Journal of Nonlinear and Convex Analysis
Volume 26, Number 6, 2025, 1899–1908

Yokohama Publishers
ISSN 1880-5221 ONLINE JOURNAL

PREDICTIVE WEIGHTED LINEAR REGRESSION MODEL AND ECONOMIC APPLICATION

JUNFU CUI, YICONG HE, SU GUO, YANG LIU, AND MIAO HAO*

Abstract. Linear regression model is widely used in various economic and social fields because of its simple operation, intuitive nature, and powerful explanatory power. The primary purpose of most models is prediction. The linear regression model is mainly used for internal prediction rather than external prediction. In this paper, according to partial modeling idea, with the help of bootstrap method to expand the sample, the predictive effect is used to weight and establish the predictive weighted linear regression model. The predictive ability of the model is evaluated using the normalized mean square error ($NMSE$), and the predictive weighted linear regression model exhibits a smaller $NMSE$ and better predictive ability. In conclusion, this study is an attempt to improve the external predictive power of linear regression model to some extent. Moreover, this partial modeling idea can also be applied to nonlinear modeling and widely extended.

1. Introduction

Linear regression model is the most widely used statistical model. The most important role of modeling is used for prediction, for a certain variable y, the actual value of the variable is y_i, but the actual value may be difficult to obtain for a while, so a statistical model can be constructed to get the predicted value $\hat{y}_i$, and use the predicted value to carry out research [6]. Depending on the scope of prediction, prediction can be categorized into internal prediction (interpolation) and external prediction (extrapolation). The linear regression model is mainly used for internal prediction, while it may be difficult to ensure the accuracy for external prediction [13]. As shown in Figure 1, the linear regression model l can predict the data points relatively well when $x_1 \leq x \leq x_2$, while the prediction of the linear regression model l will be invalid when $x_2 \leq x \leq x_3$.

This occurs mainly because the modeling starting point of the current linear regression model is based on the holistic modeling idea, and there are some limitations of this modeling idea. The linear regression model is of the following form:

$$Y = X\beta + \varepsilon. \tag{1.1}$$

The key to modeling is to estimate the parameter β, which can be estimated using ordinary least squares estimation, maximum likelihood estimation, and moment

2020 *Mathematics Subject Classification.* 91B74, 91B80, 91B86.

Key words and phrases. Prediction, weight, linear regression, bootstrap method.

This work was supported in part by Shandong Women's University High-Level Project and Talent Fund (2018GSPGJ03, 2018RCYJ03), Shandong Women's University International Business and Digital Economy Fund(2303) and International Collaborative Science and Innovation Finance (Jinan) Lab Fund (JNSX202378).

*Corresponding author.

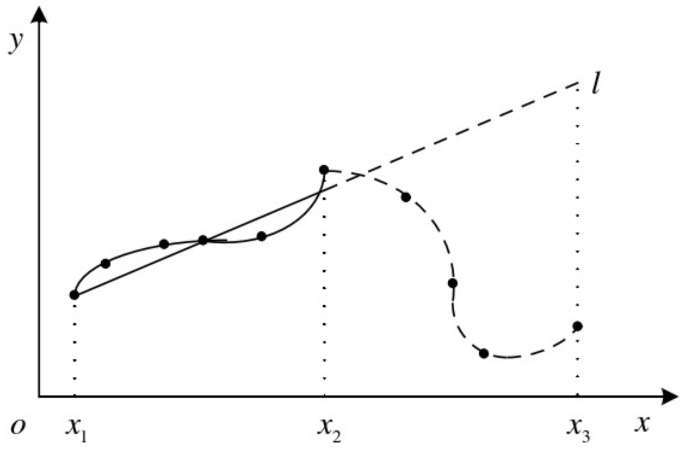

FIGURE 1. Possible failure of external prediction of linear regression model

estimation.

