UDC 519.237.5 Вестник СПбГУ. Прикладная математика. Информатика... 2022. Т. 18. Вып. 4
MSC 91A25

New application of multiple linear regression method-A case in China air quality

Y. He[1], *D. Qi*[1], *V. M. Bure*[1,2]

[1] St Petersburg State University, 7–9, Universitetskaya nab., St Petersburg, 199034, Russian Federation
[2] Agrophysical Research Institute, 14, Grazhdanskiy pr., St Petersburg, 195220, Russian Federation

For citation: He Y., Qi D., Bure V. M. New application of multiple linear regression method-A case in China air quality. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2022, vol. 18, iss. 4, pp. 516–526.
https://doi.org/10.21638/11701/spbu10.2022.406

In this paper, we propose an econometric model based on the multiple linear regression method. This research aims to evaluate the most important factors of the dependent variable. To be more specific, we consider the properties of this model, model quality, parameters test, checking the residual of the model. Then, to ensure that the prediction model is optimal, we use the backward elimination stepwise regression method to get the final model. At the same time, we also need to check the properties in each step. Finally, the results are illustrated by a real case in China air quality. The achieved model was applied to predict the 31 capital cities in China's air quality index (AQI) during 2013–2019 per year. All calculations and tests were achieved by using *R*-studio. The dependent variable is the China's AQI. The control variables are six pollutant factors and four meteorological factors. In summary, the model shows that the most significant influencing factor of the AQI in China is $PM_{2.5}$, followed by O_3.

Keywords: multiple linear regression, air pollution, AQI, hypothesis test, $PM_{2.5}$, O_3.

1. Introduction. Applied statistical methods are widely used in natural science and social science, such as ecology, astronomy, physics, medicine and other fields [1–4]. One most popular methods are linear regression, which are simple linear regression, multiple linear regressions, logistic regression, ordinal regression, multinominal regression and discriminant analysis [5]. This method's application typically breaks up into two groups, prediction and category problems [6]. However, there are applications in which the controlled variable is the most crucial variable between the dependent variable and controlled variable during past times. It can also be seen as a prediction problem based on multiple linear regression [7].

To the best of the author's knowledge of multiple linear regression, regression analysis was first developed in the latter part of the XIX century by Sir Francis Galton [8]. The honour of the first publication of the least square (LS) method belongs to Legendre in Paris in 1805 [9]. Later, the first proof of the LS method was given by Dr. Robert in 1808 [10]. Simultaneously, Gauss further developed the theory of least squares in 1821, including a version of the Gauss — Markov theorem [11]. Later, the term t-statistic is abbreviated as hypothesis test statistic, it was used to test whether the controlled variable is significant or not — the theorem from William Sealy Gosset. He first published it in English in 1908 [12]. Also, f-statistic usually tests the linear regression model is significant or not [13].

In all the above described references, whether the model and parameters are significant or not. However, some controlled variables are not significant. Then, we have several choices to ensure that all variables are significant: Forward selection, backward elimination and stepwise regression. After that, we could get the final prediction model. Consider, at each step, the residuals should be normality and have non-autocorrelation between each other. Residual normality can be seen by a histogram, plot $Q-Q$, kurtosis, and skewness. Some hypothesis tests can also be tested, such as the Kolmogorov—Smirnov test, Anderson—Darling test, etc. In contrast to the autocorrelation between residuals, we use the Durbin—Watson statistic test. Also, to avoid heteroscedasticity, we use $\mathrm{Ln}(y)$ to replace Y. Another part for multicollinearity, we use stepwise to get the final model.

Therefore, it makes sense to determine the final predicting model, which can also be used to determine the essential factors among dependent variables. In this paper, we discuss the details of each step. In particular, model quality checked by R-square [14], model is significant or not checked by Fisher's test. Controlled variables are significant or not checked by Student's test. Consider, 31 capital cities to represent China, the residual normality test by Shapiro—Wilk test [15], Autocorrelation test by Durbin—Watson statistic test [16]. On the other hand, this application also can be helpful when determining the most important factors among controlled variables.

This contribution is organized as follows. In Section 2 the basic model and properties are introduced. In Section 3 China air quality index (AQI) and its computation. In Sections 4 and 5 appiled model at different years and result can be show at two different ways. In Section 6 some concluding discussions are provided.

2. Multiple regression models. This study specified a model for AQI prediction that predicts the impact of AQI on local air quality in target cities through pollutants and meteorological factors to quantify the analysis of air pollution levels and the main influencing factors during 2013–2019. Here i represents the cities and t is the time:

$$\mathrm{Ln}(y_{it}) = \beta_0 + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \beta_3 x_{3,it} + \beta_4 x_{4,it} + \beta_5 x_{5,it} + \\ + \beta_6 x_{6,it} + \beta_7 x_{7,it} + \beta_8 x_{8,it} + \beta_9 x_{9,it} + \beta_{10} x_{10,it} + \delta_{it},$$

