Advances in Applied Mathematics 应用数学进展, 2018, 7(6), 731-742
Published Online June 2018 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2018.76088

Hans汉斯

Regression Analysis of Air Quality Data in Xi'an

Jinyu Tang, Xiang Gao*

School of Mathematical Sciences, Ocean University of China, Qingdao Shandong
Email: 18354222836@163.com, *gaoxiangshuli@126.com

Received: Jun. 3rd, 2018; accepted: Jun. 19th, 2018; published: Jun. 26th, 2018

Abstract

Air is the condition for human beings and organisms to survive, so clean air is especially important to people. However, in recent years, with the development of China's industry and transportation industry, a large number of pollutants have been discharged into the air, the quality of air is getting worse and worse, and the problem of air quality has been paid more and more attention by the government and the public. In order to explore the close correlation between air quality and which pollutants and the relationship between them, this article selected Xi'an as an example to collect Xi'an air quality data from December 2013 to April 2018, mainly including air mass index (AQI) and PM2.5, PM10, SO_2, CO, O_3 and so on. R software was used for regression analysis. First, the air quality data of December 2013-2017 year June are used to model the model, and the model is tested. Then the data of April July 2107-2018 are used to predict the AQI index, in order to test the model.

Keywords

Air Quality, Regression Analysis, Significance Test, The Prediction of AQI Index

西安空气质量数据的回归分析

唐金玉，高　翔*

中国海洋大学数学科学学院，山东　青岛
Email: 18354222836@163.com, *gaoxiangshuli@126.com

收稿日期：2018年6月3日；录用日期：2018年6月19日；发布日期：2018年6月26日

摘　要

空气是人类及生物赖以生存的条件，因此清洁的空气对人类显得尤为重要。但是，近年来随着我国工业

*通讯作者。

文章引用: 唐金玉, 高翔. 西安空气质量数据的回归分析[J]. 应用数学进展, 2018, 7(6): 731-742.
DOI: 10.12677/aam.2018.76088

和交通运输业的发展，大量的污染物被排放到空气中，空气质量每况愈下，空气质量问题也越来越受到政府和公众的关注。为了探究空气质量与哪些污染物密切相关以及它们之间的关系，本文选取西安作为实例，收集了2013年12月~2018年4月西安空气质量数据[1]，主要包括空气质量指数(AQI)及PM2.5、PM10、SO_2、CO、O_3等含量，应用R软件进行回归分析。首先利用2013年12月~2017年6月的空气质量数据进行建模，并对模型进行显著性检验，然后利用2107年7月~2018年4月的数据进行AQI指数的预测，以此来检验模型的好坏。

关键词

空气质量，回归分析，显著性检验，AQI指数预测

1. 引言

工业的快速发展与城市化导致我国的空气污染形势日益严峻。近年来，我国有越来越多的城市被雾霾问题所困扰，这一问题逐渐引起了国内外媒体与广大学者的关注，我国也开始在治理空气污染方面加大投入。党的十九大报告指出，我们要着力解决突出环境问题。国内外学者用不同的方法对我国的空气质量情况进行了分析。李丹[2]基于聚类分析和多元回归的思想对空气质量问题进行了研究，建立了回归模型；李晓童[3]等人，基于 Bootstrap 方法，对北京市空气质量的影响因素进行回归分析及预测；肖正[4]等人基于多元线性回归分析对合肥市的空气质量进行了实证研究。基于以上研究，本文利用回归分析对西安空气质量数据进行了分析。首先利用 2013 年 12 月~2017 年 6 月的空气质量数据进行建模，并对模型进行显著性检验，然后利用 2107 年 7 月~2018 年 4 月的数据进行 AQI 指数的预测，以此来检验模型的好坏。

2. 回归分析建模

回归分析[5]是研究变量间函数关系的一种方法。

2.1. 建立一般多元线性回归模型

假设 AQI 为响应变量，PM2.5、PM10、SO_2、CO、NO_2、O_3 为预测变量，响应变量与各预测变量的建模过程如下：

1) 散点图

如图 1 的散点图可以看出，响应变量与各预测变量之间大致呈线性关系，因此建立一般多元线性回归模型。

2) 拟合模型

如表 1，从拟合结果来看，只有变量 PM2.5，PM10 和 O_3 的系数显著，其他几个参数估计结果均不显著，R-squared = 0.9701 是一个很大的值，F 检验的 p-value: < 2.2e−16 是一个很小的值，这说明模型可能存在过拟合的情况。

