  \tag{77}

and the 95% confidence interval is

\[ [80647, 80862]. \]  \tag{78}

That is, we predict that the cumulative number on March 24, 2020 will be 80755, and we are 95% sure that the number falls into [80647, 80862].

### 4 Comparative analysis

In this section, we compare uncertain regression model with autoregressive time series errors with uncertain regression model (Liu 2021) and difference plus UAR model (Ye and Yang 2021). The estimated standard deviation (see Table 5)

obtained by the uncertain logistic growth model, 183.82, is too large to be accepted while the estimated standard deviation obtained by the uncertain logistic growth model with autoregressive time series errors, 53.4133, makes more sense. From Table 5, the uncertain logistic growth model with autoregressive time series errors can predict better. This model has less information loss, but more computation.

Ye and Yang (2021) modeled the second difference of the cumulative cases series using the UAR(5) model. However, so far, there are no definitions of stationary and difference in uncertain time series analysis. In Reference (Ye and Yang 2021), the differencing operation was based on stochastic time series analysis, and this methodology (i.e., difference plus UAR model) used for analyzing time series data was flawed. But in this paper, uncertainty theory supplies the theoretical justifications for the uncertain regression model with autoregressive time series errors, and this method can be used easily.

In 2019, Aslam and Albassam (2019) used the neutrosophic regression model to study the relationship between prostate cancer and dietary fat level. The essential difference between neutrosophic regression and uncertain regression lies in statistical techniques. The former uses neutrosophic statistics which was introduced based on the idea of neutrosophic logic, while the latter uses uncertain statistics which was introduced based on uncertainty theory. In other words, the difference between neutrosophic regression and uncertain regression is that the former deals with the data having Neutrosophy, inexact values, unclear observations, and interval values, while the latter deals with the data containing precise and exact observations. The former provides the parameters, confidence interval and \( p \)-values in the indeterminacy interval range, while the latter provides the determined values of all parameters.

### 5 Why is stochastic regression model with autoregressive time series errors not suitable for modeling cumulative number?

The difference between traditional regression model with autoregressive time series errors and uncertain regression model with autoregressive time series errors lies in how the disturbance terms are assumed. The former assumes the disturbance term is a stochastic variable, while the latter assumes the disturbance term is an uncertain variable. Since random

### Table 4 Results of data modification

| \( z_t \) | The original data | The revised data |
|---|---|---|
| \( z_5 \) | 60.3347 | 232.9860 |
| \( z_7 \) | 276.2352 | 157.9230 |
| \( z_{11} \) | -363.0434 | -125.2114 |
| \( z_{12} \) | -381.9831 | -295.0974 |
| \( z_{17} \) | 155.3367 | -12.1848 |

### Table 5 Effect of \( Z_t \)

| \( Z_t \) | Actual value | Forecast value | Standard deviation |
|---|---|---|---|
| \( \varepsilon \) | 80744 | 80807 | 183.82 |
| UAR(4) | 80744 | 80755 | 53.413 |

...
variables and uncertain variables obey different operational laws, wrong assumptions may mislead the decision-maker.

In the example of COVID-19, we use the Lilliefors test for testing the normality of the residuals (see Fig. 4). The test results show that the null hypothesis is rejected. That is, the disturbance term cannot be characterized as a normal random variable. Therefore, stochastic regression model with autoregressive time series errors is not suitable for modeling cumulative number.

6 Conclusions

This paper firstly proposed a new model, i.e., the uncertain regression model with autoregressive time series errors. Then, the principle of least squares was used to estimate the unknown parameters. Finally, we made a comparative analysis. The conclusion was that the uncertain regression model with autoregressive time series errors can improve the accuracy of predictions compared with the uncertain regression model.

In future research, we will investigate how to deal with imprecise observations using neutrosophic statistics or uncertain statistics. In addition, referring to goodness of fit test (Aslam 2021a), analysis of means (Aslam 2021c), and skewness and kurtosis estimators (Aslam 2021b), these techniques can be introduced into uncertain statistics as a future research endeavor.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grants No.61873329 and 61873084).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Data availability All data generated or analyzed during this study are included in Tables 1 and 2.

Code availability All the codes implemented during this study are available from the corresponding author on reasonable request.

Ethical approval This paper does not contain any studies with human participants or animals performed by any of the authors.