where $\mathrm{Ln}(y_{it}) = \mathrm{Ln}(\mathrm{AQI}_{it})$ the dependent variable is the logarithm of AQI of each city per year, with AQI data being logarithmically transformed to avoid the potential heteroscedacity; $x_{1,it} = \mathrm{SO}_{2,it}$ is the average per year concentration of SO_2 ($\mu\mathrm{g/m}^3$) in city i at time t; $x_{2,it} = \mathrm{NO}_{2,it}$ is the average per year concentration of NO_2 ($\mu\mathrm{g/m}^3$) in city i at time t; $x_{3,it} = \mathrm{PM}_{10,it}$ is average per year concentration of $\mathrm{PM}_{10,it}$ in city i at time t; $x_{4,it} = \mathrm{CO}_{it}$ is the 95th percentile daily average concentration of CO ($\mu\mathrm{g/m}^3$) in city i at time t; $x_{5,it} = \mathrm{O}_{3,it}$ is 95th percentile daily maximum 8 hours average concentration of Ozone ($\mu\mathrm{g/m}^3$) in city i at time t; $x_{6,it} = \mathrm{PM}_{2.5,it}$ is the average year concentration of $\mathrm{PM}_{2.5}$ ($\mu\mathrm{g/m}^3$) in city i at time t; $x_{7,it} = T_{it}$ is the average temperature per year ($^\circ$C) in city i at time t; $x_{8,it} = \mathrm{HU}_{it}$ is the average relative humidity per year (%) in city i at time t; $x_{9,it} = \mathrm{PR}_{it}$ indicate precipitation (milimeters) in city i at time t; $x_{10,it} = \mathrm{SH}_{it}$ is the sunshine per year (hours) in city i at time t; $\beta_0, ..., \beta_{10}$ are the 11 parameters need to be determined; δ_{it} is an unobservable random variable in city i at time t.

Furthermore, the residuals must satisfy the following four Gauss—Markov conditions and additional condition.

Assumption 1. The expected value of the error term is zero for all observations, i. e. $E(\delta_{it} = 0), \ \ i = 1, 2, ..., 31.$

Assumption 2. Homoskedasticity. The variance of the error term in constant for all $i = 1, 2, ..., 31$: $\mathrm{var}(\delta_{it}) = \sigma^2$.

Assumption 3. Error term is independently distributed and not correlated: $\operatorname{cov}(\delta_{it}, \delta_{jt}) = 0,\ i \neq j$.

Assumption 4. x_{it} is independently deterministic: independent variable x_{it} is uncorrelated with the error term, where x_{it} represent above controlled variable.

Additional condition:

Assumption 5. Observation errors are normally distributed random variables.

2.1. Model quality test. Due to the assumption above, the first step is to test whether the model is highly significant. Typically, we use R^2 to test whether the model has high quality or not if R^2 is more than 0.8, which means the model has a high quality. Else if R^2 is less than 0.8, the model is not suitable for an estimate:

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}, \tag{1}$$

among them

$$\text{SST} = \sum_{i=1}^{n} (\text{Ln}(\text{AQI})_{it} - \text{Ln}(\bar{\text{A}}\text{QI})_{it})^2, \tag{2}$$

$$\text{SSE} = \sum_{i=1}^{n} (\text{Ln}(\hat{\text{A}}\text{QI})_{it} - \text{Ln}(\bar{\text{A}}\text{QI})_{it})^2, \tag{3}$$

$$\text{SSR} = \sum_{i=1}^{n} (\text{Ln}(\text{AQI})_{it} - \text{Ln}(\hat{\text{A}}\text{QI})_{it})^2. \tag{4}$$

In formulas (1)–(4) SST is the total sum of squares, SSE is the explained sum of squares, SSR is the residual sum of the squares or sum of squared residuals, $\text{Ln}(\bar{\text{A}}\text{QI})_{it}$ and $\text{Ln}(\hat{\text{A}}\text{QI})_{it}$ corresponding to the average value of $\text{Ln}(\text{AQI})_{it}$ and the fitted values. Moreover, it is interpreted as the proportion of the sample variation in $\text{Ln}(\text{AQI})_{it}$ that is explained by the ols regression line. By definition, R^2 is a number between zero and one. If R^2 is equal to 1, it means the model has a high significance.

2.2. F-test. After that, we use Fisher's test or f-test to test model is significant or not, which can be used to compare the values of the sample variances of two independent samples. When executing the null hypothesis, test statistics of the criterion have a Fisher distribution (f-distribution).

Null hypothesis: $\beta_{1,t} = \beta_{2,t} = ... = \beta_{10,t} = 0$.

Alternative hypothesis: at least one of the coefficient is not equal to zero, where n is the observation number; k is the number of the parameters. In this test, we define the significance level α as equal to 0.05. If the f-test value is more than the f_{tab} value, then the model is significant. Else, corresponding to the p-value (f-test) being less than 0.05, we can also conclude that the model is significant:

$$F = \frac{R^2}{1 - R^2} \frac{n - k - 1}{k}.$$

2.3. Student t-test. Null hypothesis: $\beta_{1,t} = 0$, $\beta_{2,t} = 0$,..., $\beta_{10,t} = 0$ (coefficient is not significant).

Alternative hypothesis: $\beta_{1,t} = 0$, $\beta_{2,t} = 0$,..., $\beta_{10,t} = 0$.

The significance level is also defined as $\alpha = 0.05$. If the Student's t-test value is more than the t_{tab} value, then the coefficient is significant. Else, corresponding to the p-value (t-test) being less than 0.05, we conclude that the coefficient is significant.

2.4. Residual normality test. Null hypothesis: all of residuals with normality.

Alternative hypothesis: the residuals are not normal distribution.