3) 回归诊断

绘制上述模型的边际模型图：

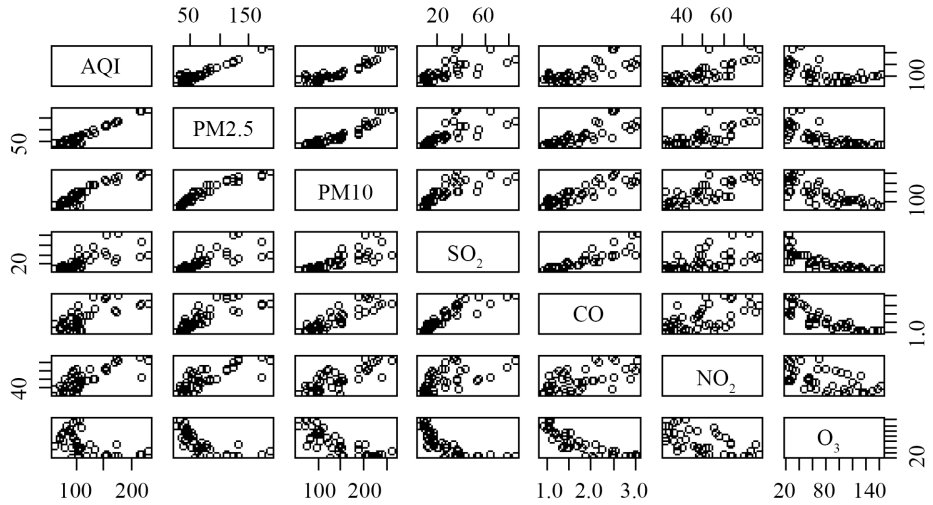

Figure 1. Scatter plot
图 1. 散点图

Table 1. Model fitting result
表 1. 模型拟合结果

Call:

lm(formula = AQI ~ PM2.5 + PM10 + SO_2 + CO + NO_2 + O_3, data = air)

Residuals:

Min	1Q	Median	3Q	Max
−11.9898	−3.9216	−0.2669	3.6843	18.2069

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−13.52998	10.94688	−1.236	0.2245
PM2.5	0.86999	0.08431	10.319	2.67e−12 ***
PM10	0.22467	0.07914	2.839	0.0074 **
SO_2	0.10969	0.12524	0.876	0.3869
CO	1.84763	5.41873	0.341	0.7351
NO_2	0.01416	0.14484	0.098	0.9227
O_3	0.33644	0.04969	6.770	6.59e−08 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.118 on 36 degrees of freedom

Multiple R-squared: 0.9744, Adjusted R-squared: 0.9701

F-statistic: 228.1 on 6 and 36 DF, p-value: < 2.2e−16

如图 2，边际模型图显示各个变量的拟合效果比较好。

绘制上述模型的诊断图(图 3)。

如图 3 拟合模型的诊断图：用 Q-Q 图判断正态性，若满足正态假设，那么图上的点应该落在呈 45 度角的直线上[6]，可以看出模型的标准化残差不满足正态性；用位置尺度图判断同方差性，若满足不变方差假设，那么在“位置尺度图”中，水平线周围的点应该随机分布[6]，该图不满足同方差性假设。

