

文章编号: 1009-4490(2017)02-0088-06

# 芜湖市 $PM_{2.5}$ 的影响因素分析与预测

黄婷婷, 朱家明\*, 刘丹丹

安徽财经大学统计与应用数学学院, 安徽 蚌埠 233030

**摘 要:** 本文针对芜湖市  $PM_{2.5}$  的影响因素分析与预测研究, 分别建立了多元线性回归和 ARMA 预测等模型, 使用 EVIEWS、EXCEL 等软件编程求解. 研究得出:  $PM_{10}$  和 CO 对  $PM_{2.5}$  浓度的变化影响较大, 符合芜湖市  $PM_{2.5}$  浓度变化的 ARMA(1 2) 模型. 最后为政府和环境部门有针对性地治理  $PM_{2.5}$  污染问题提供较为科学的建议.

**关键词:** 芜湖市;  $PM_{2.5}$ ; 多元线性回归; ARMA; EVIEWS

**中图分类号:** X513 **文献标识码:** A

**DOI:** 10.16207/j.cnki.1009-4490.2017.02.015

新世纪以来, 随着我国  $PM_{2.5}$  污染问题的日益严重以及社会公众环境意识的增强,  $PM_{2.5}$  污染问题已成为政府、环境保护部门及人民群众重点关注的问题之一. 为此新修订的《环境空气质量标准》将  $PM_{2.5}$  纳入空气质量 AQI 基本检测的指标之一, 并在 2016 年开始全面实施,  $PM_{2.5}$  的控制可以提高社会的经济效益, 改善人们的生活环境质量, 因此对  $PM_{2.5}$  的检测和研究十分重要. 本文通过对芜湖市的  $PM_{2.5}$  影响因素的分析和建立  $PM_{2.5}$  预测模型, 为政府和环境部门有针对性地治理  $PM_{2.5}$  污染问题提供了建议.

## 1 数据的获取与指标选取

数据来源于 2015 年芜湖市空气质量指数统计网站<sup>[1]</sup>. 除  $PM_{2.5}$  外, 空气质量 AQI 的其他 5 项基本监测指标分别是可吸入颗粒物  $PM_{10}$ 、二氧化硫  $SO_2$ 、二氧化氮  $NO_2$ 、一氧化碳 CO 和臭氧  $O_3$ . 研究表明 AQI 监测指标中的  $SO_2$ 、 $NO_2$ 、CO 和  $O_3$  等是在一定环境条件下形成  $PM_{2.5}$  前的主要气态物质, 因此本文选取的指标包括  $PM_{2.5}$  和其他 5 项基本检测监测指标.

## 2 基于多元线性回归的 $PM_{2.5}$ 影响因素分析

### 2.1 研究思路

首先, 以  $PM_{2.5}$  为因变量,  $PM_{10}$ 、 $SO_2$ 、 $NO_2$ 、CO 和  $O_3$  为自变量, 其中  $y$  表示  $PM_{2.5}$ ,  $x_1$  表示  $PM_{10}$ ,  $x_2$  表示  $SO_2$ ,  $x_3$  表示  $NO_2$ ,  $x_4$  表示 CO,  $x_5$  表示  $O_3$ . 借助 EVIEWS 软件对  $PM_{2.5}$  进行多元线性回归方程拟合, 并对模型进行检验, 分析各因素对  $PM_{2.5}$  影响程度的大小.

### 2.2 研究方法

多元线性回归模型可以描述被解释变量与多个解释变量之间的线性关系, 通过最小二乘参数估计方法得到各个回归系数, 定量研究各变量之间的因果关系<sup>[2]</sup>.

(1) 多元线性回归模型的一般表现形式

收稿日期: 2016-11-02

基金项目: 国家自然科学基金资助项目(11601001); 安徽财经大学教研发助项目(acjyzd201429).

作者简介: 黄婷婷(1995—), 女, 安徽芜湖人, 安徽财经大学统计与应用数学学院本科生, 主要从事经济统计方面的研究.