The normality can be checked by using several tests by Kolmogorov — Smirnov (if the sample size is more than 50), Shapiro — Wilk (if the sample size is less than 50) and the Anderson — Darling test. Since the sample size in this research is $n = 31$ less than 50, we choose the Shapiro — Wilk test. The significant level we define $\alpha = 0.05$. When the p-value is more than 0.05, we reject the alternative hypothesis, and the residual is the normal distribution.

2.5. Residual autocorrelation test. Autocorrelation means that the residuals satisfy the equation

$$\delta_{i,t} = \rho\delta_{i-1,t} + v_{i,t},$$

where $\varepsilon_{i,t}$ is the residual; I is the observation number; ρ is autocorrelation of residual; $v_{i,t}$ is the intercept.

Null hypothesis: $\rho = 0$ (autocorrelation is zero).

Alternative hypothesis: $\rho \neq 0$ (autocorrelation is not zero).

From here

$$\mathrm{D-W} = \frac{\sum(\delta_{i,t} - \delta_{i-1,t})^2}{\sum \delta_{i,t}^2}.$$

In our model, we use the D — W (Durbin — Watson) statistic test to check residuals autocorrelation is zero or not. If the p-value is more than 0.05, we could accept the null hypothesis, the autocorrelation of the residuals is zero.

2.6. Data. Six pollutant such as O_3, CO, SO_2, $PM_{2.5}$, NO_2 and PM_{10}, temperature, humidity, light and precipitation for 31 provincial capitals cities from the China Statistical Yearbook. AQI per year data calculated by AQI-calculator at the platform website by the authors.

3. China AQI. This AQI was first proposed by the Ministry of Environment of the People's Republic of China and used to measure the degree of air pollution. Government agencies use an AQI to communicate how polluted the air currently is and how polluted it is forecast to become. Public health risks increase as the AQI rises. The AQI is a dimensionless index that quantitatively describes the condition of air quality. For example, as shown in Figure 1, 31 capital cities AQI situation at 2019, the different colours represent different air quality levels — the darker the colour, the more serious the air pollution. As we see, most cities' air quality level belongs to good and light pollution.

3.1. AQI calculation. The primary pollutants involved in the air quality evaluation were fine particles, inhalable particulate matter, nitrogen dioxide, ozone, sulfur dioxide and carbon monoxide. China' AQI is calculated by below:

$$\mathrm{AQI} = \max\left\{I\mathrm{AQI}_1, I\mathrm{AQI}_2, I\mathrm{AQI}_3, ..., I\mathrm{AQI}_n\right\},$$

where AQI is air quality index; $I\mathrm{AQI}_n$ is individual air quality index; N is pollution project, which is from one to six. For explicitly, 1 is sulfur dioxide, 2 is nitrogen dioxide, 3 is inhalable particulate matter PM_{10}, 4 is carbon monoxide, 5 is Ozone, 6 is fine particles $PM_{2.5}$.

3.2. IAQI calculation. In formula

$$I\mathrm{AQI}_n = \frac{I\mathrm{AQI}_{h_i} - I\mathrm{AQI}_{l_o}}{BP_{h_i} - BP_{l_o}}(C_n - BP_{l_o}) + I\mathrm{AQI}_{l_o},$$

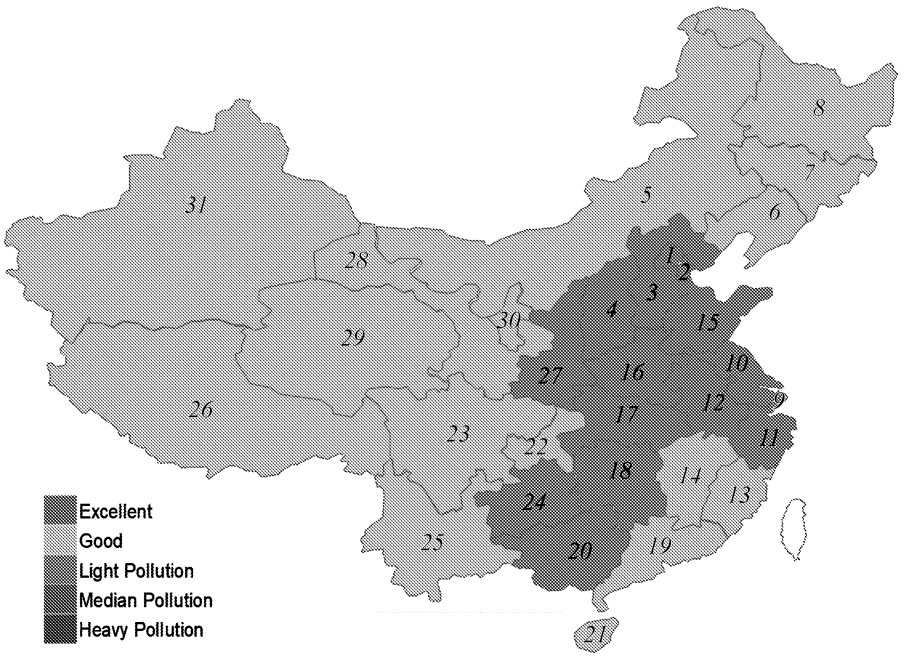

Figure 1. 31 capital cities AQI situation at 2019
1 — Beijing; *2* — Tianjin; *3* — Shijiazhuang; *4* — Taiyuan; *5* — Hohhot; *6* — Shenyang; *7* — Changchun; *8* — Harbin; *9* — Shanghai; *10* — Nanjing; *11* — Hangzhou; *12* — Hefei; *13* — Fuzhou; *14* — Nanchang; *15* — Jinan; *16* — Zhengzhou; *17* — Wuhan; *18* — Changsha; *19* — Guangzhou; *20* — Nanning; *21* — Haikou; *22* — Chongqing; *23* — Chengdu; *24* — Guiyang; *25* — Kunming; *26* — Lhasa; *27* — Xi'an; *28* — Lanzhou; *29* — Xining; *30* — Yinchuan; *31* — Urumqi.