进一步绘制上述回归模型的变量添加图评估一下每个变量对其他变量的影响。

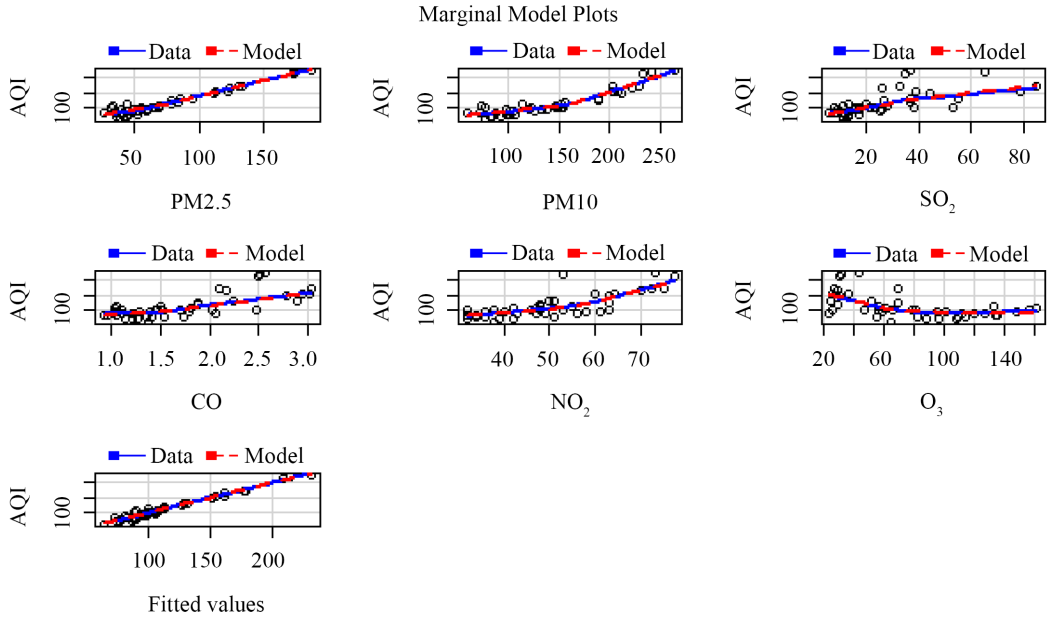

Figure 2. Marginal model diagram
图 2. 边际模型图

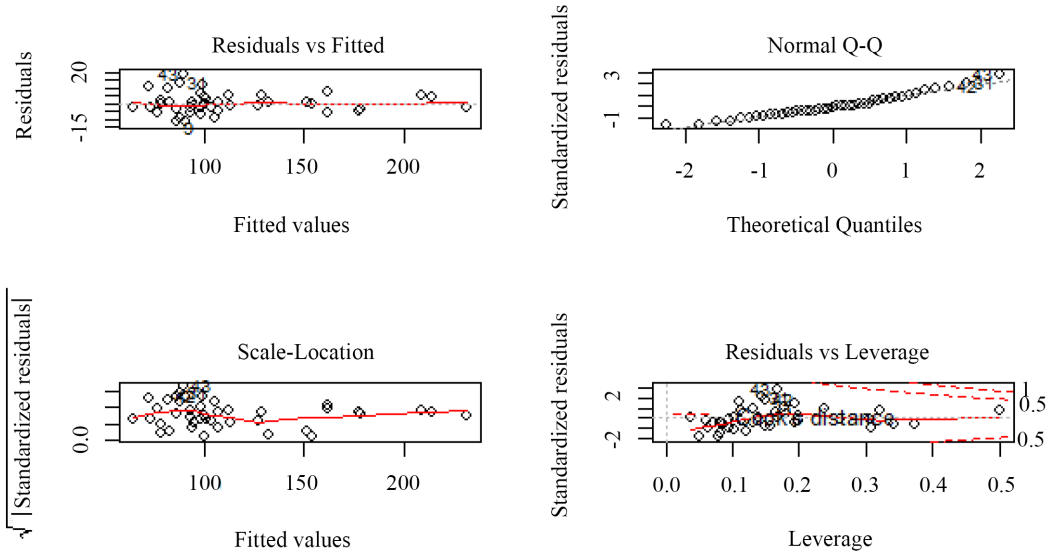

Figure 3. Regression diagnosis chart
图 3. 回归诊断图

如图 4 变量添加图可以看出，PM2.5，PM10 和 O_3 对其他变量影响较大，NO_2，SO_2 和 CO 对其他变量的影响都不是太大，这也与估计结果中 NO_2，SO_2 和 CO 的系数估计的结果不显著相对应，所以我们考虑将这三个变量去除之后再做模型的拟合。

如表 2 模型拟合结果来看，去掉变量后所建模型中变量的系数均显著，并且 R-squared 没有减小且 F 检验的 p-value 依旧显著。绘制这个模型的标准化残差诊断图。

如图 5 回归诊断图可以看出，虽然模型拟合较好但是诊断图形呈现的问题仍然没有得到解决。于是认为，一般的线性模型不足以表达变量间的关系，我们考虑对变量进行变换继续建立更加有效地模型。

