Uncertain Box-Cox Regression Analysis With Rescaled Least Squares Estimation

SHIQIN LIU1, LIANG FANG2, ZAIYING ZHOU3, AND YIPING HONG4

1College of Mathematics and Computer Science, Hengshui University, Hengshui 053000, China
2School of Economics and Management, Beijing Forestry University, Beijing 100083, China
3Center for Statistical Science, Tsinghua University, Beijing 100084, China
4Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

Corresponding author: Liang Fang (tisnghuafl@163.com)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant BLX2019446, and in part by the National Science Foundation of China under Grant 61702165.

ABSTRACT
Under the uncertain statistical framework by Liu [19], there is still a lack of an effective fitting method for uncertain linear models with Box-Cox transformation of response variables. For example, for the transformation parameter λ, the uncertain least squares estimation will produce a severely low estimation result. In this paper, we propose uncertain Box-Cox regression analysis by utilizing the uncertainty theory to model the imprecise data and applying a generalized Box-Cox transformation indexed by its parameter to validate classic regression assumptions. We use rescaled least squares to estimate unknown parameters and provide an estimate for noises followed by residual analysis for these uncertain Box-Cox regression models. We also give the forecast values and confidence intervals and use a numerical example to demonstrate our methodology. Our work sets a uniformed framework for Box-Cox transformation on the uncertain regression, and extends such regression from linear to nonlinear cases, taking the Johnson-Schumacher growth model as an example.

INDEX TERMS
Uncertain box-cox regression model, rescaled least squares, residual analysis, uncertain box-cox linear regression model, uncertain box-cox Johnson-Schumacher growth model.

I. INTRODUCTION
Regression analysis is a useful tool for studying how a response variable is impacted by other explanatory variables. This mathematical model has become prevalent in all scientific branches. The essential part of the model is in the prescribed assumptions, which is, to be specific, that the response variable, also named as the dependent variable, is linearly dependent on the explanatory variables, or independent variables, or predictors. Such linear regression models are often helpful in either estimating the underlying relationships of these variables or predicting the responses given new values of predictors.

Classical statistical inferences mainly include estimation for unknown quantities and significance tests for hypotheses. The most seminal estimation methods are the least squares by Legendre [14] and Gauss [12], and the maximum likelihood by Wilks [30]. With data collected from the real world, it is not suitable for direct applications of statistical models due to the nature of imprecision and fuzziness [19]. To handle such imprecise observations or fuzzy observations, people at first turn to the fuzzy set theory. Very influential literature of fuzzy sets concerning this aspect includes Tanaka et al. [28], Corral and Gil [4], [5], Sakawa and Yano [26], and Casals et al. [2], [3], to name a few.

An important improvement over fuzzy variables was the uncertainty theory proposed by Liu [18], which targeted specifically at the uncertain data problems. Similar with the classic probability framework, the uncertainty theory views the imprecise observations as the uncertain variables, which follow various uncertainty distributions (Liu [15]). Uncertain statistics is a new axiom system established by Liu [19], which can be used to analyze the expert reliability data and other data types not suitable for classical statistical inference methods. Data envelopment analysis (DEA) is the field where many advances of the application of the uncertain theory happened. For instance, Wen et al. [29] proposed some new ranking criteria and the uncertain DEA model,
Lio and Liu [21] considered the imprecise input and output case, then Nejad and Ghaflari-Hadigheh [22] further refined the model so that the decision making unit could be evaluated efficiently. The important progress in uncertain statistical inference started from Yao [32]. Then Yao and Liu [33] extended the principle of least squares and thus proposed uncertain regression analysis. Lio and Liu [20] developed residual analysis as well as prediction for the regression models. As we all know that classic linear regression requires the normality assumption, which is often violated by real data, a Box-Cox transformation was then recommended by Box and Cox [1] as a pre-treatment, which has now become a popular and standard method in statistical inference. With the imprecise observations, Fang and Hong [9] adopted the uncertain regression analysis and considered three most frequently used transformations on the dependent variables. We notice that these three types of transformations, namely the logarithmic, square root and reciprocal transformations, can all be regarded as special cases of Box-Cox transformations, thus in this article, we generalize the results there to any Box-Cox transformation indexed by its parameter. Furthermore, we point out that the Box-Cox transformation could also be applied on some nonlinear models.

To be specific, we propose the generalized uncertain Box-Cox revised regression analysis to model the quantitative relationship between uncertain dependent variables after Box-Cox transformation and uncertain independent variable with imprecisely observations, which can relax the assumptions required by classic regression analysis. Since the Box-Cox transformation is applicable on both a linear model and a nonlinear model, such as the Johnson-Schumacher growth model, which was first noticed by Fang et al. [10], we also show the estimation of the relevant noises and describe how to perform residual analysis. Besides, estimation of forecast value and confidence interval will be explained in detail.