$I\text{AQI}_n$ is the Sub AQI of pollutant project n, n represent pollutant item; C_n is the mass concentration value of the pollutant item n; BP_{h_i} is the high value of the pollutant concentration limit close to C_n in the pollutant item concentration corresponding to the air quality sub-index; BP_{l_o} is the low value of the pollutant concentration limit close to C_n in the pollutant item concentration corresponding to the air quality sub-index; $I\text{AQI}_{h_i}$ is the air quality sub-index corresponding to BP_{h_i} in the pollutant item concentration corresponding to the air quality sub-index; $I\text{AQI}_{l_o}$ is the air quality sub-index corresponding to BP_{l_o} in the pollutant item concentration corresponding to the air quality sub-index.

4. Empirical result and explanations.

4.1. Applied model at 2019. The approach was applied to the linear regression by calling the OLS method in *R*-studio, and the parameters and significance tests were obtained after the run (Table 1). A regression analysis with a constant was performed using Ln(AQI) as the dependent variable, pollutant items, meteorological factors as explanatory variables. It can be seen that degree of freedom is 20, which means the sample variables is 31. $R^2 = 0.9959$, is more than 0.8 and extremely close 1. On the one hand, there are 99.5 per cent changes in the response variable Ln(AQI) because of the combination of ten controlled variables. On the other hand, it means the model has a high quality. Adjust $R^2 = 0.993$, indicating that the regression equation determines that 99.3 % of the variance of the dependent variable within the observed value.

When we set significance level as 0.05, the $F_{\text{table}} = F_{(0.05,10.31-10-1)} = 2.348$. From the Table 1, the f-statistic value equally 487.4 and p-value less $2.2e^{-16}$ is less than 0.05,

which means we could reject the null hypothesis, we could conclude that the model is significant.

Table 1. **Economic model result at 2019**

Ln(AQI)	Coefficient	st.err	t-value	p-value	sig
x_1	$-3.519 \cdot 10^{-04}$	$1.168 \cdot 10^{-03}$	-0.301	0.7664	—
x_2	$1.581 \cdot 10^{-03}$	$7.105 \cdot 10^{-04}$	2.226	0.0377	*
x_3	$-2.700 \cdot 10^{-04}$	$6.146 \cdot 10^{-04}$	-0.439	0.6651	—
x_4	$4.550 \cdot 10^{-03}$	$1.216 \cdot 10^{-02}$	0.374	0.7123	—
x_5	$8.798 \cdot 10^{-03}$	$2.659 \cdot 10^{-04}$	33.092	$< 2e^{-16}$	***
x_6	$-1.078 \cdot 10^{-03}$	$9.288 \cdot 10^{-04}$	-1.160	0.2596	—
x_7	$5.152 \cdot 10^{-04}$	$1.670 \cdot 10^{-03}$	0.309	0.7608	—
x_8	$6.998 \cdot 10^{-05}$	$6.860 \cdot 10-^{04}$	0.102	0.9198	—
x_9	$-3.953 \cdot 10^{-06}$	$1.204 \cdot 10^{-05}$	-0.328	0.7461	—
x_{10}	$-2.982 \cdot 10^{-06}$	$1.189 \cdot 10^{-05}$	-0.251	0.8046	—
Constant	$3.182 \cdot 10^{+00}$	$6.706 \cdot 10^{-02}$	47.458	$< 2e - 16$	***
res st.err	0.01776	Degrees of freedom	20	Mul R-squared	0.9959
f-statistic	487.4	p-value	$< 2.2 \cdot 10^{-16}$	Adj R-squared	0.9939
sig.codes	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 '-'

N o t e. If $P(t) > 0.1$, then the significance level signal is —; if $0.05 < P(t) < 0.1$, then the significance level signal is .; if the $0.01 < P(t) < 0.05$, then the significance level signal is $*$; if the $0.001 < P(t) < 0.01$, then the significance level signal is $**$; if the $0 < P(t) < 0.001$, then the significance level signal is $***$.