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y. \tag{1.2}$$

In order to improve the effectiveness of the fit, it is common to estimate the model by including all the sample points in the model, i.e., estimating it as a whole. In this way the estimation results contain all sample point information, which can maximize the internal predictive effect of the model, but this modeling idea needs to meet the assumption conditions that the model is set correctly, the explanatory variables are uncorrelated and have variability, and the residuals have conditional zero mean, homoskedasticity, and conform to normal distribution [3]. These assumption conditions in the model is that the parameters are invariant or stable throughout the sample space, which can effectively improve the predictive ability of the model. But the reality of economic and social operation is constantly changing, the model structure must also be changed, the parameters are not stable, so the model established based on the holistic modeling idea is difficult to accurately simulate the economic and social operation [9].

Existing studies have used a variety of methods to improve the predictive ability of linear regression models. The exploration of the existing literature on the improvement of linear regression models have provided a good reference and foundation to successfully conduct this study. In this paper, from the partial modeling idea, using bootstrap method to expand the sample, the predictive effect is used to weight. We constructed the predictive weighted linear regression model. The empirical results shows that the predictive ability of this model improved to some extent. The rest of the study is organized as follows. Section 2 discusses the holistic and partial modeling ideas. Section 3 describes the steps of constructing a predictive weighted linear regression model. Section 4 discusses the validity of the model constructed in this study using China macro data. Section 5 provides the discussion and conclusion of the study.

2. Partial modeling idea and predictive weighted linear regression model

The linear regression model fit can be measured through various ways, such as coefficient of determination and hypothesis testing, both of which are applications of the holistic modeling idea. The counterpart to the holistic modeling idea is the partial modeling idea, which uses external prediction as the main goal of modeling [4].

2.1. **Coefficient of Determination, Hypothesis Testing and Holistic Modeling Idea.** The total sum of squares (TSS) of the sample and the sample mean are decomposed into explained sum of squares (ESS), which can be explained by the sample regression line, and residual sum of squares (RSS), which cannot be explained by the sample regression line. ESS as a proportion of TSS is the coefficient of determination. The larger the coefficient of determination, the more effectively the model fits.

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{2.1}$$

Hypothesis testing of the linear regression model is to test whether the linear relationship between the independent variable and the dependent variable is significant, i.e. whether the estimated parameters are equal to zero. If it is not zero, it means that the linear relationship is significant and the established model has better simulation effect. The original hypothesis and alternative hypotheses for the test T and the test F

$$H_0 : \beta_h = 0; H_1 : \beta_h \neq 0, \tag{2.2}$$

$$H_0 : \beta_1 = \cdots \beta_k = 0; H_1 : \beta_1 \cdots \beta_k \,\text{is not all zero}. \tag{2.3}$$

Constructing T statistic and F statistic

$$T_h = \frac{\hat{\beta}_h - \beta_h}{SE\left(\hat{\beta}_h\right)} \sim t\left(n - k - 1\right), \tag{2.4}$$

$$F = \frac{ESS/k}{RSS/\left(n - k - 1\right)} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2/k}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2/\left(n - k - 1\right)} \sim F\left(k, n - k - 1\right). \tag{2.5}$$

According to the P-value of the T statistic and the F statistic, the judgment is made under the set significance level. From the coefficient of determination and hypothesis testing construction methods, it can be found that the measurements are based on the entire sample space and are not extrapolated beyond the sample space, which means that these two evaluation methods are mainly applicable to internal prediction evaluation.