This paper for uncertain observed data is presented in the following order: The generalized uncertain Box-Cox regression models and uncertain Box-Cox Johnson-Schumacher growth models are defined in Section 2, followed by the rescaled least squares approach to the estimation problems. The residual analysis is handled in Section 3. The corresponding forecast values and confidence intervals are built in Section 4. For Section 5, we simulate relevant numerical examples to explain how our theories and methods work. Finally, we make some conclusion remarks in Section 6.

II. UNCERTAIN BOX-COX REGRESSION MODELS

In this section, the uncertain Box-Cox regression model will be established. The biggest difference between uncertainty analysis and classical statistics is the definition of independence between variables. A brief introduction to the uncertainty theory, including imprecise observations, uncertain variables, uncertain distributions, etc. (see chapter 2 in [19]). The vector \((x_1, x_2, \cdots, x_q)\) is made of uncertain explanatory variables, and the scalar \(y\) is the uncertain response variable. Following the traditional assumptions, we use function \(h\) to characterize how \(y\) is dependent on \((x_1, x_2, \cdots, x_p)\), then the uncertain Box-Cox regression model can be defined by

\[
L(y; \lambda) = h(x_1, x_2, \cdots, x_p; \beta) + \epsilon, \tag{1}
\]

where

\[
L(y; \lambda) = \begin{cases} 
\frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\
\log y, & \text{if } \lambda = 0 
\end{cases}
\]

is the Box-Cox transformation with respect to the transformation parameter \(\lambda\). Here \(h\) is a given function, vector \(\beta\) contains the regression coefficients and \(\epsilon\) represents the white noise, i.e., noise with zero mean and finite variance, as in the classic regression settings only that the variables are uncertain variables instead of random variables. When

\[
h(x_1, x_2, \cdots, x_q; \beta) = \beta_0 + \sum_{k=1}^{q} x_k \beta_k,
\]

this model is the uncertain Box-Cox linear regression model. This model with a transformation to the response \(y\) offers a more flexible parametrization of the relationship between the response and the covariates, compared to the uncertain regression model

\[
y = h(x_1, x_2, \cdots, x_q; \beta) + \epsilon
\]

proposed by Lio and Liu [20].

Fang and Hong [9] considered the special case of model (1) where \(\lambda\) is predetermined, such as \(\lambda = -1, 0\) or 0.5. For imprecise observations

\[
(\tilde{x}_{j1}, \tilde{x}_{j2}, \cdots, \tilde{x}_{jq}, \tilde{y}_j), \quad j = 1, 2, \cdots, n,
\]

assuming that \(\tilde{x}_{j1}, \tilde{x}_{j2}, \cdots, \tilde{x}_{jq}, \tilde{y}_j\) are uncertain variables following uncertainty distributions \(\Phi_{j1}, \Phi_{j2}, \cdots, \Phi_{jq}, \Psi_j, j = 1, 2, \cdots, n\), respectively, they proposed

\[
\beta^* = \arg \min_{\beta} \sum_{j=1}^{n} E[L(\tilde{y}_j; \lambda) - h(\tilde{x}_{j1}, \tilde{x}_{j2}, \cdots, \tilde{x}_{jq}; \beta)]^2 \\
\Rightarrow \arg \min_{\beta} L_n(\beta, \lambda), \tag{2}
\]

by generalizing the least square estimation (LSE) of Yao and Liu [33]. However, when \(\lambda\) is unknown, the LSE easily fails. For instance, in the uncertain Box-Cox linear regression model (1) where

\[
h(\tilde{x}_{j1}, \tilde{x}_{j2}, \cdots, \tilde{x}_{jq}; \beta) = \sum_{k=1}^{q} \tilde{x}_{jk} \beta_k \defeq \tilde{x}_j^t \beta,
\]

we have

\[
\lim_{\lambda \to -\infty} \sum_{j=1}^{n} E[L(\tilde{y}_j; \lambda) - \tilde{x}_j^t \beta]^2 = 0 \tag{3}
\]

for \(\beta = 0\) provided that \(\tilde{y}_j > 1\) for all \(j\). Since if \(y > 1\), then \(L(y, \lambda) \to 0\) when \(\lambda \to -\infty\). In other words, the minimum point of \(L(\beta, \lambda)\) is \((\beta^*, \lambda^*) = (0, -\infty)\), which is not a reasonable estimate.
Powell [25] pointed out a similar problem occurs for applying a generalized method-of-moments (GMM) estimator for the Box-Cox regression model in the statistical framework. Therefore, a rescaled method-of-moments estimation was introduced there for a Box-Cox regression model by replacing the objective function $Q_n(\beta, \lambda)$ with

$$Q_n(\beta, \lambda) = S_n(\beta, \lambda)(\hat{y})^{-2\lambda},$$

where

$$\hat{y} = \exp \left\{ \frac{1}{n} \sum_{j=1}^{n} \log |y_j| \right\} := \hat{y}_n,$$

being the geometric mean of the absolute values of observations $y_j, j = 1, 2, \ldots, n$ in the statistical sense, $Q_n$ and $S_n$ are defined in Powell [25]. Similarly, in the framework of uncertain variables, we consider rescaling $L_n(\beta, \lambda)$ in (2) to get the new objective function