通讯作者: 朱家明(1973—), 男, 安徽泗县人, 安徽财经大学统计与应用数学学院副教授, 硕士, 主要从事应用数学与数学建模方面的研究.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_k x_{ki} + \varepsilon_i$$

其中  $y$  为解释变量  $x_1, x_2, \cdots, x_k$  表示  $k$  个解释变量  $\varepsilon$  为随机误差项.

## (2) 模型显著性检验—F 检验

对于多元回归模型  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_k x_{ki} + \varepsilon_i$  提出原假设和备择假设分别为

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad H_1: \beta_i \text{ 不全为 } 0 \quad i = 1, 2, \cdots, k$$

由于  $y_i$  服从正态分布, 故有

$$ESS = \sum (\hat{y}_i - \bar{y})^2 \sim \chi^2(k) \quad RSS = \sum (y_i - \hat{y}_i)^2 \sim \chi^2(n - k - 1)$$

构造 F 统计量用于检验原假设  $H_0$

$$F = \frac{ESS/k}{RSS/(n - k - 1)} = \frac{\sum (\hat{y}_i - \bar{y})/k}{\sum e_i^2/(n - k - 1)} \sim F(k, n - k - 1)$$

给定一个显著性水平  $\alpha$ , 可以查 F 分布表得到一个临界值  $F_\alpha$ . 在原假设  $H_0$  下是  $F > F_\alpha$  一个小概率事件, 如果根据样本数据计算得到  $F > F_\alpha$ , 则表示在“原假设  $H_0$  是正确”的假设下发生了一个小概率错误, 故拒绝原假设  $H_0$ , 接受了备择假设  $H_1$ , 说明模型的线性关系显著成立. 模型通过了方程显著性检验. 反之, 则不能拒绝原假设  $H_0$ , 即认为模型的线性关系不显著<sup>[3]</sup>.

## (3) 参数显著性检验—T 检验

解释变量  $x_i$  对被解释变量的作用显著, 说明参数  $\beta_i$  显著不为 0. 在该检验中, 原假设为  $H_0: \beta_i = 0$ , 当原假设不成立时, 解释变量对被解释变量作用显著. 当原假设成立时, 可得

$$t_{\hat{\beta}_i} = \hat{\beta}_i / s_{\hat{\beta}_i} \sim t(n - k - 1)$$

给定一个显著性水平  $\alpha$ , 一般取值 0.05, 通过查  $t$  分布表来确定检验结果. 如果发生了  $|t| > t_{\alpha/2}(n - k - 1)$ , 则在  $1 - \alpha$  水平下拒绝原假设  $H_0$ , 说明解释变量  $x_i$  是显著的, 通过变量的显著性检验. 反之, 说明解释变量  $x_i$  是不显著的, 未通过变量的显著性检验<sup>[4]</sup>.

## 2.3 数据处理

根据芜湖市 2015 年空气质量指数中 PM<sub>2.5</sub>、PM<sub>10</sub>、SO<sub>2</sub>、NO<sub>2</sub>、CO 和 O<sub>3</sub> 这 6 个指标的数据, 为了便于观察 PM<sub>2.5</sub> 与各指标间的相关关系, 分别做出 PM<sub>2.5</sub> 与空气质量 AQI 的其他 5 项基本监测指标散点图 (图 1).

由图 1 可知, PM<sub>2.5</sub> 与 PM<sub>10</sub> 间的相关性很强, PM<sub>2.5</sub> 与其他指标间相关程度相对低点, 可以初步判断 PM<sub>2.5</sub> 的变化趋势依赖于其他五项指标的变化. 接下来, 以 PM<sub>2.5</sub> 为因变量, PM<sub>10</sub>、SO<sub>2</sub>、NO<sub>2</sub>、CO 和 O<sub>3</sub> 为自变量进行多元线性回归 (图 2).

由图 2 可知, 回归方程调整的  $R^2$  为 0.960 3, 且 Prob(F-statistic) 值为 0, 模型通过显著性检验. 但  $x_3$  的  $P$  值为 0.288 3, 未通过  $t$  检验, 故将 NO<sub>2</sub> 这项指标剔除, 然后重新进行多元回归方程的拟合 (图 3).