However, the initial regression equation is significant as a whole, but this does not mean that every controlled variable is significant. According to Table 1, when absolute t-value is more than $T_{\text{tab}} = T_{(0.05,10.31-10-1)} = 2.086$ or the p-value is less than 0.05, we could reject the Student test null hypothesis, the variable is significant. In the initial model, O_3, NO_2 are significant, and the other eight controlled variables are not significant. According to the regression results in Table 1, we can get the regression model:

$$\text{Ln}(y) = 3.18 - 3.51 \cdot 10^{-4} x_1 + 1.58 \cdot 10^{-3} x_2 + 2.7 \cdot 10^{-4} x_3 - 4.55 \cdot 10^{-3} x_4 + 8.8 \cdot 10^{-03} x_5 -$$

$$- 1.08 \cdot 10^{-3} x_6 + 5.15 \cdot 10^{-4} x_7 + 7 \cdot 10^{-5} x_8 - 2.95 \cdot 10^{-6} x_9 - 2.98 \cdot 10^{-6} x_{10}.$$

Table 2. **The dynamic processes of stepwise regression at 2019**

Variable	R^2	AdjR^2	f-value	$P(F)$	S—W	P(S—W)	D—W	P(D—W)
$-x_8$	0.995	0.993	487.35	$1.3 \cdot 10^{-21}$	0.984	0.908	1.899	0.225
$-x_1$	0.995	0.994	569.28	$5.2 \cdot 10^{-23}$	0.983	0.879	1.894	0.213
$-x_4$	0.995	0.994	667.07	$2 \cdot 10^{-24}$	0.981	0.836	1.890	0.213
$-x_{10}$	0.995	0.994	792.40	$7.2 \cdot 10^{-26}$	0.982	0.883	1.885	0.243
$-x_9$	0.995	0.994	960.40	$2.3 \cdot 10^{-27}$	0.987	0.962	1.852	0.238
$-x_7$	0.995	0.994	1192.90	$6.9 \cdot 10^{-29}$	0.98	0.817	1.896	0.281
$-x_3$	0.995	0.995	1508.87	$2.3 \cdot 10^{-30}$	0.98	0.816	1.992	0.407
x_2, x_5, x_6	0.995	0.995	1897.39	$9.1 \cdot 10^{-32}$	0.978	0.751	2.094	0.556

N o t e. S—W mean Shapiro — Wilk test, D—W mean Durbin — Watson statistic test.

In order to make that all the variables are significant, we choose the backward elimination stepwise regression method. In each step, we need to delete one variable which is not significant by the Student t-test. From Table 2, the f-statistic value continues to increase in each step. We delete the maximal p-value by Student t-test and get the maximal f statistic value. In column 1, the bolded numbers are the control variables with the lowest levels of significance and the ones we eliminate at each stage. The stop condition remains

controlled variables are significant. We also test the residual autocorrelation, residual normality in each step. From the Shapiro — Wilk test, each step p-value is more than 0.05. We could reject the alternative hypothesis. The residuals are normal distribution. As for the Durbin — Watson test result, all the p-value are more than 0.05. We could reject the alternative hypothesis. The autocorrelation between residuals is zero. The final model in this year are as follows:

$$\underset{P(t\text{-test})}{\mathrm{Ln}(y)} = \underset{(2.29\cdot 10^{-42})}{3.18} + \underset{(8.14\cdot 10^{-3})}{1.47\cdot 10^{-3}x_2} + \underset{(9.06\cdot 10^{-30})}{8.85\cdot 10^{-3}x_5} - \underset{(1.5\cdot 10^{-3})}{1.46\cdot 10^{-3}x_6}\,.$$

4.2. Applied model at 2013–2018. First, we check the model quality at 2013–2018, as shown in Figure 2, a, b, for both the initial model and the stepped final model after backward elimination, with all R, R^2 values exceeding 0.8, which indicates a high quality of the model. In contrast to the initial model, the distance between R^2 and adjust R^2 in the final model gradually decreases, indicating that we have obtained a more stable model.

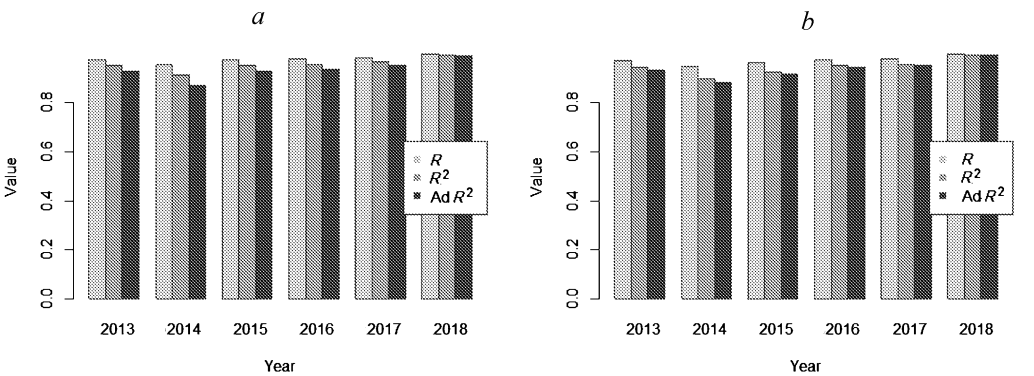

Figure 2. Initial (a) and final (b) economic model quality at 2013–2018

Then we use the same method to get the final model for the 2013–2018. Table 3 presents the final model after stepwise regression from 2013–2018. The elimination variable is step by step until all variables are significant. According to the Fisher test result, all F-statistic values are more than F_{tab} value, corresponding to all p-values are less than 0.05, which means we could reject the null hypothesis, all final models are significant. Else, from the Shapiro — Wilk test, all p-value is more than significant level 0.05, we could reject the alternative hypothesis, all residuals are from the normal distribution. From the Durbin — Watson test result, all p-value is more than significant level 0.05. Then we could reject the alternative hypothesis. The autocorrelation of the residuals is zero.

