基于多元线性回归模型的邯郸市空气质量分析

刘欣欣 安笑洁 余志鹏

（指导教师：张艳萍）

河北工程大学

一、研究背景概述

随着城市工业化的发展，各种尾气、废气的排放让空气质量问题成了即温室效应之后，人们关注的又一大热点。空气质量的好坏反映了空气污染程度，它是依据空气中污染物浓度的高低来判断的。空气污染是一个复杂的现象，在特定时间和地点空气污染物浓度受到许多因素影响。来自固定和流动污染源的人为污染物排放大小是影响空气质量的最主要因素之一，其中包括车辆、船舶、飞机的尾气、工业企业生产排放、居民生活和取暖、垃圾焚烧等。空气污染的污染物主要包括，烟尘、总悬浮颗粒物、可吸入颗粒物（PM2.5）、细颗粒物（PM10）、二氧化氮、二氧化硫、一氧化碳、臭氧、挥发性有机化合物等等。而人们常常谈论的雾霾，就是属于可吸入颗粒物，空气质量的重要指标之一。

近年来，雾霾天气在我国频繁出现，空气质量问题已引起全社会高度关注。我国也认识到这一问题的严重性，开始在治理空气污染方面加大投入。党的十九大报告指出，着力解决突出环境问题，包括“持续实施大气污染防治行动，打赢蓝天保卫战”等。针对于此，国内众多学者用不同的方法对我国城市空气质量进行了分析：杨新兴等[1]对大气颗粒物 PM2.5 及其危害进行了讨论；白洋等[2]对“雾霾”成因的深层法律思考及防治对策进行了分析；王红磊等[3]对武汉市三类不同大气污染过程下大气污染物特征及潜在源区进行了分析；王丽华对多元线性回归模型进行了实例分析[4]李晓童等[5]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；李丹[6]基于聚类分析和多元回归分析的思想对空气质量问题进行了研究，建立了回归模型；肖正等[7]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。基于上述研究本文对邯郸市空气质量数据进行了分析。将 2013 年 12 月至 2018 年 4 月的邯郸空气质量数据分成两部分，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 2017 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。最后，根据得到的模型对邯郸未来五个月的空气质量进行预测。

二、研究方法简介

（一）多元线性回归模型的简介

本文通过回归分析来研究各变量间的函数关系[8-10]。设随机变量Y与响应变量$X_1, X_2, \ldots, X_p$的线性回归模型一般形式为：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \tag{1}$$

式中，$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$式$p+1$个未知参数，$\beta_0$称为回归常数，$\beta_1, \beta_2, \ldots, \beta_p$称为回归系数。$Y$被称为解释变量（因变量），$X_1, X_2, \ldots, X_p$是$p$个可以精确测量并控制的一般变量，称为解释变量（自变量）。$p=1$时，模型（1）为一元线性回归模型；$p \geq 2$时，称（1）式为多元线性回归模型。ε是随机误差，常假定

$$\begin{cases} E(\varepsilon) = 0 \\ Var(\varepsilon) = \sigma^2 \end{cases}$$

称

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

为理论回归方程。用矩阵表示式（1）为：

$$Y = X\beta + \varepsilon \tag{2}$$

（二）回归系数的最小二乘估计

对（2）式表示的回归模型，寻找参数$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$的估计值$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$使离差平方和$Q(\beta_0, \beta_1, \beta_2, \ldots, \beta_p) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip})^2$达到极小，即寻找$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$满足：

$$\begin{aligned} Q(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p) &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_p x_{ip})^2 \\ &= \min_{\beta_0, \beta_1, \beta_2, \ldots, \beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip})^2 \end{aligned} \tag{3}$$

依照式（3）求出的$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_p$就称为回归参数$\beta_0,\beta_1,\beta_2,\ldots,\beta_p$的最小二乘估计[10]。

三、数据预处理

本文所用到的关于邯郸市的空气质量数据及其各项指标数据均来自于《中中国空气质量在线监测分析平台》。

数据标准化的处理方法有$Z-Score$值标准化、最小最大值标准化、归一化标准化方法等。本文处理方法所采用的的是$Z-Score$标准化方法，其表达式如下：

$$X_i^{'}=\frac{X_i-\bar{X}^{'}}{\delta_i} \qquad Y^{'}=\frac{Y-\bar{Y}}{\sigma_Y} \tag{4}$$

其中$\bar{X}_i,\bar{Y}$分别代表其均值,δ_i,σ_Y则代表其对应的方差。

四、建立模型

基于以上研究，本文利用多元回归分析对邯郸市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。

（一）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，因变量与各自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^2=0.9701 是一个很大的值，F 检验的 $p<2.2\times10^{-16}$