$$R_n(\beta, \lambda) = L_n(\beta, \lambda)(\hat{y})^{-2\lambda},$$

where

$$\hat{y} = \exp \left\{ \frac{1}{n} \sum_{j=1}^{n} E(\log |y_j|) \right\}$$

being the result of replacing log $|y_j|$ in (4) by uncertain observations, taking logarithmic and expectation. Hereby, we propose the rescaled least squares estimation (RSLSE) for the parameters $(\beta, \lambda)$ within the uncertain Box-Cox regression model (1) by

$$(\beta^*, \lambda^*) = \arg \min_{\beta, \lambda} R_n(\beta, \lambda),$$

or equivalently,

$$(\beta^*, \lambda^*) = \arg \min_{\beta, \lambda} \log R_n(\beta, \lambda)$$

$$= \arg \min_{\beta, \lambda} \left\{ \log (T_n(\beta, \lambda)) - \frac{2\lambda}{n} \sum_{j=1}^{n} E(\log |y_j|) \right\}. \tag{5}$$

Here

$$L_n(\beta, \lambda) = \sum_{j=1}^{n} E[L(\tilde{y}_j; \lambda) - h(\tilde{x}_j, \tilde{x}_2, \ldots, \tilde{x}_q; \beta)]^2.$$ 

The fitted model is then denoted by

$$L(y; \lambda^*) = h(x_1, x_2, \ldots, x_q; \beta^*).$$

The expectations in (5) can be computed using the inverse uncertainty distributions of the datasets, with the help of Theorem 2.15 [17]. The following Theorem 1 reveal the explicit forms of the log-objective function log $R_n(\beta, \lambda)$ with respect to the inverse uncertainty distribution for two special models corresponding to different $f(x_1, x_2, \ldots, x_q; \beta)$, i.e., the linear and the Johnson-Schumacher growth model. Note that we assume $\tilde{x}_j, \tilde{x}_2, \ldots, \tilde{x}_q, \tilde{y}$ are independent uncertain variables for every $j = 1, \ldots, n$ in the following theorems.

**Theorem 1:** For imprecise observations $(\tilde{x}_j, \tilde{x}_2, \ldots, \tilde{x}_q, \tilde{y}), j = 1, 2, \ldots, n$, where $\tilde{x}_j, \tilde{x}_2, \ldots, \tilde{x}_q, \tilde{y}$ are independent uncertain variables with regular uncertainty distributions $\Phi_{j1}, \Phi_{j2}, \ldots, \Phi_{jq}, \Psi_j, j = 1, 2, \ldots, n$, respectively, the RLS of $\beta_0, \beta_1, \ldots, \beta_q$ in the uncertain Box-Cox linear model

$$L(y; \lambda) = \beta_0 + \sum_{k=1}^{q} \beta_k x_k + \epsilon$$

is the solution to the optimization formula:

$$\min_{\beta_0, \beta_1, \ldots, \beta_q, \lambda} \left\{ \frac{n}{2} \log \sum_{j=1}^{n} (A_{\alpha}) - \lambda \sum_{j=1}^{n} (B_{\alpha}) \right\}, \tag{6}$$

where

$$A_{\alpha} = \int_{0}^{1} \left( L(\Psi_j^{-1}(\alpha); \lambda) - \beta_0 - \sum_{k=1}^{q} \beta_k \Upsilon_j^{-1}(\alpha, \beta_k) \right)^2 d\alpha,$$

$$B_{\alpha} = \int_{0}^{1} \log(\Psi^{-1}(\alpha)) d\alpha.$$ 

Here

$$\Upsilon_j^{-1}(\alpha, \beta_k) = \begin{cases} \Phi_{jk}^{-1}(1 - \alpha), & \text{if } \beta_k \geq 0 \\
\Phi_{jk}^{-1}(\alpha), & \text{if } \beta_k < 0 \end{cases}$$

for $j = 1, 2, \ldots, n$, and $k = 1, 2, \ldots, q$, here $\epsilon$ is the white noise.

**Proof:** The RLS of $\beta_0, \beta_1, \ldots, \beta_q, \lambda$ in the Box-Cox linear model can be viewed as the solution to the optimization problem

$$\min_{\beta, \lambda} \left\{ \frac{n}{2} \left\{ \log \sum_{j=1}^{n} E(\beta, \lambda) - \lambda \sum_{j=1}^{n} E[\log(\tilde{y}_j)] \right\} \right\}. \tag{7}$$

Here

$$E(\beta, \lambda) = E \left[ L(\tilde{y}_j; \lambda) - \beta_0 - \sum_{k=1}^{q} \beta_k \tilde{x}_k \right]^2.$$ 

For any given $j$, it follows from Theorem 2.15 [17] that the inverse uncertainty distribution of

$$L(\tilde{y}_j; \lambda) - \beta_0 - \sum_{k=1}^{q} \beta_k \tilde{x}_k$$

is

$$F_j^{-1}(\alpha) = L(\Psi_j^{-1}(\alpha); \lambda) - \beta_0 - \sum_{k=1}^{q} \beta_k \Upsilon_j^{-1}(\alpha, \beta_k).$$