Table 3. **The economic model result after stepwise regression at 2013–2018**

Year	Elimination variable	f-value	$p(f)$	S—W	p(S—W)	D—W	p(D—W)
2013	$4, 3, 2, 10, 5, 9$	85.82	$6.45\cdot 10^{-15}$	0.975	0.666	2.24	0.653
2014	$3, 4, 7, 1, 5, 10$	57.77	$1.47\cdot 10^{-12}$	0.974	0.624	2.637	0.943
2015	$4, 3, 2, 10, 8, 9$	83.04	$2.03\cdot 10^{-14}$	0.966	0.418	2.016	0.412
2016	$2, 3, 10, 4, 8$	100.45	$1.01\cdot 10^{-15}$	0.945	0.11	1.231	0.371
2017	$9, 8, 1, 7, 4, 3, 10, 2$	313.63	$6.77\cdot 10^{-20}$	0.935	0.06	1.894	0.213
2018	$10, 9, 8, 1, 4, 7, 3$	1828.98	$2.77\cdot 10^{-31}$	0.984	0.919	1.532	0.083

Finally, the model all variables are significant and have the following form at different years:

in 2013:

$$\underset{(P(t\text{-test}))}{\mathrm{Ln}(y)} = \underset{(5.9\cdot10^{-25})}{4.48} \underset{(0.048)}{- 1.5\cdot10^{-3}x_1} + \underset{(7\cdot10^{-15})}{1.079\cdot10^{-2}x_6} \underset{(0.0198)}{- 9.3\cdot10^{-3}x_7} - \\ \underset{(6.9\cdot10^{-5})}{- 9.6\cdot10^{-3}x_8} + \underset{(6.8\cdot10^{-5})}{1.8\cdot10^{-4}x_9}, \tag{5}$$

in 2014:

$$\underset{(P(t\text{-test}))}{\mathrm{Ln}(y)} = \underset{(4.9\cdot10^{-24})}{4.17} + \underset{(2.1\cdot10^{-4})}{8.35\cdot10^{-3}x_2} + \underset{(8\cdot10^{-9})}{8.56\cdot10^{-3}x_6} \underset{(2.4\cdot10^{-5})}{- 1.14\cdot10^{-2}x_8} + \underset{(1.9\cdot10^{-4})}{2.34\cdot10^{-4}x_9}, \tag{6}$$

in 2015:

$$\underset{(P(t\text{-test}))}{\mathrm{Ln}(y)} = \underset{(9\cdot10^{-25})}{3.58} \underset{(0.047)}{- 1.90\cdot10^{-3}x_1} + \underset{(7\cdot10^{-9})}{4.99\cdot10^{-3}x_5} + \underset{(2.8\cdot10^{-8})}{6.69\cdot10^{-3}x_6} \underset{(7.53\cdot10^{-3})}{- 8.15\cdot10^{-3}x_7}, \tag{7}$$

in 2016:

$$\underset{(P(t\text{-test}))}{\mathrm{Ln}(y)} = \underset{(5.2\cdot10^{-27})}{3.46} \underset{(0.033)}{- 1.75\cdot10^{-3}x_1} + \underset{(1.2\cdot10^{-12})}{6.23\cdot10^{-3}x_5} + \underset{(1.5\cdot10^{-7})}{5.62\cdot10^{-3}x_6} - \\ \underset{(1.2\cdot10^{-4})}{- 1.42\cdot10^{-2}x_7} + \underset{(8.98\cdot10^{-5})}{7.47\cdot10^{-5}x_9}, \tag{8}$$

in 2017:

$$\underset{P(t\text{-test})}{\mathrm{Ln}(y)} = \underset{(4.2\cdot10^{-31})}{3.31} + \underset{(3.2\cdot10^{-16})}{7.41\cdot10^{-3}x_5} + \underset{(3.59\cdot10^{-3})}{2.25\cdot10^{-3}x_6}, \tag{9}$$

in 2018:

$$\underset{P(t\text{-test})}{\mathrm{Ln}(y)} = \underset{(1.9\cdot10^{-41})}{3.18} + \underset{(9.5\cdot10^{-4})}{2.12\cdot10^{-3}x_2} + \underset{(4\cdot10^{-27})}{8.67\cdot10^{-3}x_5} \underset{(4.7\cdot10^{-3})}{- 1.4\cdot10^{-3}x_6}\,. \tag{10}$$

5. Results. From the initial values and the estimated values of Y in the resulting final equation, the fit of the economic model can be plotted for each year (see (5)–(10)), as shown in Figure 3, I—VII. Combined with the previous model quality R^2, the final predicting model fits very well.

Table 4. **Final model results at 2013–2019**

Year	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
2013	★	—	—	—	-	★	★	★	★	—
2014	—	★	—	—	—	★	—	★	★	—
2015	★	—	—	—	★	★	★	—	—	—
2016	★	—	—	—	★	★	★	—	★	—
2017	—	—	—	—	★	★	—	—	—	—
2018	—	★	—	—	★	★	—	—	—	—
2019	—	★	—	—	★	★	—	—	—	—
All	3	3	0	0	5	7	3	2	3	0

N o t e. — mean this variable is deleted, ★ mean this variable stall remain in the final model at each year.

On the other hand, the final model results can also evaluate important factors between dependent and controlled variables. From Table 4, we set each row as a final model result.