修正后的回归方程式为  $y = -12.189 5 + 0.739 2x_1 - 0.224 6x_2 + 5.818 5x_4 + 0.108 7x_5$ , 余下的四个指标均通过  $t$  检验, 方程调整后的  $R^2$  为 0.960 7, 说明拟合效果好, 模型通过显著性检验.

对修正后的回归方程做出残差分布图, 如图 4 所示.

由图 4 可知, 三条曲线分别表示 PM<sub>2.5</sub> 序列的实际值、拟合值以及序列的残差. 可以看出, 拟合值与观测值十分接近, 说明模型的拟合预测效果较好.

## 2.4 结果分析

由模型 II 可知, NO<sub>2</sub> 指标对 PM<sub>2.5</sub> 浓度的变化影响不显著, 故将其剔除, 重新建立多元回归方程. 由残差分布图、 $t$  检验和  $F$  检验可知方程的拟合效果较好. 通过分析 PM<sub>2.5</sub> 与各变量间的相关程度和变化方向可知: PM<sub>10</sub> 和 CO 对 PM<sub>2.5</sub> 浓度的变化影响较大, O<sub>3</sub> 对 PM<sub>2.5</sub> 浓度的变化影响最弱, 且 PM<sub>2.5</sub> 和 SO<sub>2</sub> 这两个指标值的变化方向相反等.

## 3 基于 ARMA 模型的 PM<sub>2.5</sub> 预测研究

### 3.1 研究思路

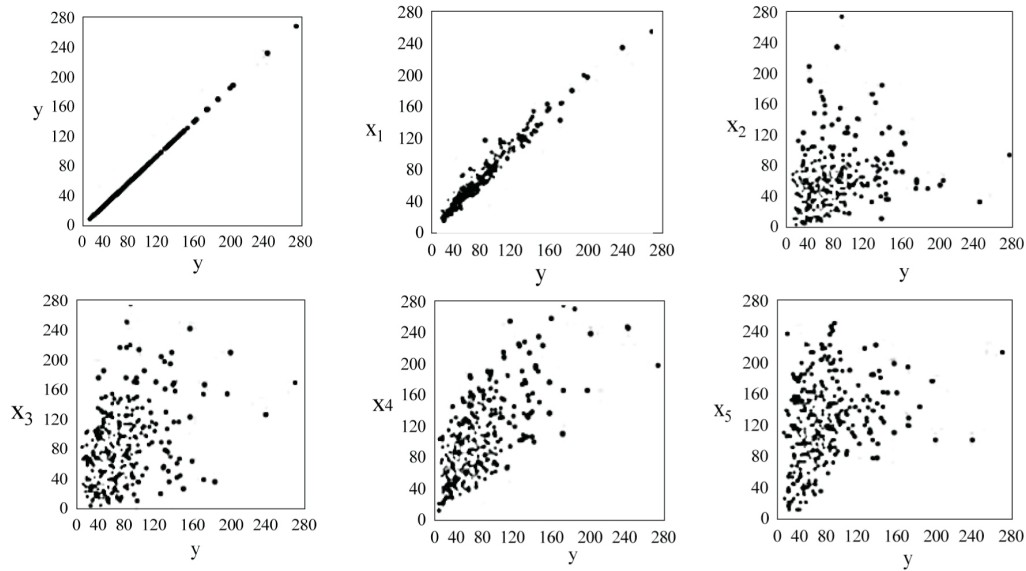


图 1 散点图

Fig. 1 Scatterplot

Variable	Coefficient	Std Error	t-Statistic	Prob
$C$	-11.580 93	1.704 872	-6.792 844	0.000 0
$x_1$	0.744 770	0.012 817	58.107 95	0.000 0
$x_2$	-0.205 928	0.041 808	-4.925 595	0.000 0
$x_3$	-0.031091	0.029 234	-1.063 521	0.288 3
$x_4$	5.389 913	1.351 599	3.987 805	0.000 1
$x_5$	0.110 750	0.037 182	2.978 561	0.003 1
R-squared	0.960 812	Mean dependent var		55.901 37
Adjusted R-squared	0.960 267	S. D. dependent var		34.857 81
S. E of regression	6.948 289	Akaike info criterion		6.731 170
Sum squared resid	17 332.06	Schwarz criterion		6.795 278
Log likelihood	-1 222.438	F-statistic		1 760.413
Durbin-Watson stat	1.005 581	Prob(F-statistic)		0.000 000