Then from equations (2.163) and (2.206) [17], we obtain

$$E[\log(\tilde{y}_j)] = \int_{0}^{1} \log(\Psi^{-1}(\alpha)) d\alpha,$$

$$E \left( \left( L(\tilde{y}_j; \lambda) - \beta_0 - \sum_{k=1}^{q} \beta_k \tilde{x}_k \right)^2 \right) = \int_{0}^{1} \left( L(\Psi_j^{-1}(\alpha); \lambda) - \beta_0 - \sum_{k=1}^{q} \beta_k \Upsilon_j^{-1}(\alpha, \beta_k) \right)^2 d\alpha.$$
Hence, the minimization formula (7) becomes
\[
\min_{\beta_0, \beta_1, \ldots, \beta_q, \lambda} \left\{ \frac{n}{2} \log \sum_{j=1}^{n} (A_d) - \lambda \sum_{j=1}^{n} (B_d) \right\}.
\]
Thus the above theorem is proved.

**Theorem 2**: For imprecise observations \((\tilde{x}_j, \tilde{y}_j),\ j = 1, 2, \ldots, n\), where \(\tilde{x}_j, \tilde{y}_j\) are independent uncertain variables with regular uncertainty distributions \(\Phi_j, \Psi_j, j = 1, 2, \ldots, n\), respectively, the RLSE of \(\beta_0, \beta_1, \beta_2\) in the uncertain Box-Cox Johnson-Schumacher growth model
\[
L(y; \lambda) = \beta_0 \exp \left( -\frac{\beta_1}{x + \beta_2} \right) + \epsilon
\]
is the solution to the optimization formula:
\[
\min_{\beta_0, \beta_1, \beta_2, \lambda} \left\{ \frac{n}{2} \log \sum_{j=1}^{n} (C_d) - \lambda \sum_{j=1}^{n} (D_d) \right\}.
\]
Here
\[
C_d = \int_{0}^{1} \left( L(\Phi_j^{-1}(\alpha); \lambda) - \beta_0 \exp \left( -\frac{\beta_1}{\Phi_j^{-1} (1 - \alpha) + \beta_2} \right) \right)^2 d\alpha,
\]
\[
D_d = \int_{0}^{1} \log(\Psi_j^{-1}(\alpha)) d\alpha,
\]
and \(\epsilon\) is the white noise.

**Proof**: The RLSE of \(\beta_0, \beta_1, \ldots, \beta_p, \lambda\) in the Box-Cox linear model can be considered as the solution to the optimization formula
\[
\min_{\beta, \lambda} \left\{ \frac{n}{2} \log \sum_{j=1}^{n} E_{\beta, \lambda}^\prime \right\} - \lambda \sum_{j=1}^{n} E[\log(\tilde{y}_j)]\right\).
\]
Here
\[
E_{\beta, \lambda}^\prime = \mathbb{E}\left[ L(\tilde{y}_j; \lambda) - \beta_0 \exp \left( -\frac{\beta_1}{\tilde{x}_j + \beta_2} \right) \right]^2.
\]
For any given \(j\), it can follow from Theorem 2.15 [17] that the inverse uncertainty distribution of
\[
L(\tilde{x}_j; \lambda) - \beta_0 \exp \left( -\frac{\beta_1}{\tilde{x}_j + \beta_2} \right)
\]
is
\[
F_j^{-1}(\alpha) = L(\Phi_j^{-1}(\alpha); \lambda) - \beta_0 \exp \left( -\frac{\beta_1}{\Phi_j^{-1} (1 - \alpha) + \beta_2} \right).
\]
Then from equations (2.163) and (2.206) [17], we obtain
\[
E[\log(\tilde{y}_j)] = \int_{0}^{1} \log(\Psi_j^{-1}(\alpha)) d\alpha,
\]
\[
E\left[ L(\tilde{y}_j; \lambda) - \beta_0 \exp \left( -\frac{\beta_1}{\tilde{x}_j + \beta_2} \right) \right]^2
\]
\[
= \int_{0}^{1} \left( L(\Phi_j^{-1}(\alpha); \lambda) - \beta_0 \exp \left( -\frac{\beta_1}{\Phi_j^{-1} (1 - \alpha) + \beta_2} \right) \right)^2 d\alpha.
\]
So (9) is equivalent to
\[
\min_{\beta_0, \beta_1, \beta_2, \lambda} \left\{ \frac{n}{2} \log \sum_{j=1}^{n} (C_d) - \lambda \sum_{j=1}^{n} (D_d) \right\}.
\]
Thus the above theorem is proved.