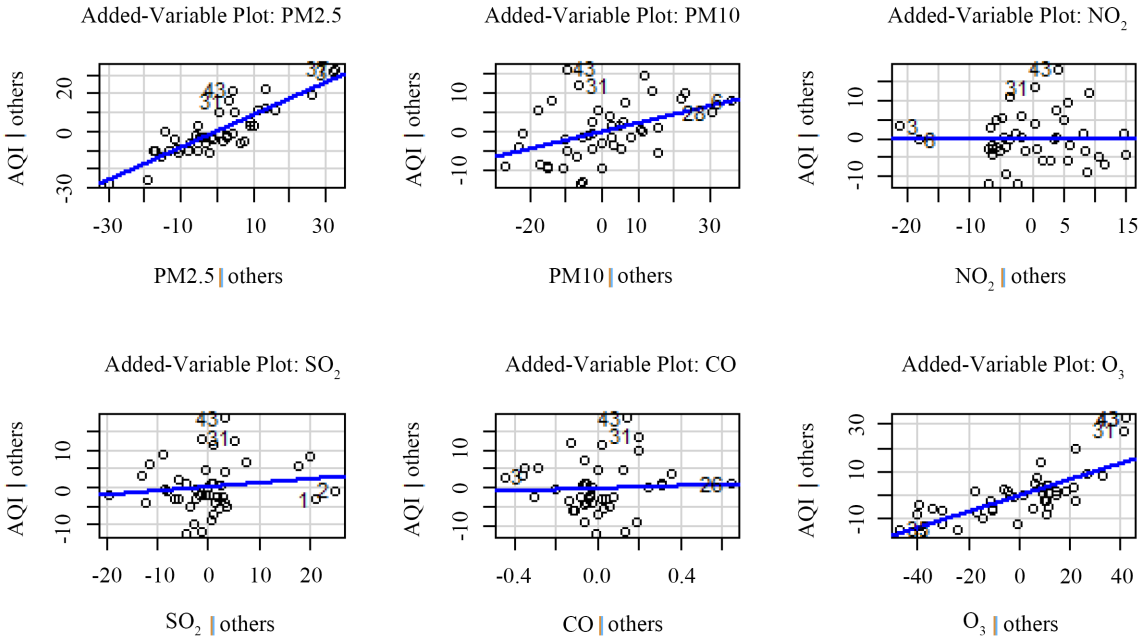

Figure 4. Variable addition graph
图 4. 变量添加图

Table 2. Model fitting result
表 2. 模型拟合结果

Call:				
lm(formula = AQI ~ PM2.5 + PM10 + O_3)				
Residuals:				
Min	1Q	Median	3Q	Max
−12.2394	−4.0724	−0.1166	3.5825	19.8939
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−9.12503	7.90795	−1.154	0.255561
PM2.5	0.87644	0.08104	10.815	2.66e-13 ***
PM10	0.25403	0.06878	3.693	0.000677 ***
O_3	0.30719	0.04108	7.477	4.79e−09 ***

Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			
Residual standard error: 7.021 on 39 degrees of freedom				
Multiple R-squared:	0.973,	Adjusted R-squared:	0.9709	
F-statistic: 468.4 on 3 and 39 DF,	p-value: < 2.2e−16			