2.2. **Prediction Guidelines and Partial Modeling Idea.** For a specific linear regression model, the predictive value is $\hat{y}_i$, the more information the predictive value $\hat{y}_i$ contains, the better the model simulation, the less information the predictive value $\hat{y}_i$ contains, the worse the model simulation. The common evaluation indexes of predictive effect are mean error, mean absolute error, mean square error and so on. These indicators can be used for internal prediction evaluation as well as external prediction evaluation, but these evaluation indicators are all absolute evaluation indicators and lack relative evaluation standards. In order to compare the predictive effect of the model with the predictive effect of the mean value, this paper selects the normalized mean square error ($NMSE$) to evaluate the predictive effect of the model.

$$NMSE = \frac{\overline{(y_i - \hat{y}_i)^2}}{\overline{(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{2.6}$$

If the normalized mean square error is less than 1, it means that the predictive effect of the model is better than the mean prediction; if the normalized mean square error is equal to 1, it means that the predictive effect of the model is comparable to the use of the mean prediction; if the normalized mean square error is greater than 1, it means that the predictive effect of the model is not as good as the mean prediction. For the latter two cases, the model has lost its application value, and it is simpler and faster to use mean value prediction directly.

With the idea of partial modeling, the entire sample space is divided into two sample spaces, the training set and the testing set. The training set is used to build the model, then the testing set is used to predict, and the normalized mean square error is applied to evaluate the model. That is, the model is built based on the training set (internal sample space), and the evaluation is based on the testing set (external sample space). If the normalized mean square error is small, it means that the model is good for external prediction. The partial modeling idea can also be used for internal prediction evaluation, just the part is extended to the whole, the training set and the test set are the same, the same sample space is used for modeling and prediction, and the normalized mean square error is still used to evaluate the model predictive effect.

3. Predictive weighted linear regression model construction

Partial modeling idea requires attention to key issues such as how the training and testing sets are determined and how parameter estimation is performed. This study used the bootstrap method to expand the sample space, and the estimated parameters were determined by weighting them according to the predictive effect [8].

3.1. **Parameter Estimation Steps Based on Predictive Weighting.** Bootstrap methods are sampling methods developed using computers to eliminate the subjectivity of sampling. In reality, economic data are relatively small and difficult to analyze accurately and comprehensively. The bootstrap method expands the sample size and improves the comprehensiveness and accuracy of the analysis. This study builds a model using a sample expanded with the help of the bootstrap method. More sub-models can be built on expanding the sample size, and these sub-models are weighted to form a more accurate comprehensive model. The weights of parameter estimators of the sub-models are determined according to the predictive effect, and the predictive effect of the sub-models is determined using the $NMSE$ [10]. The better the predictive effect of the sub-models, the larger the parameter estimator weights, which are weighted to form the parameter estimators of the final model [7]. The specific steps used are as follows:

Step 1: Expand the samples with the bootstrap method.

Step 2: Determine the number of random sample m.

Step 3: Randomize the total sample space X into training set (internal sample space) $X_j, j = 1, ..., m$ and testing set (external sample space) $X_h, h = 1, ..., m$, $X = X_j + X_h$.

Step 4: Use the training set X_j to build the model, i.e. $Y_j = X_j\hat{\beta}_j$; apply the model to predict the data in the testing set X_h, i.e. $\hat{Y}_h = X_h\hat{\beta}_j$.

Step 5: Measure the normalized mean square error $NMSE_j$ of the testing set and the weights $w_j = 1 - NMSE_j$ of the parameter estimates $\hat{\beta}_j$. If the normalized mean square error $NMSE_j$ of a model is greater than 1, it means that the predictive ability of the model is not as good as using the mean directly, and the predictive ability is weak, so it will be discarded and not included in the final modeling .

Step 6: Repeat m times steps 3-5 to obtain the parameter estimates $\hat{\beta}_1 \ldots \hat{\beta}_m$ and their corresponding weights $w_1 \ldots w_m$.

Step 7: Measure the weighted average of the parameter estimators , which is the final parameter estimate of the model, i.e. $\hat{\beta} = \sum_{j=1}^{m} \frac{w_j\hat{\beta}_j}{\sum_{j=1}^{m} w_j}$.

3.2. **Properties of Parameter Estimators of Predictive Weighted Linear Regression Model.** The parameter estimators of the predictive weighted linear regression model have excellent properties such as convergence, consistency, unbiasedness, and robustness.