图 2 回归分析图

Fig. 2 Regression analysis diagram

Variable	Coefficient	Std Error	t-Statistic	Prob
$C$	-12.189 53	1.606 256	-7.588 787	0.000 0
$x_1$	0.739 213	0.011 706	63.148 50	0.000 0
$x_2$	-0.224 557	0.037 969	-5.914 288	0.000 0
$x_4$	5.818 534	1.290 349	4.509 273	0.000 0
$x_5$	0.108 746	0.037 141	2.927 904	0.003 6
R-squared	0.960 689	Mean dependent var		55.901 37
Adjusted R-squared	0.960 252	S. D. dependent var		34.857 81
S. E of regression	6.949 554	Akaike info criterion		6.728 836
Sum squared resid	17 386.67	Schwarz criterion		6.782 259
Log likelihood	-1 223.013	Hannan-Quinn criterion		6.750 067
F-statistic	2 199.432	Durbin-Watson stat		1.010 276
Prob(F-statistic)	0.000 000			

图 3 修正后回归分析图

Fig. 3 The revised regression analysis diagram

首先,通过 PM<sub>2.5</sub>序列图和 ADF 检验得出原始序列是平稳的;其次,根据序列的自相关图和偏自相关图都没有明显的截尾性确定使用 ARMA 模型,并根据 AIC 准则确定未知参数  $p, q$ ;然后,对模型进行参数估计与检验,并给出模型的表达式;最后,运用 ARMA 模型,对芜湖市 2015 年 1 月 1 日至 12 月 31 日 PM<sub>2.5</sub> 浓度进行拟合,通过与实际值进行比较,验证了模型的准确性。

### 3.2 研究方法

**时间序列预测**是通过时间序列的历史数据,探寻现象随时间变化的规律,进而预测现象未来走势的方法。ARMA 模型是时间序列分析中最常用的模型之一,该模型可分解为两种基本模型:自回归模型 AR 和移动平均模型 MA,这三种基本模型都仅适用于平稳时间序列<sup>[5]</sup>。

#### (1) 自回归模型 AR( $p$ )

自回归模型 AR 描述了时间序列数据内部的线性回归关系,它是用现时的干扰和有限项过去的观测值来预测当前值。 $p$  阶自回归模型 AR( $p$ ) 的表达式为

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \mu_t$$

其中  $\varphi_i (i = 1, 2, \cdots, p)$  为模型的待定系数,随机扰动项  $\mu_t$  是一个白噪声( $\mu_t = \varepsilon_t$ ),  $Y_t$  为一个时间序列。

#### (2) 移动平均模型 MA( $q$ )

移动平均模型 MA 的含义是用现时的干扰及过去各个时期干扰的有限项的线性组合来预测当前值。 $q$  阶移动平均模型 MA( $q$ ) 的数学表达方式如下:

$$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

其中  $\theta_i (i = 1, 2, \cdots, q)$  为模型的待定系数,  $\varepsilon_t$  是一个白噪声( $\mu_t = \varepsilon_t$ )。

#### (3) 自回归移动平均模型 ARMA( $p, q$ )

自回归移动平均模型 ARMA 以 AR 和 MA 模型为基础组合而成,数学表达式为

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

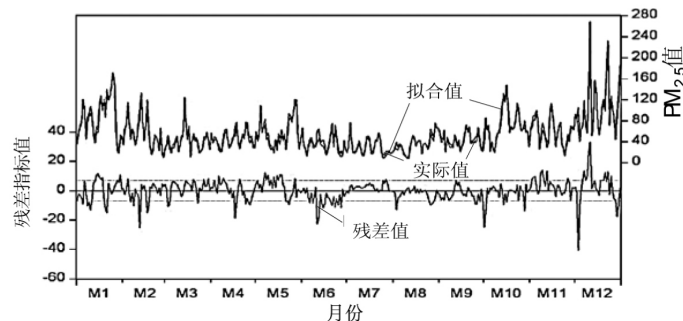


图4 残片分布

Fig. 4 Residual distribution

### 3.3 数据处理

首先,用 EViews 软件画出芜湖市 2015 年 1 月 1 日至 12 月 31 日 PM<sub>2.5</sub> 浓度的序列图(图 5),可以大致认为该序列没有明显的趋势和周期性,初步判断该序列是平稳的。