2.2. 变换变量建模

首先，我们从验证变量为正态分布的假设入手，我们需要画出各变量高斯核密度估计图、箱线图、QQ 图。

从图 6~12 中观察可得各变量数据存在偏态非正态的情况，所以考虑对其预测变量和相应变量同时进行 BOX-COX 变换。

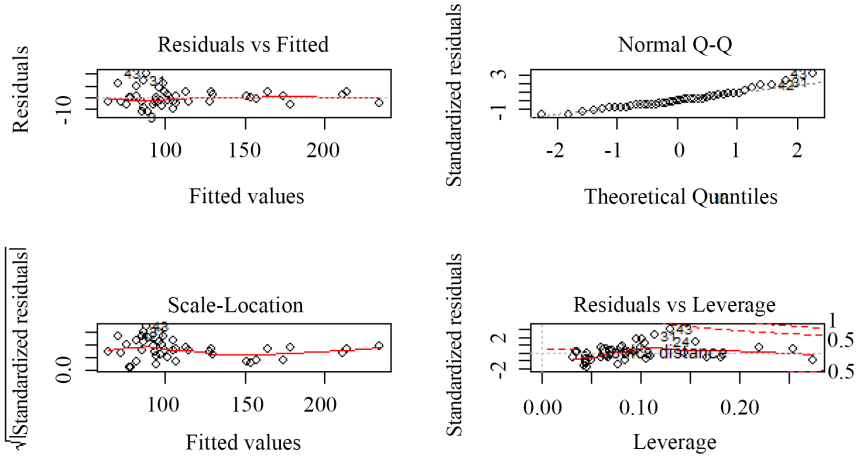

Figure 5. Regression diagnosis chart
图 5. 回归诊断图

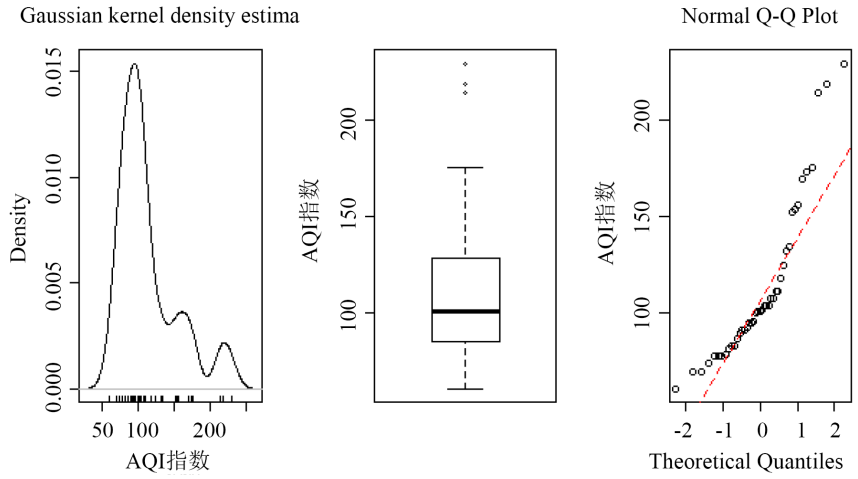

Figure 6. AQI: normal distribution test
图 6. AQI 正态分布检验

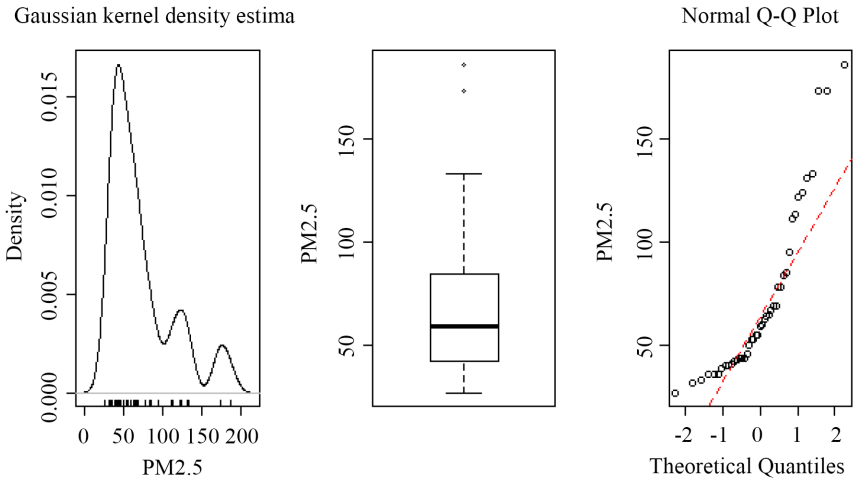

Figure 7. PM2.5: normal distribution test
图 7. PM2.5 正态分布检验

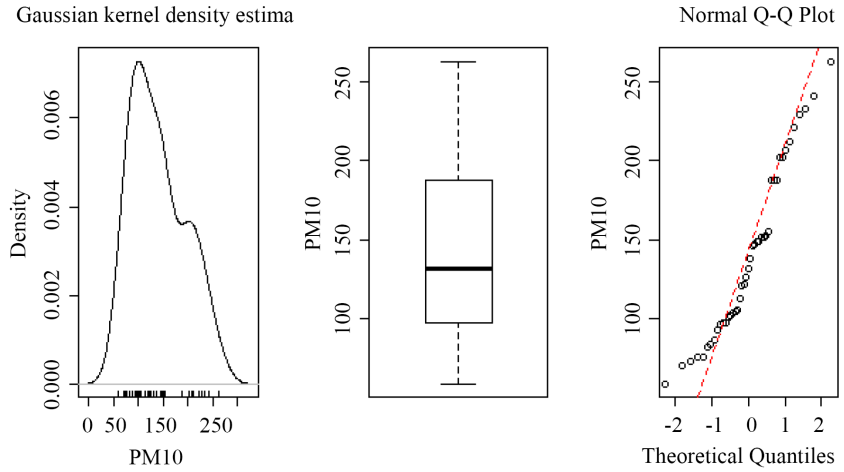

Figure 8. PM10: normal distribution test
图 8. PM10 正态分布检验

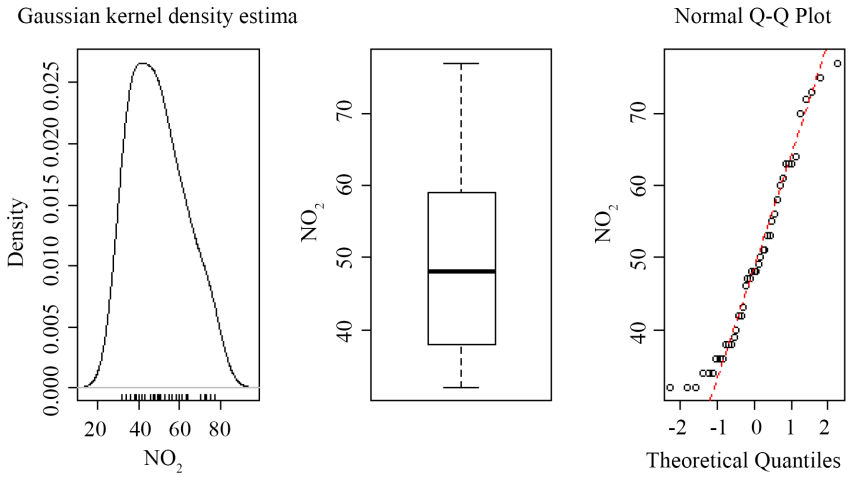