3.2.1. *The parameter estimators $\hat{\beta}$ are convergent.* This is because

$$\left|\hat{\beta}_j\right| = \left|\left(X_j^T X_j\right)^{-1} X_j^T Y_j\right| < +\infty. \tag{3.1}$$

Then

$$\left|\hat{\beta}\right| = \left|\sum_{j=1}^{m} \frac{w_j\hat{\beta}_j}{\sum_{j=1}^{m} w_j}\right| < +\infty. \tag{3.2}$$

3.2.2. *As the sample size n increases, the parameter estimators $\hat{\beta}$ become closer and closer to the overall parameters β, i.e., the limit is β.* This is because

$$p\lim\nolimits_{n\to\infty}\left(\beta_j - \hat{\beta}_j\right) = \left[E\left(x_i x_i^T\right)\right]^{-1} \bullet \left[E\left(x_i\varepsilon_i\right)\right] = \left[E\left(x_i x_i^T\right)\right]^{-1} \bullet 0 = 0. \tag{3.3}$$

That is

$$p\lim_{n\to\infty}\left(\hat{\beta}_j\right) = \beta_j. \tag{3.4}$$

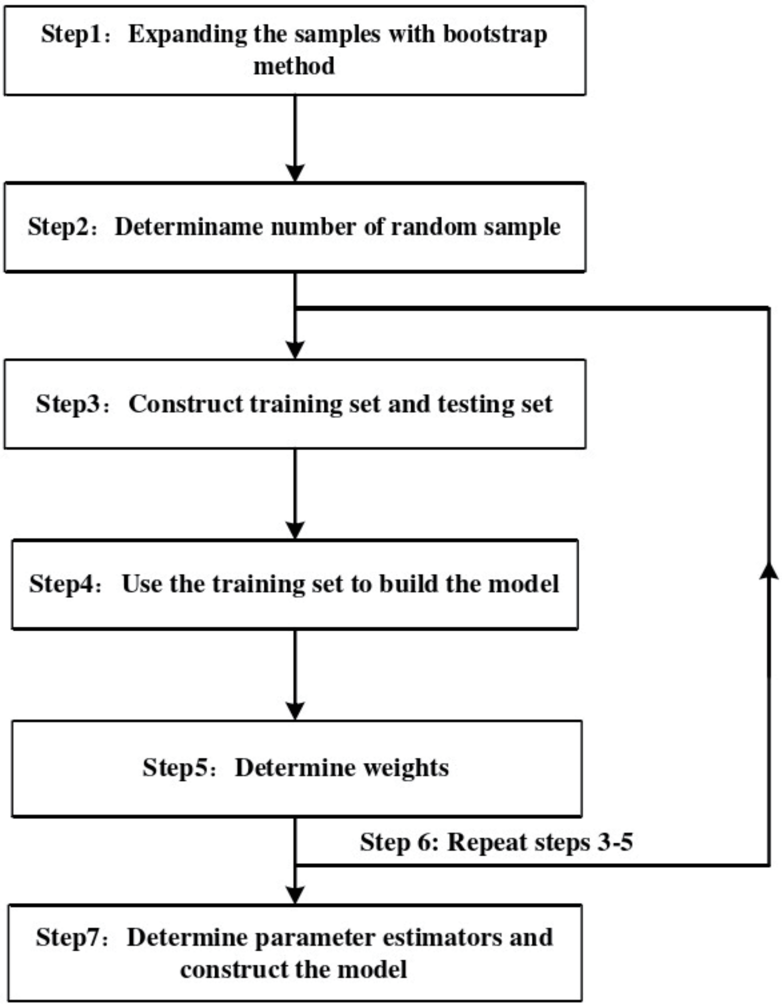
Step1：Expanding the samples with bootstrap method
Step2：Determiname number of random sample
Step3：Construct training set and testing set
Step4：Use the training set to build the model
Step5：Determine weights
Step 6: Repeat steps 3-5
Step7：Determine parameter estimators and construct the model