然后采用 ADF 检验对 PM<sub>2.5</sub> 序列进行平稳性检验,若数据通过  $t$  检验,则说明原始序列是平稳的,结果见表 1。

由表 1 可知,ADF 检验的  $t$  统计量值为 -5.128 9, 1%、5%、10% 的显著性水平临界值分别为 -3.448 3、-2.869 3、-2.571 0,  $t$  统计量值小于显著性水平下的

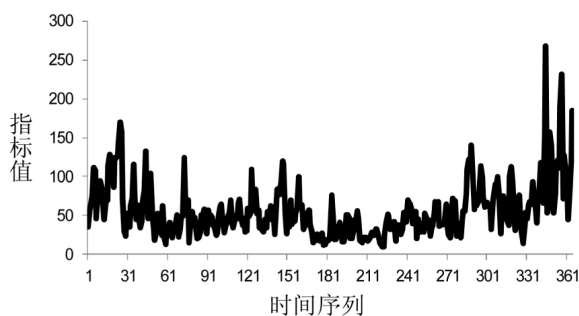


图5 PM<sub>2.5</sub>时间序列图

Fig. 5 PM<sub>2.5</sub> time sequence diagram

临界值,故拒绝原假设,认为  $PM_{2.5}$  序列不存在单位根,即序列平稳。

运用 EViews 软件绘出  $PM_{2.5}$  序列的自相关图和偏自相关图(图 6)。

由图 6 可知,  $PM_{2.5}$  序列的自相关图与偏自相关图都没有明显的截尾性,故需要建立 ARMA 模型。对于未知参数  $p、q$  的求解,本文采取最佳准则函数定阶法<sup>[6]</sup>,即 AIC 准则。AIC 准则是在模型参数极大似然估计的基础上,选取使得 AIC 达到最小的那一组参数为理想参数。基于该准则经过反复试验,求得使 AIC 达到最小的模型为  $ARMA(1, 2)$ 。

由表 2 可知,  $ARMA(1, 2)$  模型调整后的  $R^2$  为 0.507 853,而 AIC 和 SC 值分别为 9.244 2 和 9.287 1,其对应的模型表达式为  $x_t = 62.179\ 2 + 0.984\ 3x_{t-1} - 0.357\ 6\varepsilon_{t-1} - 10.433\ 9\varepsilon_{t-2}$ 。由图 7 可知,表 2 中的根都在单位圆内,说明该序列是平稳的。

运用已建立好的  $ARMA(1, 2)$  模型,通过 EViews 软件对芜湖市 2015 年 1 月 1 日至 12 月 31 日  $PM_{2.5}$  浓度进行拟合(图 8)。由图 8 可知,三条曲线分别表示  $PM_{2.5}$  序列的实际值、拟合值以及序列的残差。可以看出,拟合值与观测值较接近,模型的拟合预测效果好。

3.4 结果分析

通过对芜湖市 2015 年 1 月 1 日至 12 月 31 日的  $PM_{2.5}$  浓度时间序列进行模型的建立和实证分析,得到符合芜湖市  $PM_{2.5}$  浓度变化的  $ARMA(1, 2)$  模型,该模型的预测精度

表 1 序列  $PM_{2.5}$  的 ADF 检验

Tab.1 ADF test of sequence  $PM_{2.5}$

	t-Statistic	Prob
Augmented Dickey-Fuller test statistic	-5.128 861	0.000 0
1 % level	-3.448 211	
Test critical values:	5 % level -2.869 307	
	10 % level -2.570975	