Figure 9. NO_2: normal distribution test
图 9. NO_2 正态分布检验

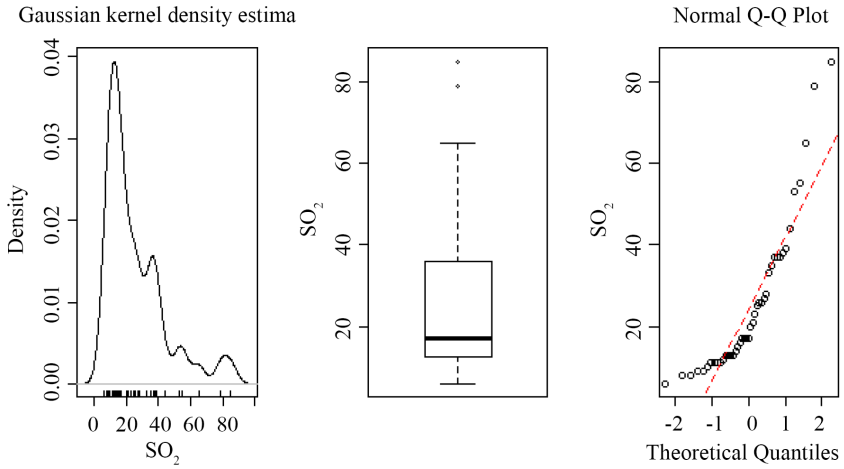

Figure 10. SO_2: normal distribution test
图 10. SO_2 正态分布检验

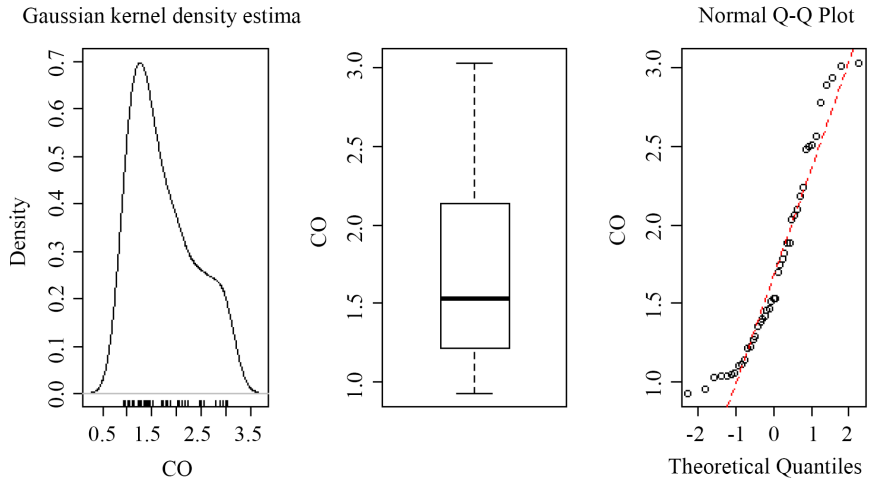

Figure 11. CO: normal distribution test
图 11. CO 正态分布检验

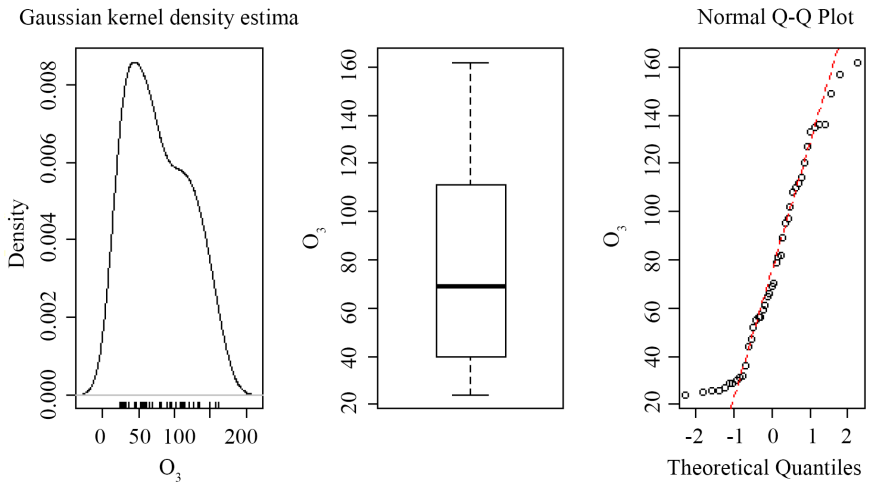

Figure 12. O_3: normal distribution test
图 12. O_3 正态分布检验

如表 3 的结果显示对变量进行如下变换：log(AQI)，log(PM2.5)，$1/\sqrt{SO_2}$，1/CO 对变换后的变量进行建模。

从表 4 的模型拟合结果来看，变量 NO_2，tSO_2，tCO 系数不显著，这有可能是由多重共线性引起的，下面进行多重共线性的判别。

tPM2.5	PM10	NO_2	tSO_2	tCO	O_3
28.211357	24.069357	3.452873	10.288414	14.752069	6.674461