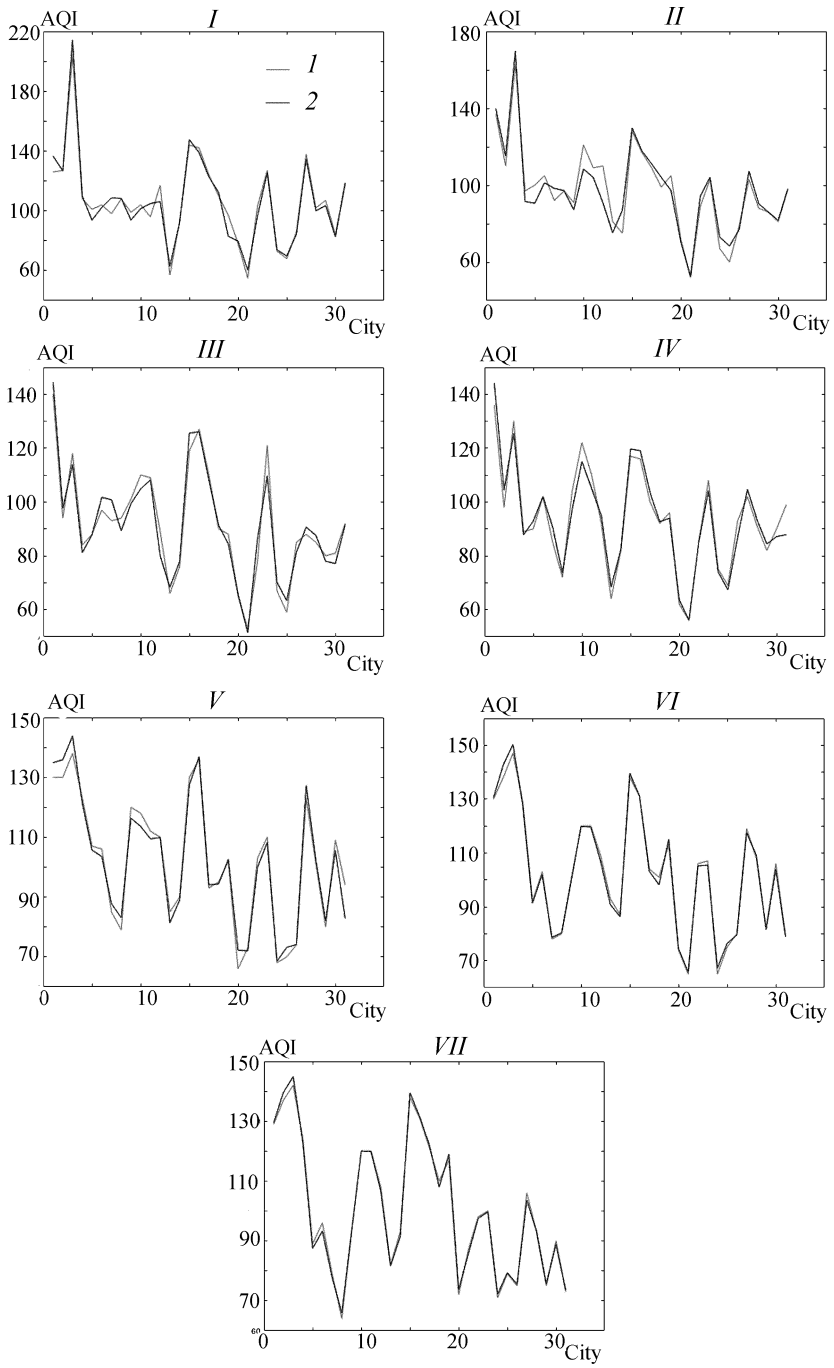

Figure 3. Final economic models at 2013 (*I*), 2014 (*II*), 2015 (*III*), 2016 (*IV*), 2017 (*V*), 2018 (*VI*) and 2019 (*VII*)
1 — actual value; *2* — predicted value.

We use big stars and short horizontal lines to represent the variables that are significant and not significant at this year. The bottom number is the sum of the total influence factors

among seven years. Therefore, it is clear to find that the control variable x_6 ($PM_{2.5}$) is the objective influence on the logarithm of the AQI, followed by x_5 (O_3).

6. Conclusions and discussions. Multiple linear regression is not only used for forecasting. It can also be used to assess which one or more are the most influential factor among controlled variables. A real case of conclusion can be drawn from the obtained result. First, evaluate the model quality based on the R^2 more than 0.8, which means the model has high quality and deserves application. However, when variables are not significant, we choose backward elimination regression to get the final model. It can be implemented in Matlab, *R*-studio, Stata, Spss, Excel and MedCale, etc. In addition, we also test the residual normality and autocorrelation is zero or not in each step until all variables are significant and get the final model.

Furthermore, it is worth thinking about there are different ways to get the results: (i) Forward selection stepwise regression; (ii) Akaike information criterion; (iii) Bayesian information criterion. These questions provide good ideas for the future research program.

References

1. Nassiri M., Elahi T. M., Ghovvati S. Evaluation of different statistical methods using SAS software: an in silico approach for analysis of real-time PCR. *Journal of Applied Statistics*, 2018, vol. 45, iss. 2, pp. 306–319.