### III. NOISE ESTIMATION AND RESIDUAL ANALYSIS

This section focuses on the noise \(\epsilon\) in the uncertain Box-Cox regression model (1), linear or nonlinear. We propose an estimation of the variance for this term and perform residual analysis by estimating the mean and investigating the residual plot. For the imprecise observations
\[
(\tilde{x}_{j1}, \tilde{x}_{j2}, \ldots, \tilde{x}_{jq}, \tilde{y}_j), \ j = 1, 2, \ldots, n
\]
and the parameter estimation \((\beta^*, \lambda^*)\), the unobserved noise \(\epsilon\) can be approximately estimated by the residuals defined by the differences of \(L(\tilde{y}_j; \lambda^*)\) and \(h(\tilde{x}_{j1}, \tilde{x}_{j2}, \ldots, \tilde{x}_{jq}; \beta^*)\). For this case, we obtain the following definition.

**Definition 1**: If \((\tilde{x}_{j1}, \tilde{x}_{j2}, \ldots, \tilde{x}_{jq}, \tilde{y}_j), \ j = 1, 2, \ldots, n\) is a set of imprecise observations, and the fitted regression model is
\[
L(\tilde{y}; \lambda^*) = h(x_1, x_2, \ldots, x_q; \beta^*),
\]
then the \(j\)-th residual is defined by
\[
\hat{\epsilon}_j = L(\tilde{y}_j; \lambda^*) - h(\tilde{x}_{j1}, \tilde{x}_{j2}, \ldots, \tilde{x}_{jq}; \beta^*),
\]
where \(j = 1, \ldots, n\).

For the estimation of the moments of \(\epsilon\), we have
\[
\hat{\epsilon} = \frac{1}{n} \sum_{j=1}^{n} E[\hat{\epsilon}_j],
\]
that is, the expected noise can naturally be estimated by the average of the expected residuals, and the variance of \(\epsilon\)
\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} E[(\hat{\epsilon}_j - \hat{\epsilon})^2],
\]
where \(\hat{\epsilon}_j\) is the \(j\)-th residual, \(j = 1, 2, \ldots, n\), respectively. Note that the assumption \(E(\epsilon) = 0\) can be validated by the estimate \(\hat{\epsilon}\).

Next we give the following theorems on computing estimations \(\hat{\epsilon}\) and \(\hat{\sigma}^2\) for two special cases of the uncertain Box-Cox regression models, namely the linear and the Johnson-Schumacher growth model.

**Theorem 3**: For imprecise observations \((\tilde{x}_{j1}, \tilde{x}_{j2}, \ldots, \tilde{x}_{jq}, \tilde{y}_j), j = 1, 2, \ldots, n\), where \(\tilde{x}_{j1}, \tilde{x}_{j2}, \ldots, \tilde{x}_{jq}, \tilde{y}_j\) are independent uncertain variables following uncertainty distributions \(\Phi_{j1}, \Phi_{j2}, \ldots, \Phi_{jq}, \Psi_j, j = 1, 2, \ldots, n\), respectively, the fitted uncertain Box-Cox linear regression model is
\[
L(y; \lambda^*) = \beta_0^* + \sum_{k=1}^{q} \beta_k^* x_k.
\]
Then the estimated expected value of the noise $\epsilon$ is

$$
\hat{\epsilon} = \frac{1}{n} \sum_{j=1}^{n} \int_{0}^{1} (L_\alpha) \, d\alpha
$$

and the estimated variance is

$$
\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} \int_{0}^{1} (L_\alpha)^2 \, d\alpha.
$$

where

$$
L_\alpha = L(\Psi^{-1}_j(\alpha); \lambda^*) - \beta_0^* - \sum_{k=1}^{q} \beta_k^* \Upsilon_{jk}^{-1}(\alpha, \beta_k^*) - \hat{\epsilon},
$$

$$
\Upsilon_{jk}^{-1}(\alpha, \beta_k^*) = \left\{ \begin{array}{ll}
\Phi_{jk}^{-1}(1-\alpha), & \text{if } \beta_k^* \geq 0 \\
\Phi_{jk}^{-1}(\alpha), & \text{if } \beta_k^* < 0
\end{array} \right.
$$

for $j = 1, 2, \cdots, n$ and $k = 1, 2, \cdots, q$.

**Proof:** For any given $j$, it can follow from Theorem 2.15 [17] that the inverse uncertainty distribution of

$$
L(\tilde{\gamma}_j; \lambda^*) - \beta_0^* - \sum_{k=1}^{q} \beta_k^* \tilde{x}_{jk}
$$

is

$$
F_j^{-1}(\alpha) = L(\Psi_j^{-1}(\alpha); \lambda^*) - \beta_0^* - \sum_{k=1}^{q} \beta_k^* \Upsilon_{jk}^{-1}(\alpha, \beta_k^*).
$$

The theorem can follow from equations (2.163) and (2.206) [17] immediately.