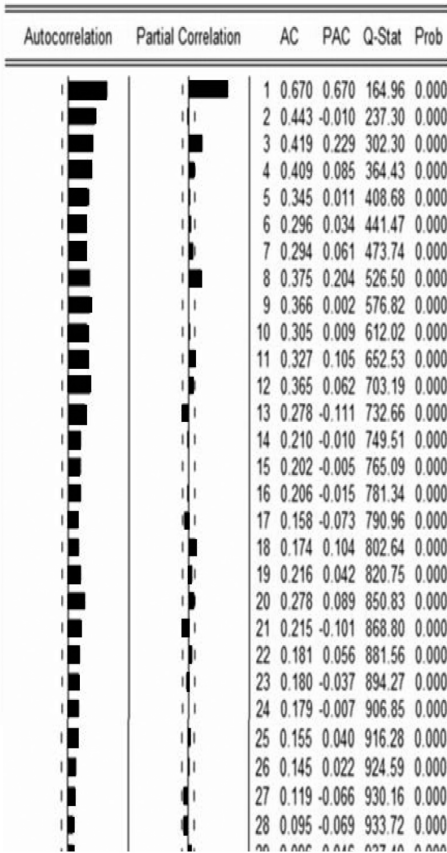


图 6  $PM_{2.5}$  序列自相关与偏自相关分析图

Fig.6 The analysis diagram of  $PM_{2.5}$  sequence autocorrelation and partial autocorrelation

表 2  $ARMA(1, 2)$  模型结果

Tab.2  $ARMA(1, 2)$  model results

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	62.179 15	19.007 83	3.271 238	0.001 2
AR(1)	0.984 289	0.016 677	59.021 11	0.000 0
MA(1)	-0.357 632	0.052 114	-6.862 541	0.000 0
MA(2)	-0.433 865	0.051 354	-8.448 533	0.000 0
R-squared	0.511 920	Mean dependent var		55.956 04
Adjusted R-squared	0.507 853	S. D. dependent var		34.890 12
S. E. of regression	24.476 53	Akaike info criterion		9.244 234
Sum squared resid	215 676.1	Schwarz criterion		9.287 060
Log likelihood	-1 678.451	Hannan-Quinn criter		9.261 256
F-statistic	125.861 6	Durbin-Watson stat		1.902 738
Prob(F-statistic)	0.000 000			
Inverted AR Roots	.98			
Inverted MA Roots	.86	-.50		

较高,故可以通过该模型对芜湖市 PM<sub>2.5</sub> 浓度进行预测,了解芜湖市短期内 PM<sub>2.5</sub> 浓度变化的基本特征。

#### 4 结束语

本文从定量分析的角度,建立多元线性回归模型,分析影响 PM<sub>2.5</sub> 浓度的主要因素,得出 PM<sub>10</sub> 和 CO 对 PM<sub>2.5</sub> 浓度的变化影响较大, O<sub>3</sub> 对 PM<sub>2.5</sub> 浓度的变化影响最弱, PM<sub>2.5</sub> 和 SO<sub>2</sub> 这两个指标值的变化方向相反等结论。建立 PM<sub>2.5</sub> 的 ARMA 预测模型,得到符合芜湖市 PM<sub>2.5</sub> 浓度变化的 ARMA(1,2) 模型。我们可以将这两个模型进行横向和纵向的推广,通过模型 I 分析 PM<sub>2.5</sub> 与经济、气象等因素间的关系,得到引起 PM<sub>2.5</sub> 浓度增大的主要因素,从根源上治理 PM<sub>2.5</sub> 污染问题;通过模型 II 对 PM<sub>2.5</sub> 进行短期预测,提前做好预防准备,为人们的生活和生产活动带来便利。