References

Aslam M (2021a) A new goodness of fit test in the presence of uncertain parameters. Complex Intell Syst 7(1):359–365
Aslam M (2021b) A study on skewness and kurtosis estimators of wind speed distribution under indeterminacy. Theo Appl Climatol 143(3–4):1227–1234
Aslam M (2021c) Analyzing wind power data using analysis of means under neutrosophic statistics. Soft Comput 25(10):7087–7093
Aslam M, Albassam M (2019) Application of neutrosophic logic to evaluate correlation between prostate cancer mortality and dietary fat assumption. Symmetry 11(3):330
Chen D (2020a) Tukeys biweight estimation for uncertain regression model with imprecise observations. Soft Comput 24(22):16803–16809
Chen D, Yang X (2021) Maximum likelihood estimation for uncertain autoregressive model with application to carbon dioxide emissions. J Intell Fuzzy Syst 40(1):1391–1399
Chen D, Yang X (2021) Ridge estimation for uncertain autoregressive model with imprecise observations. Int J Uncertain Fuzzy Knowl-Based Syst 29(1):37–55
Chen X, Li J, Xiao C, Yang P (2021) Numerical solution and parameter estimation for uncertain SIR model with application to COVID-19. Fuzzy Optim Decis Mak 20(2):189–208
Cochrane D, Orcutt GH (1949) Application of least squares regression to relationships containing autocorrelated error terms. J Am Stat Assoc 44(245):32–61
Durbin J (1960) Estimation of parameters in time-series regression models. J Royal Stat Soc Series B 22(1):139–153
Jia L, Chen W (2021) Uncertain SEIAR model for COVID-19 cases in China. Fuzzy Optim Decis Mak 20(2):243–259
Liu W, Liu B (2018) Residual and confidence interval for uncertain regression model with imprecise observations. J Intell Fuzzy Syst 35(2):2573–2583
Liu W, Liu B (2020a) Uncertain maximum likelihood estimation with application to uncertain regression analysis. Soft Comput 24(13):9351–9360
Liu W, Liu B (2021) Initial value estimation of uncertain differential equations and zero-day of COVID-19 spread in China. Fuzzy Optim Decis Mak 20(2):177–188
Liu B (2007) Uncertainty theory, 2nd edn. Springer-Verlag, Berlin
Liu S (2019) Leave-p-out cross-validation test for uncertain Verhulst-Pearl model with imprecise observations. IEEE Access 7:131705–131709
Liu Z (2021) Uncertain growth model for the cumulative number of COVID-19 infections in China, Fuzzy Optim Decis Mak 20(2):229–242
Liu Z (2021a) Generalized moment estimation for uncertain differential equations. Appl Math Comput 392:125724
Liu Z, Jia L (2020) Cross-validation for the uncertain Chapman-Richards growth model with imprecise observations. Int J Uncertain Fuzzy Knowl-Based Syst 28(5):769–783
Liu Z, Yang X (2020) Cross validation for uncertain autoregressive model. Commun Stat Simul Comput. https://doi.org/10.1080/03610918.2020.1747077
Liu Z, Yang Y (2020) Least absolute deviations estimation for uncertain regression with imprecise observations. Fuzzy Optim Decis Mak 19(1):33–52
Liu Y, Liu B (2020) Estimating unknown parameters in uncertain differential equation by maximum likelihood estimation, Technical Report
Sheng Y, Yao K, Chen X (2020) Least squares estimation in uncertain differential equations. IEEE Trans Fuzzy Syst 28(10):2651–2655
Song Y, Fu Z (2018) Uncertain multivariable regression model. Soft Comput 22(17):5861–5866
Tang H (2020) Uncertain vector autoregressive model with imprecise observations. Soft Comput 24(22):17001–17007
Yang X, Liu B (2019) Uncertain time series analysis with imprecise observations. Fuzzy Optim Decis Mak 18(3):263–278
Yang X, Ni Y (2021) Least-squares estimation for uncertain moving average model. Commun Stat Theory Methods 50(17):4134–4143
Yang X, Liu Y, Park G (2020) Parameter estimation of uncertain differential equation with application to financial market. Chaos Solitons Fract 139:110026

Springer
Yang X, Park G, Hu Y (2020) Least absolute deviations estimation for uncertain autoregressive model. Soft Comput 24(23):18211–18217

Yao K, Liu B (2018) Uncertain regression analysis: an approach for imprecise observations. Soft Comput 22(17):5579–5582

Yao K, Liu B (2020) Parameter estimation in uncertain differential equations. Fuzzy Optim Decis Mak 19(1):1–12

Ye T, Liu Y (2020) Multivariate uncertain regression model with imprecise observations. J Ambient Intell Human Comput 11(11):4941–4950

Ye T, Liu B (2021) Uncertain hypothesis test with application to uncertain regression analysis. Fuzzy Optim Decis Mak. https://doi.org/10.1007/s10700-021-09365-w

Ye T, Yang X (2021) Analysis and prediction of confirmed COVID-19 cases in China with uncertain time series. Fuzzy Optim Decis Mak 20(2):209–228

Zhang C, Liu Z, Liu J (2020) Least absolute deviations for uncertain multivariate regression model. Int J General Syst 49(4):449–465

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