**Theorem 4:** For imprecise observations $(\tilde{x}_j, \tilde{y}_j)$, $j = 1, 2, \cdots, n$, where $\tilde{x}_j, \tilde{y}_j$ are independent uncertain variables following uncertainty distributions $\Phi_j, \Psi_j$, $j = 1, 2, \cdots, n$, respectively, the fitted uncertain Box-Cox Johnson-Schumacher growth model is

$$
L(y; \lambda^*) = \beta_0^* \exp \left( -\frac{\beta_1^*}{x + \beta_2^*} \right).
$$

Then the estimated expected value of the noise $\epsilon$ is

$$
\hat{\epsilon} = \frac{1}{n} \sum_{j=1}^{n} \int_{0}^{1} (L'_\alpha) \, d\alpha
$$

and the estimated variance is

$$
\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} \int_{0}^{1} (L'_\alpha)^2 \, d\alpha.
$$

where

$$
L'_\alpha = L(\Psi_j^{-1}(\alpha); \lambda^*) - \beta_0^* \exp \left( -\frac{\beta_1^*}{\Phi_j^{-1}(1-\alpha) + \beta_2^*} \right).
$$

**Proof:** For any given $j$, it can follow from Theorem 2.15 [17] that the inverse uncertainty distribution of

$$
L(\tilde{\gamma}_j; \lambda^*) = \beta_0^* \exp \left( -\frac{\beta_1^*}{\tilde{x}_j + \beta_2^*} \right)
$$

is

$$
F_j^{-1}(\alpha) = L(\Psi_j^{-1}(\alpha); \lambda^*) - \beta_0^* \exp \left( -\frac{\beta_1^*}{\Phi_j^{-1}(1-\alpha) + \beta_2^*} \right).
$$

The theorem can follow from equations (2.163) and (2.206) [17] immediately.

We conclude this section by proposing a residual plot, which is an intuitive assessment of how well the model fits the data. This plot can also verify the homoskedasticity assumption, namely, the variance of the noise $\epsilon$ is irrelevant with the fitted uncertain variable $h(x_1, x_2, \cdots, x_q|\beta^*)$.

Similar to the residual plot in statistics, we plot the residuals $\hat{\epsilon}_i$ on the vertical axis and the fitted variables $h(\tilde{x}_{j1}, \tilde{x}_{j2}, \cdots, \tilde{x}_{jq}|\beta^*)$ on the horizontal axis. Unlike the single plot in statistics, where the residuals and fitted values are real numbers, here both of them are uncertain variables with distributions, therefore a more complicated visual result is need. For observation $i$, first we plot the mean value of the residuals $E(\hat{\epsilon}_j)$ against the mean value of the fitted variables $E(h(\tilde{x}_{j1}, \tilde{x}_{j2}, \cdots, \tilde{x}_{jq}|\beta^*))$, then we add a rectangle around this point to show the bound of this pair of uncertain variables. The bottom left corner of a rectangle shows the lower bounds of the residual and the corresponding fitted variable, while the upper right corner of the rectangle shows their upper bounds.

![Residual Plot](image)

**Figure 1.** Residual plot of the uncertain Box-Cox linear regression model for the numerical example.

Two examples of the residual plot are shown in Figure 1 and Figure 2. The plot shows that the residuals are evenly distributed around zero without clear patterns, which suggests that our model fits the data well and the homoskedasticity assumption holds.

**IV. FORECASTING AND INTERVAL ESTIMATION**

Using vector $(\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_q)$ to denote the new uncertain explanatory variables, with each following uncertainty distribution $\Phi_1, \Phi_2, \cdots, \Phi_q$, respectively, the goal is often to predict the corresponding response variable. A typical solution is via the fitted linear model

$$
L(y; \lambda^*) = \beta_0^* + \sum_{k=1}^{q} \beta_k^* x_k
$$
where the noise $\epsilon$ has estimated expected value $\hat{\epsilon}$ and variance $\hat{\sigma}^2$, and is independent of $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_q$. The forecast uncertain variable of $y$ with respect to $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_q$ becomes

$$L(\hat{y}; \lambda^*) = \beta_0^* + \sum_{k=1}^{q} \beta_k^* \tilde{x}_k + \hat{\epsilon}.$$ 

If normality for the noise $\hat{\epsilon}$ is further assumed, then the inverse uncertainty distribution of $\hat{\epsilon}$ is

$$\hat{\Psi}^{-1}(\alpha) = L^{-1} \left[ \beta_0^* + \sum_{j=1}^{p} \beta_j^* \lambda_j^{-1}(\alpha, \beta_j^*) + \Phi^{-1}(\alpha); \lambda^* \right],$$

where $\Phi^{-1}(\alpha)$ is the inverse uncertainty distribution of $\epsilon$, i.e.,

$$\Phi^{-1}(\alpha) = \hat{\epsilon} + \frac{\hat{\sigma} \sqrt{3}}{\pi} \log \frac{\alpha}{1-\alpha},$$

and

$$\lambda_j^{-1}(\alpha, \beta_j^*) = \begin{cases} \Phi_j^{-1}(\alpha), & \text{if } \beta_j^* \geq 0 \\ \Phi_j^{-1}(1 - \alpha), & \text{if } \beta_j^* < 0 \end{cases}$$