2. Bure V. M., Petrushin A. F., Mitrofanov E. P., Mitrofanova O. A., Denisov V. Experience with the use of mathematical statistics methods for assessment of agricultural plants status. *Sel'skokhozyaistvennaya Biologiya* [*Agricultural Biology*], 2019, vol. 54, iss. 1, pp. 84–90. https://doi.org/10.15389/agrobiology.2019.1.84eng

3. Iakushev V. P., Bure V. M., Mitrofanova O. A., Mitrofanov E. P. Theoretical foundations of probabilistic and statistical forecasting of agrometeorological risks. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2021, vol. 17, iss. 2, pp. 174–182. https://doi.org/10.21638/11701/spbu10.2021.207

4. Iakushev V. P., Bure V. M., Mitrofanova O. A., Mitrofanov E. P. On the issue of semivariograms constructing automation for precision agriculture problems. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2020, vol. 16, iss. 2, pp. 177–185. https://doi.org/10.21638/11701/spbu10.2020.209

5. Ghani I. M., Ahmad S. Stepwise multiple regression method to forecast fish landing. *Procedia-Social and Behavioral Sciences*, 2010, vol. 8, pp. 549–554.

6. Bure V. M., Parilina E. M., Sedakov A. A. Metody prikladnoi statistiki v R i Excel. 3-izd. [Applied statistics methods in R and Excel]. 3rd ed. St Petersburg, Lan' Publ., 2019, 196 p. (In Russian)

7. Qi D. Study of the investment attractiveness of China's regions. *Management Processes and Sustainability*, 2020, vol. 7, iss. 1, pp. 423–427.

8. Karim S. A., Kamsani N. F. Water quality index prediction using multiple linear fuzzy regression model: Case study in Perak River, Malaysia. *Springer Nature*, 2020, pp. 31–35.

9. Adrain R. Research concerning the probabilities of the errors which happen in making observations. *George Long*, 1814, vol. 1, no. 4, pp. 93–107.

10. Merriman M. On the history of the method of least squares. *The Analyst*, 1877, vol. 4, iss. 2, pp. 33–36.

11. Zyskind G., Martin F. B. On best linear estimation and general Gauss — Markov theorem in linear models with arbitrary nonnegative covariance structure. *SIAM Journal on Applied Mathematics*, 1969, vol. 17(6), pp. 1190–1202.

12. Quandt R. E. Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association*, 1960, vol. 55, iss. 290, pp. 324–330.

13. Pope P. T., Webster J. T. The use of an *F*-statistic in stepwise regression procedures. *Technometrics*, 1972, vol. 14, iss. 2, pp. 327–340.

14. Bure V. M., Parilina E. M. *Teoriia veroiatnosti i matematicheskaia statistika* [*Probability theory and mathematical statistics*]. 1st ed., St Petersburg, Lan' Publ., 2013, 416 p. (In Russian)

15. Royston P. Approximating the Shapiro — Wilk test for non-normality. *Statistics and Computing*, 1992, vol. 2, iss. 3, pp. 117–119.

16. Wilford L. L., Taylor D. The power of four tests of autocorrelation in the linear regression model. *Journal of Econometrics*, 1975, vol. 3, iss. 1, pp. 1–21.

Received: March 19, 2022.
Accepted: September 1, 2022.

Authors' information:

Yang He — Postgraduate Student; hy1186867324@outlook.com

Dongfang Qi — Postgraduate Student; st073409@student.spbu.ru

Vladimir M. Bure — Dr. Sci. in Engineering Sciences, Professor; vlb310154@gmail.com

Новое приложение множественной линейной регрессии — случай качества воздуха в Китае

Я. Хе[1], *Д. Ци*[1], *В. М. Буре*[1,2]

[1] Санкт-Петербургский государственный университет, Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

[2] Агрофизический научно-исследовательский институт, Российская Федерация, 195220, Санкт-Петербург, Гражданский пр., 14

Для цитирования: *He Y., Qi D., Bure V. M.* New application of multiple linear regression method-A case in China air quality // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2022. Т. 18. Вып. 4. С. 516–526. https://doi.org/10.21638/11701/spbu10.2022.406

Предлагается эконометрическая модель, основанная на методе множественной линейной регрессии. Это исследование направлено на оценку наиболее важных факторов зависимой переменной. Например, рассматриваются свойства этой модели, качество модели, тест параметров, проверка остатков модели. Затем, чтобы убедиться, что модель прогнозирования оптимальна, используется метод пошаговой регрессии с обратным исключением, чтобы получить окончательную модель. В то же время также необходимо проверять свойства на каждом шаге. Наконец, результаты иллюстрируются реальным случаем качества воздуха в Китае. Полученная модель была применена для прогнозирования индекса качества воздуха (AQI) в 31 городе Китая в течение 2013–2019 гг. Все расчеты и тесты проводились с использованием *R*-studio. Величина AQI характеризует индекс качества воздуха в Китае. К контрольным переменным относятся шесть факторов загрязнения и четыре метеорологических фактора. Таким образом, модель показывает, что наиболее значительным фактором, влияющим на AQI в Китае, является $PM_{2,5}$, за которым следует O_3.

Ключевые слова: множественная линейная регрессия, загрязнение воздуха, AQI, проверка гипотез, $PM_{2.5}$, O_3.

Контактная информация:

Хе Ян — аспирант; hy1186867324@outlook.com

Ци Дунфан — аспирант; st073409@student.spbu.ru

Буре Владимир Мансурович — д-р техн. наук, проф.; vlb310154@gmail.com