由上面的多重共线性的输出结果可以看出，方差膨胀因子存在大于 10 的值，因此数据存在多重共线性。下面采用向后剔除法进行变量选择。

由表 5 可以看出，变量选择后的模型中各个变量都是显著的。

绘制上述模型的诊断图(图 13)。

由图 13 的回归诊断图可以看出残差与拟合图中的点更加随机的分布在水平线周围，说明建立的模型合适。位置尺度图中的点也随机分布在水平线周围满足同方差性。因此该模型的拟合效果比较好。

Table 3. BOX-COX transformation results
表 3. BOX-COX 变换结果

bcPower Transformations to Multinormality				
	Est Power	Rounded Pwr	Wald Lwr Bnd	Wald Upr Bnd
AQI	−0.1534	0.0	-0.7781	0.4714
PM2.5	−0.2468	0.0	-0.6046	0.1109
PM10	0.9526	1.0	0.4643	1.4409
NO_2	0.1247	1.0	−1.0050	1.2543
SO_2	−0.3271	−0.5	−0.6019	−0.0523
CO	−0.6059	−1.0	−1.1548	−0.0570
O_3	1.0138	1.0	0.5136	1.5140
Likelihood ratio test that transformation parameters are equal to 0				
(all log transformations)				
			LRT df	pval
LR test, lambda = (0 0 0 0 0 0 0)			49.20338 7	2.07e-08
Likelihood ratio test that no transformations are needed				
			LRT df	pval
LR test, lambda = (1 1 1 1 1 1 1)			133.5278 7	< 2.22e−16

Table 4. Model fitting result
表 4. 模型拟合结果

Call:				
lm(formula = tAQI ~ tPM2.5 + PM10 + NO2 + tSO2 + tCO + O3)				
Residuals:				
Min	1Q	Median	3Q	Max
−0.14048	−0.05670	−0.02133	0.04512	0.23899
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.2113008	0.4725635	4.679	3.99e−05 ***
tPM2.5	0.3727720	0.1483575	2.513	0.01660 *
PM10	0.0043372	0.0012748	3.402	0.00165 **
NO_2	0.0001276	0.0020187	0.063	0.94995
tSO_2	−0.1923691	0.6134640	−0.314	0.75565
tCO	0.1674015	0.2425960	0.690	0.49459
O_3	0.0031013	0.0008675	3.575	0.00102 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error: 0.09108 on 36 degrees of freedom				
Multiple R-squared: 0.9318, Adjusted R-squared: 0.9205				
F-statistic: 82.02 on 6 and 36 DF, p-value: < 2.2e−16				

Table 5. Variable selection results
表 5. 变量选择结果

Call:
lm(formula = tAQI ~ tPM2.5 + PM10 + O3)
Residuals:
Min 1Q Median 3Q Max
−0.14596 −0.05394 −0.01348 0.03781 0.23737
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.187334 0.409397 5.343 4.22e−06 ***
tPM2.5 0.399506 0.128671 3.105 0.00354 **
PM10 0.003997 0.001127 3.546 0.00104 **
O_3 0.003511 0.000544 6.455 1.21e−07 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.08817 on 39 degrees of freedom
Multiple R-squared: 0.9308, Adjusted R-squared: 0.9255
F-statistic: 174.8 on 3 and 39 DF, p-value: < 2.2e−16

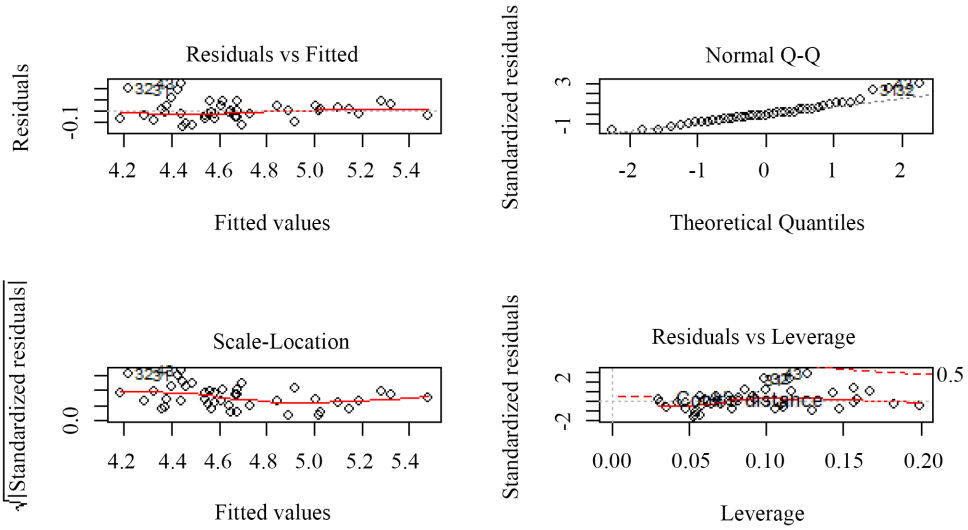

Figure 13. Regression diagnosis chart
图 13. 回归诊断

绘制上述模型的边际模型图(图 14)。

由图 14 的边际模型图也可以看出拟合效果很好，因此该模型为有效模型。

综上，最终确定模型为：

$$\log(\mathrm{AQI}) = 0.399506\log(\mathrm{PM2.5}) + 0.003997\mathrm{PM10} + 0.003511\mathrm{O}_3 + 2.187334$$