for $k = 1, 2, \ldots, q$. The uncertainty distribution for $\hat{\epsilon}$, i.e., $\hat{\Psi}$, can thus be derived from $\hat{\Psi}^{-1}$. And we have the forecast value of $y$ as

$$\mu = \int_0^1 \hat{\Psi}^{-1}(\alpha) d\alpha.$$ 

Notice that here forecast value $\mu$ is the point estimate of $y$. When the interval estimation is required, the confidence interval with a specified confidence level $\alpha$, for instance, 95%, is

$$[\hat{\Psi}^{-1}(b), \hat{\Psi}^{-1}(a + b)],$$

where $b$ can be chosen satisfying $b \in [0, 1-\alpha]$. Following the classic statistical inference, $b$ usually results in the interval

$$\hat{\Psi}^{-1}(a + b) - \hat{\Psi}^{-1}(b)$$

with the minimum length. We illustrate this in the next section.

### V. RELEVANT NUMERICAL EXAMPLES

In this section, we will fit the uncertain Box-Cox Johnson-Schumacher growth model and the Box-Cox linear model for two imprecise observation sets, providing the forecast value and the confidence interval for a given new uncertain explanatory vector, respectively.

First we consider the uncertain Box-Cox linear regression model. The imprecisely observed data $(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{y}_j)$, where $j = 1, \ldots, n$ with $n = 24$, is seen as in Fang [9] and they are independent uncertain variables following the linear uncertain distributions $\Phi_j, \Phi_{j_2}, \Phi_{j_3}, \Psi_j$, respectively. True $\beta$ is set as $(2,1,0,-0.5)$ and $\lambda$ 0.5.

Firstly, we fit the following uncertain Box-Cox linear regression model

$$L(y; \lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$ 

The rescaled least squares estimation of $\beta = (\beta_0, \ldots, \beta_3)$ and $\lambda$ can be calculated from Theorem 1. The result is

$$\beta^* = (\beta_0^*, \beta_1^*, \beta_2^*, \beta_3^*) = (2.4892, 0.8741, 0.0185, -0.4453), \lambda^* = 0.5422.$$ 

We have the fitted model

$$L(y; 0.5422) = 2.4892 + 0.8741 x_1 + 0.0185 x_2 - 0.4453 x_3.$$ 

Then we apply the residual analysis. According to Theorem 3, the estimated expected noise of $\epsilon$ is $\hat{\epsilon} = 0$ with the variance $\hat{\sigma}^2 = 1.4711$. We can see here the mean of $\epsilon$ is almost zero, which validates the zero-mean assumption.

Figure 1 is the plot of the residual versus fitted variables. We can see that the points are evenly distribute around the vertical line of zero without outliers, and the rectangles do not have clear patterns, which supports the homoskedasticity assumption very well.

Then we carry out prediction. For a new uncertain explanatory vector

$$(\tilde{x}_1', \tilde{x}_2', \tilde{x}_3') \sim (L(5, 6), L(28, 30), L(6, 7)),$$

assuming that $\tilde{x}_1', \tilde{x}_2', \tilde{x}_3', \epsilon$ are independent, the forecast uncertain variable $\hat{y}$ should satisfy

$$L(\hat{y}; 0.5422) = 2.4892 + 0.8741 \tilde{x}_1' + 0.0185 \tilde{x}_2' - 0.4453 \tilde{x}_3' + \epsilon,$$ 

where $\epsilon \sim \mathcal{N}(0.0000, 1.4711)$. Therefore for $y$, the forecast value is

$$\mu = 22.0381,$$

and the 95% confidence interval

$$[10.5424.34.2665].$$

Secondly, we consider the uncertain Box-Cox Johnson-Schumacher growth model. The imprecisely observations $(\tilde{x}_j, \tilde{y}_j)$, where $j = 1, \ldots, n$ with $n = 30$ and $\tilde{x}_j, \tilde{y}_j$ being independent uncertain variables following linear uncertain
TABLE 1. Imprecisely observed data for the uncertain Box-Cox Johnson-Schumacher growth model, where $L$ indicates linear uncertain variable.