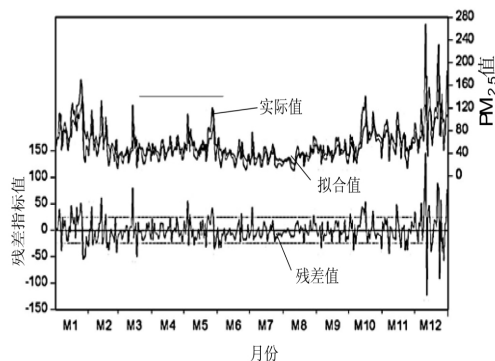


图7 单位根检验

Fig. 7 Unit root test

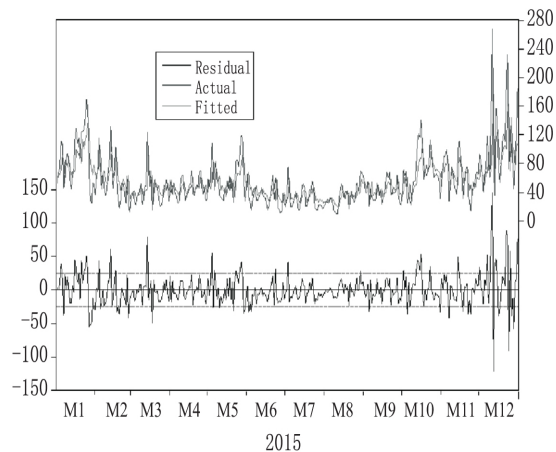


图8 模型拟合图

Fig. 8 Model fitting diagram

#### 参考文献:

- [1] 芜湖空气质量指数[EB/OL]. <http://www.tianqihoubao.com/aqi/wuhu.html> 2016-10-28.
- [2] 张云云,朱家明,高子云,等. 基于多元线性回归的 PM<sub>2.5</sub> 含量的影响因素研究[J]. 西昌学院学报, 2016, 30(1): 17~20.
- [3] 宣琳,李勇,朱家明. 蚌埠市 PM<sub>2.5</sub> 影响因素的定量研究[J]. 通化师范学院学报, 2016, 37(3): 24~26.
- [4] 王翠云,胡学平,相旭东. PM<sub>2.5</sub> 影响因素的主成分回归分析与预测[J]. 安庆师范学院学报, 2015, 11(4): 24~27.
- [5] 王丽娜,肖冬荣. 基于 ARMA 模型的经济非平稳时间序列的预测分析[J]. 武汉理工大学学报, 2004, 28(1): 133~136.
- [6] 刘玲,宋马林. 基于 ARMA 模型的南京市 PM<sub>2.5</sub> 浓度分析与预测[J]. 枣庄学院学报, 2016, 33(4): 54~62.
- [7] 苏婷,董胜伟,黄玲玲,等. 安阳市空气质量的评价与预测模型[J]. 廊坊师范学院学报, 2015, 15(4): 64~76.
- [8] 冯盼,曹显兵. 基于 ARMA 模型的股价分析与预测的实证研究[J]. 数学的实践与认识, 2011, 41(22): 85~89.
- [9] RamuR. 应用计量经济学[M]. 薛菁睿译. 北京:机械工业出版社, 2003. 237~238.
- [10] 茆诗松,程依明. 概率论与数理统计教程[M]. 北京:高等教育出版社, 2011. 236~240.
- [11] 杨桂元,朱家明. 数学建模竞赛优秀论文评析[M]. 合肥:中国科技大学出版社, 2013. 54~60.
- [12] 杨桂元. 数学建模[M]. 上海:上海财经大学出版社, 2015. 32~48.

### Analyzing and Forecasting the Influence Factor of PM<sub>2.5</sub> in Wuhu

HUANG Ting-ting, ZHU Jia-ming\*, LIU Dan-dan

(School of Statistics and Applied Mathematics,

Anhui University of Finance and Economics, Bengbu 233030, Anhui, China)

**Abstract:** In view of the research on analyzing and forecasting the influence factor of PM<sub>2.5</sub> in Wuhu, multiple linear regression and ARMA prediction model were established, respectively. Using EViews, EXCEL and other software programming solution, it concluded that the effects of PM<sub>10</sub> and CO on the change of PM<sub>2.5</sub> concentration were significant, which conformed to the ARMA(1,2) model PM<sub>2.5</sub> concentration changes in Wuhu. Finally some scientific suggestions were put forward for the government and environmental departments to deal with the problem of PM<sub>2.5</sub> pollution.

**Key words:** Wuhu; PM<sub>2.5</sub>; multiple linear regression; ARMA model; EViews