FIGURE 2. Simulation flow chart of predictive weighted linear regression model

Thus

$$\begin{aligned} p\lim_{n\to\infty}\left(\hat{\beta}\right) &= p\lim_{n\to\infty}\left(\sum_{j=1}^{m}\frac{w_j\hat{\beta}_j}{\sum\limits_{j=1}^{m}w_j}\right) = \sum_{j=1}^{m}\frac{w_j\, p\lim\limits_{n\to\infty}\left(\hat{\beta}_j\right)}{\sum_{j=1}^{m}w_j} \\ &= \sum_{j=1}^{m}\frac{w_j\beta_j}{\sum\limits_{j=1}^{m}w_j} = \beta. \end{aligned} \tag{3.5}$$

3.2.3. *The mathematical expectation of parameter estimators $\hat{\beta}$ are equal to the overall parameter β.* This is because

$$E\left(\hat{\beta}_j\right) = E\left[\left(X_j^TX_j\right)^{-1}X_j^TY_j\right] = \beta_j + \left(X_j^TX_j\right)^{-1}X_j^TE\left(\varepsilon_j\right) = \beta_j. \tag{3.6}$$

Then

$$E\left(\hat{\beta}\right)=E\left(\sum_{j=1}^{m}\frac{w_j\hat{\beta}_j}{\sum_{j=1}^{m}w_j}\right)=\sum_{j=1}^{m}\frac{w_jE\left(\hat{\beta}_j\right)}{\sum_{j=1}^{m}w_j}=\sum_{j=1}^{m}\frac{w_j\beta_j}{\sum_{j=1}^{m}w_j}=\beta. \tag{3.7}$$

3.2.4. *Parameter Estimators are More Robust and Tolerable.* This is mainly because the parameter estimators of the predictive weighted linear regression model are robust in two ways. First, the parameter estimators of the sub-models are estimated from random samples obtained by multiple random sampling, which is more objective [5]. Second, the parameter estimators are weighted, and the size of the weights is based on the accuracy of the prediction, and the more accurate the prediction, the larger the weight of the parameter estimators of the model. It is obvious that the prediction of the model constructed by the strong influence points and the outliers is poorer, and the corresponding parameter estimators have smaller weights [1]. At the same time, the predictive weighted linear regression model can also help to diagnose outliers and strong influence points, and these sample points can be deleted in some cases to enhance the simulation effect of the model.

3.3. **Advantages of Predictive Weighted Linear Regression Model.** Overall, the predictive weighted linear regression model established based on the partial modeling idea has relatively excellent characteristics compared with the linear regression model established using the holistic modeling idea. On the one hand, the extrapolation effect of the model is enhanced. Using the prediction principle to evaluate the model, the $NMSE$ is used for constructing weights to vote on the model; the more accurate the prediction, the higher the weight of the model, and the final model established is the model with the best predictive effect. On the other hand, the model is more robust. The traditional model is generally modeled only once, which is easy to be affected. However, the way randomly select samples to establish the weighted model is extremely good to simulate the stochastic factors, and the model is more scientific and stable.

In addition, the modeling idea has a better application in the field of nonlinear modeling. Linear models have a variety of theoretical support, while the development in the field of nonlinear modeling is relatively slow, and lack of corresponding theoretical exploration. Predictive weighted modeling can perfect this field, for example, for the construction of exponential models, logarithmic models, polynomial models and so on [2].