| $j$ | 1   | 2   | 3   | 4   | 5   | 6   |
|-----|-----|-----|-----|-----|-----|-----|
| $x_j$ | $L(5.7, 7.7)$ | $L(6.1, 8.1)$ | $L(6.5, 8.5)$ | $L(7.8, 9.8)$ | $L(8.6, 10.6)$ | $L(13.3, 15.3)$ |
| $y_j$ | $L(18.5, 20.5)$ | $L(21.4, 23.4)$ | $L(18.8, 20.8)$ | $L(22.0, 24.0)$ | $L(29.3, 31.3)$ | $L(37.9, 39.9)$ |
| $j$ | 7   | 8   | 9   | 10  | 11  | 12  |
| $x_j$ | $L(14.8, 16.8)$ | $L(16.6, 18.6)$ | $L(17.3, 19.3)$ | $L(17.3, 19.3)$ | $L(17.7, 19.7)$ | $L(18.2, 20.2)$ |
| $y_j$ | $L(48.3, 50.3)$ | $L(52.0, 54.0)$ | $L(55.2, 57.2)$ | $L(53.1, 55.1)$ | $L(57.9, 59.9)$ | $L(54.1, 56.1)$ |
| $j$ | 13  | 14  | 15  | 16  | 17  | 18  |
| $x_j$ | $L(21.2, 23.2)$ | $L(22.5, 24.5)$ | $L(23.3, 25.3)$ | $L(23.8, 25.8)$ | $L(25.7, 27.7)$ | $L(27.4, 29.4)$ |
| $y_j$ | $L(60.8, 62.8)$ | $L(63.9, 65.9)$ | $L(67.7, 69.7)$ | $L(68.9, 70.9)$ | $L(57.2, 59.2)$ | $L(67.4, 69.4)$ |
| $j$ | 19  | 20  | 21  | 22  | 23  | 24  |
| $x_j$ | $L(28.3, 30.3)$ | $L(28.8, 30.8)$ | $L(34.0, 36.0)$ | $L(34.9, 36.9)$ | $L(39.5, 41.5)$ | $L(43.2, 45.2)$ |
| $y_j$ | $L(68.2, 70.2)$ | $L(66.9, 68.9)$ | $L(74.7, 76.7)$ | $L(67.8, 69.8)$ | $L(84.6, 86.6)$ | $L(80.0, 82.0)$ |
| $j$ | 25  | 26  | 27  | 28  | 29  | 30  |
| $x_j$ | $L(44.5, 46.5)$ | $L(45.6, 47.6)$ | $L(47.3, 49.3)$ | $L(47.5, 49.5)$ | $L(53.5, 55.5)$ | $L(57.4, 59.4)$ |
| $y_j$ | $L(79.5, 81.5)$ | $L(84.6, 86.6)$ | $L(83.2, 85.2)$ | $L(99.0, 101.0)$ | $L(95.1, 97.1)$ | $L(92.8, 94.8)$ |

The rescaled least squares estimation of $\beta = (\beta_0, \beta_1, \beta_2)$ and $\lambda$ can be calculated by Theorem 2. As a result, $\beta^* = (\beta_0^*, \beta_1^*, \beta_2^*) = (27.2963, 11.3902, 2.5476)$, $\lambda^* = 0.5872$.

We have the fitted model $L(y; 0.5872) = 27.2963 \exp \left( \frac{-11.3902}{x + 2.5476} \right)$.

Next we perform residual analysis. According to Theorem 4, the estimated expected noise $\hat{\epsilon}$ is $\hat{\epsilon} = 0.0003$ with the estimated variance $\hat{\sigma}^2 = 0.7408$. We can see here the mean of $\epsilon$ is almost zero, which validates the zero-mean assumption.

Figure 2 is the plot of the residual versus fitted variables. Again, we see that the points are evenly distributed around the vertical line of zero without outliers, and the rectangles do not have clear patterns, which supports the homoskedasticity assumption very well.

Now suppose $\tilde{x} \sim L(11, 13)$ is a new uncertain explanatory vector. Since $\tilde{x}$ is independent of $\epsilon$, the forecast uncertain variable $\hat{y}$ should satisfy

$L(\hat{y}; 0.5872) = 27.2963 \exp \left( \frac{-11.3902}{\tilde{x} + 2.5476} \right) + \epsilon$, (12)

where $\epsilon \sim N(0.0003, 0.7408)$. Therefore, the forecast value of $y$ is $\mu = 36.9323$, with the 95% confidence interval $[32.0021, 41.7438]$.

Note that the minimization of the interval length (10) is computed by the function optimize, where the tolerance accuracy is $10^{-8}$ and the numerical integration is calculated by the function integrate, both with R software.

VI. CONCLUSION AND DISCUSSION

We introduce a generalized uncertain Box-Cox linear regression model for the uncertain observed data, which can be adapted for both linear and nonlinear regression model. The unknown parameters are estimated via the rescaled least squares method which overcomes the problem that the original least squares estimation may fail to estimate $\lambda$, caused by the limiting behavior (3). We then perform residual analysis and carry out prediction and interval estimation for the dependent variables and new independent variables. Finally, we provide relevant numerical examples to illustrate how our theories work. Our proposed methods are suitable for fitting the uncertain dataset, such as the expert reliability data, under the uncertain statistical framework. In the future work, we may make use of standardized residuals in addition to the raw residuals. Besides, we will also consider introducing a penalized term in the uncertain maximum likelihood estimation.

ACKNOWLEDGMENT

(Shiqin Liu, Liang Fang, Zaiying Zhou, and Yiping Hong contributed equally to this work.) The authors would like to thank the Laboratory of uncertainty theory at Tsinghua University.

REFERENCES

[1] G. E. P. Box and D. R. Cox, “An analysis of transformations,” J. Roy. Statist. Soc. B (Methodol.), vol. 26, no. 2, pp. 211–252, 1964.