3. 预测

得到的预测结果如下：

如表 6 的预测结果显示真实值全在预测区间内，预测效果很好，由此进一步证明了模型的有效性。

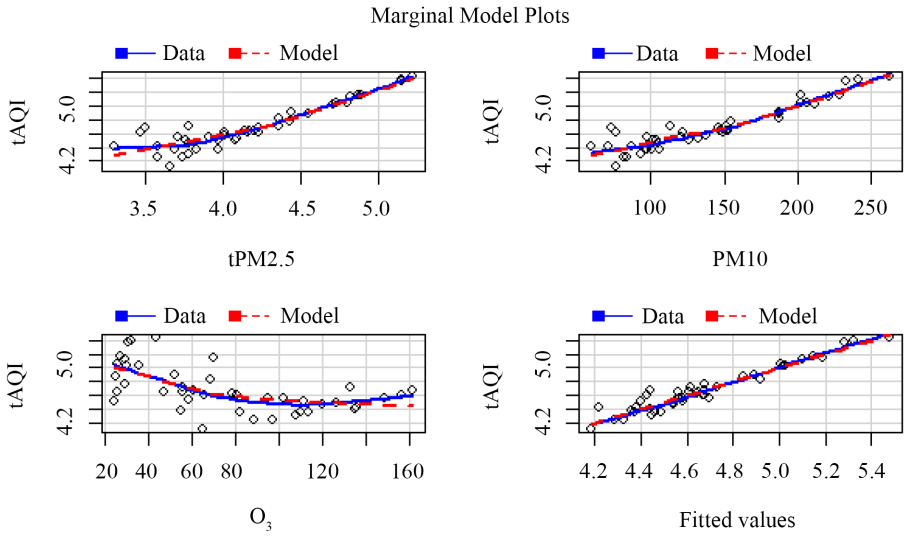

Figure 14. Marginal model diagram
图 14. 边际模型图

Table 6. Forecast result data
表 6. 预测结果数据

真实值	预测值	预测区间下限	预测区间上限	真实值是否在预测区间内
4.662174	4.479	4.286	4.671	是
4.564348	4.401	4.215	4.588	是
4.343805	4.386	4.202	4.57	是
4.330733	4.262	4.065	4.458	是
4.867534	4.775	4.586	4.964	是
5.087596	4.993	4.81	5.177	是
5.204007	5.089	4.886	5.293	是
4.828314	4.839	4.643	5.035	是
4.875197	5.001	4.81	5.192	是
4.624973	4.717	4.523	4.911	是

4. 总结

本文首先建立一般多元线性回归模型，通过模型诊断发现一般多元线性回归模型不能很好的反映响应变量与各预测变量之间的关系，拟合效果并不好，然后通过变量变换和变量选择，建立了变换后的模型，通过回归诊断发现模型的诊断效果比较好，确定了最终模型，最后用最终确定的模型进行了预测，发现预测，将预测值与真实值进行比较，发现预测效果很好，进一步证明建立的最终模型有效。

基金项目

山东省自然科学基金(ZR2018MA006)，山东省研究生教育创新计划项目(SDYY15129)，山东省研究生导师指导能力提升项目(SDYY17009)。

参考文献

[1] 西安空气质量指数 AQI-PM2.5 月统计历史数据-中国空气质量在线监测分析平台历史数据[EB/OL].

https://www.aqistudy.cn/historydata/monthdata.php?city=%E8%A5%BF%E5%AE%89

[2] 李丹. 基于聚类分析和多元回归的空气质量的分析[D]: [硕士学位论文]. 天津: 南开大学数学科学学院, 2015.

[3] 李晓童, 夏明月, 林善冬. 基于 Bootstrap 方法对北京空气质量的回归分析[J]. 河北北方学院学报(自然科学版), 2014, 30 (4): 31-34.

[4] 肖正, 祁孟阳, 朱家明. 基于多元线性回归模型的合肥市空气质量实证分析[J]. 兰州文理学院学报(自然科学版), 2017, 31(4): 13-19.

[5] Samprit Chatterjee, Ali S. Hadi. 例解回归分析[M]. 北京: 机械工业出版社, 2013.

[6] Robert I. Kabacoff. R 语言实战[M]. 北京: 人民邮电出版社, 2016.