4. Economic application of predictive weighted linear regression model

Linear regression modeling is a widely used tool in economics research, allowing economists to examine the relationship between various economic variables. The interaction between economic growth and inflation is considered a key topic in macroeconomics. Evidence states that the economic growth rate is negatively correlated with the unemployment rate, and for every 3 percentage points above the normal growth rate of Gross Domestic Product (GDP), the unemployment rate decreased by 1 percentage point [11]. Phillips [12] found a significant negative correlation between the unemployment rate and wage inflation rate from the UK data. Subsequent studies have found that there was also a significant negative correlation between the unemployment rate and the inflation rate. These studies can be combined to introduce a positive correlation between economic growth and the inflation rate, and the model more in line with the actual economic data is the inflation model that considers the expectations.

$$\pi_t = \pi_t * + \lambda \left(G_t - \bar{G}\right) + \varepsilon_t, \ \lambda > 0. \tag{4.1}$$

Where π is the inflation rate, $\pi_t *$ is the expected inflation rate, G_t , $\bar{G}$ are the economic growth rate and the potential economic growth rate respectively, and the expected inflation rate is set to the inflation rate of the previous period, then the model is

$$\pi_t = \beta_0 + \beta_1 \pi_{t-1} + \beta_2 \left(G_t - \bar{G}\right) + \varepsilon_t, \ \beta_2 > 0. \tag{4.2}$$

Data on China economic growth rate and inflation rate from 1982 to 2020 is collected from the China Statistical Yearbook, and the descriptive statistical analysis of the data is shown in Table 1.

Table 1. Descriptive statistical analysis

Varibales	Size	Mean	Standard deviation	Maximum	Minimum
Economic growth rate	39	9.411	2.906	15.19	2.3
Inflation rate	39	4.908	5.921	24.1	-1.4

Given the relatively small number of samples, the bootstrap method is used to expand the samples to 100, 1000, and 10000. Using the 10-fold cross-test, the fitting results are shown in Table 2.

Table 2. $NMSE$ for each model with different sample size

Model	100 Samples	1000 Samples	10000 Samples
Model 1	0.439	0.381	0.399
Model 2	0.415	0.389	0.407
Model 3	1.613	0.403	0.387
Model 4	0.564	0.443	0.429
Model 5	0.330	0.450	0.419
Model 6	0.541	0.421	0.410
Model 7	0.289	0.458	0.397
Model 8	3.25	0.478	0.424
Model 9	0.502	0.436	0.417
Model 10	0.666	0.403	0.422
Average	0.861	0.426	0.411
PWLRM	0.423	0.414	0.410

It can be found that the fitting results become more and more stable as the sample size increases. with 100 samples, the $NMSE$ of the different models varies greatly, with a minimum value of 0.289, a maximum value of 3.25, and an average of 0.861. the $NMSE$ of the predictive weighted linear regression model is 0.423, which is a derease of 50.8%. With 1,000 samples, the $NMSE$ of the different models has a minimum value of 0.381, a maximum value of 0.478, and an average of 0.426. The $NMSE$ of the predictive weighted linear regression model is 0.414, which is a decrease of 2.9%. With 10,000 samples, the $NMSE$ of the different models has a minimum value of 0.387, a maximum value of 0.429, and an average of 0.411. The $NMSE$ of the predictive weighted linear regression model is 0.410, which is a decrease of 0.2%. It can be found that the predictive weighted linear regression model has better predictive accuracy.

In summary, if the samples are expanded to 100, the model after weighting based on the predictive effects is

$$\pi_t = 1.057 + 0.518\pi_{t-1} + 0.743\left(G_t - \bar{G}\right). \tag{4.3}$$

If the samples are expanded to 1000, the model after weighting based on the predictive effects is

$$\pi_t = 1.435 + 0.734\pi_{t-1} + 0.691\left(G_t - \bar{G}\right). \tag{4.4}$$

If the samples are expanded to 10000, the model after weighting based on the predictive effects is

$$\pi_t = 1.430 + 0.693\pi_{t-1} + 0.708\left(G_t - \bar{G}\right). \tag{4.5}$$

5. Discussion and conclusion

The primary purpose of all models is prediction, and models can be used for both internal and external prediction. The linear regression model is the most widely used statistical model. However, it focuses more on internal prediction. If parameter instability arises due to inconsistencies in model assumptions, the external predictive effect will be greatly affected. According to the idea of partial modeling, with the help of the bootstrap method, parameter estimation is carried out using the weighting of the prediction criterion, and the higher the prediction accuracy, the higher the weight of the sub-model. The parameter estimators formed by predictive weighting have excellent properties such as convergence, consistency, unbiasedness and robustness. An empirical discussion is conducted using China macroeconomic data, and the samples are expanded to 100, 1,000, and 10,000 using the bootstrap method, and the results show that the $NMSE$ of the predictive weighted linear regression model is smaller, which decrease by 50.8%, 2.9%, and 0.2%, respectively. However, this study has some limitations. The model constructed in this study is more complex compared with the general linear regression model. For example, if the amount of data is relatively large or there are many data types, the calculation will take more time. Moreover, the validity of the modeling in this study needs more experimental verification. The model may be more effective for some datasets, and may not be particularly effective for certain other datasets, which is a direction for future exploration. Overall, this study is an attempt to improve the external predictive power of the linear regression model to some extent. Moreover, the modeling ideas proposed in this study can also be applied to nonlinear models and widely extended.

References

[1] M. S. Barrera, S. V. Aelst and V. J. Yohai, *Robust tests for linear regression models based on τ-estimates*, Computational Statistics and Data Analysis **93** (2016), 436–455.

[2] F .Belarbi, S. Chemikh and A. Laksaci, *Local linear estimate of the nonparametric robust regression in functional data*, Statistics and Probability Letters **134** (2018), 128–133.

[3] T. Boot, *Joint inference based on Stein-type averaging estimators in the linear regression model*, Journal of Econometrics **235** (2023), 1542–1563.

[4] T. Bos and J. S. Hieber, *Convergence guarantees for forward gradient descent in the linear regression model*, Journal of Statistical Planning and Inference **233** (2024): 106174.

[5] Q. R. Cui, Y. Q. Xu, Z. J. Zhang and V. Chan, *Max-linear regression models with regularization*, Journal of Econometrics **222** (2021), 579–600.

[6] J. T. Chi and I. C. F. Ipsen, *Multiplicative perturbation bounds for multivariate multiple linear regression in schatten p-norms*, Linear Algebra and its Applications **624** (2021), 87–102.

[7] R. Davidson and and M. Trokic, *The fast iterated bootstrap*, Journal of Econometrics **218** (2020), 451–475.

[8] B. Efron, *Bootstrap methods: Another look at the jackknife*, Annals of Statistics **7** (1979), 1–26.

[9] M. Friedrich, S. Smeekes and J. P. Urbain, *Autoregressive wild bootstrap inference for nonparametric trends*, Journal of Econometrics **214** (2020), 81–109.

[10] B. Funkea and M. Hirukawa, *Bias correction for local linear regression estimation using asymmetric kernels via the skewing method*, Econometrics and Statistics **20** (2021), 109–130.

[11] R. J. Gordon, *Okun's law and productivity innovations*, The American Economic Review **100** (2010), 11–15.

[12] A. W. Phillips, *The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957*, Economica **25** (1958), 283–299.

[13] Rasyidah, R. Efendi, N. M. Nawi, M. M. Deris and S. M. A. Burney, *Cleansing of inconsistent sample in linear regression model based on rough sets theory*, Systems and Soft Computing **5** (2023): 200046.

Manuscript received July 18, 2024
revised December 13, 2024

J. F. Cui
School of Economics, Shandong Women's University, China
E-mail address: 30087@sdwu.edu.cn

Y. C. He
School of Economics, Shandong Women's University, China
E-mail address: 30099@sdwu.edu.cn

S. Guo
School of Economics, Shandong Women's University, China
E-mail address: 30090@sdwu.edu.cn

Y. Liu
School of Economics, Shandong Women's University, China
E-mail address: 202201016@sdwu.edu.cn

M. Hao
Department of Business Administration, Shandong Labor Vocational and Technical College, China
E-mail address: hmsdlvtc@163.com