是一个很小的值，这说明模型可能存在过拟合的情况。

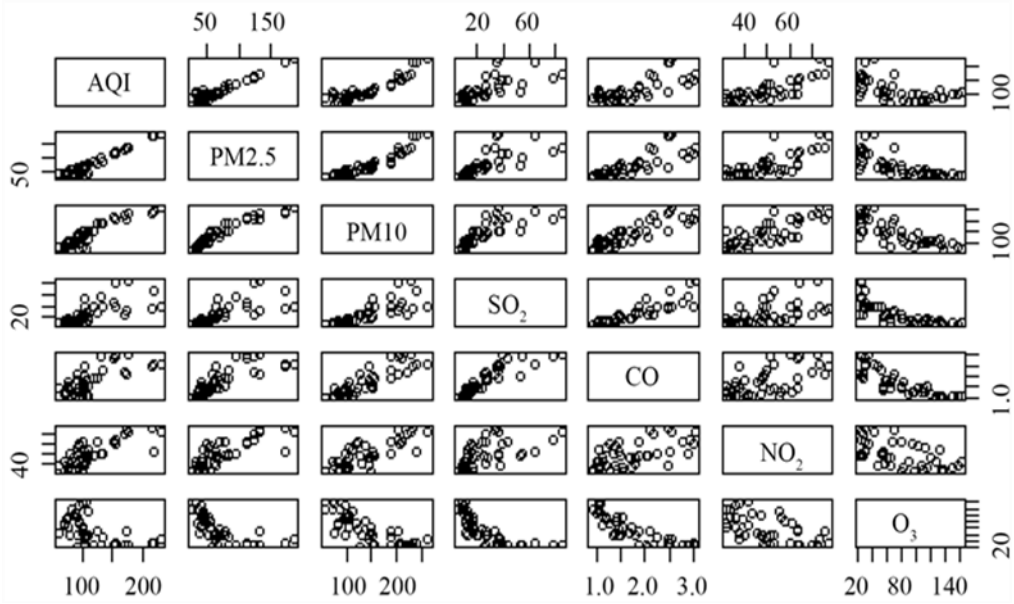

图 1　散点图

（3）回归诊断

表 1　模型拟合结果

Coefficients:	Estimate Std. Error t value Pr(>|t|)
(Intercept)	0.2245
PM2.5	2.67e-12***
PM10	0.0074 **
SO_2	0.3869
CO	0.7351
NO_2	0.9227
O_3	6.59e−08 ***
$R^2 = 0.9744$	
调整 $R^2 = 0.9701$	
$p < 2.2\times10^{-16}$	

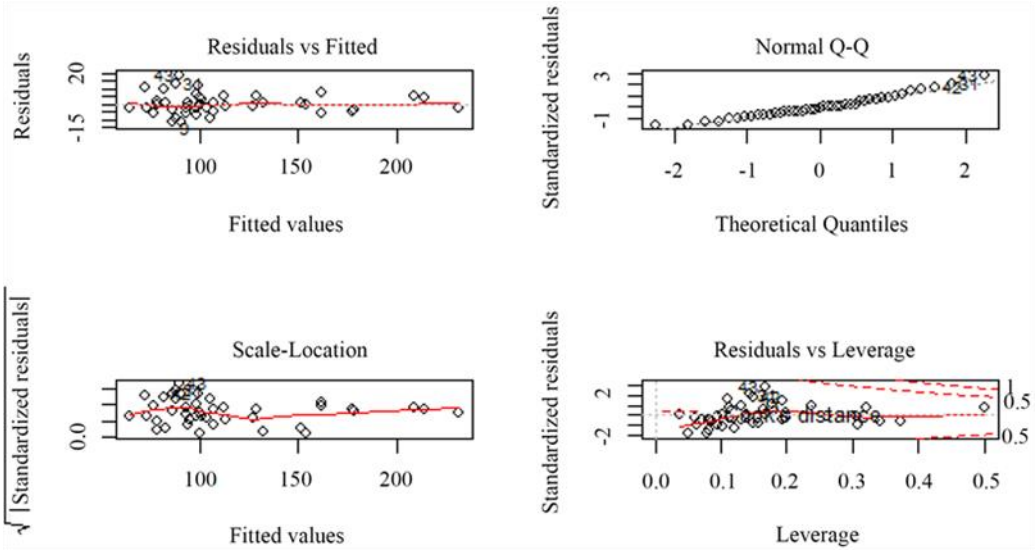

图 2　回归诊断图

估计结果中 NO_2，SO_2 和 CO 的系数估计结果不显著，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 NO_2，SO_2 和 CO 变量后所建模型中变量的系数均显著，并且 R^2 没有减小且 F 检验的 p-value 依旧显著。进而绘制这个模型的标准化残差诊断图，来判断模型的拟合情况。

如图 3 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。因此可以判定，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的模型。

表 2　模型拟合结果

Coefficients:	Estimate Std. Error t value Pr(>|t|)
(Intercept)	0.255561
PM2.5	2.66e-13 ***
PM10	0.000677 ***
O_3	4.79e−09 ***
$R^2 = 0.973$	
调整 $R^2 = 0.9709$	
$p < 2.2\times10^{-16}$	

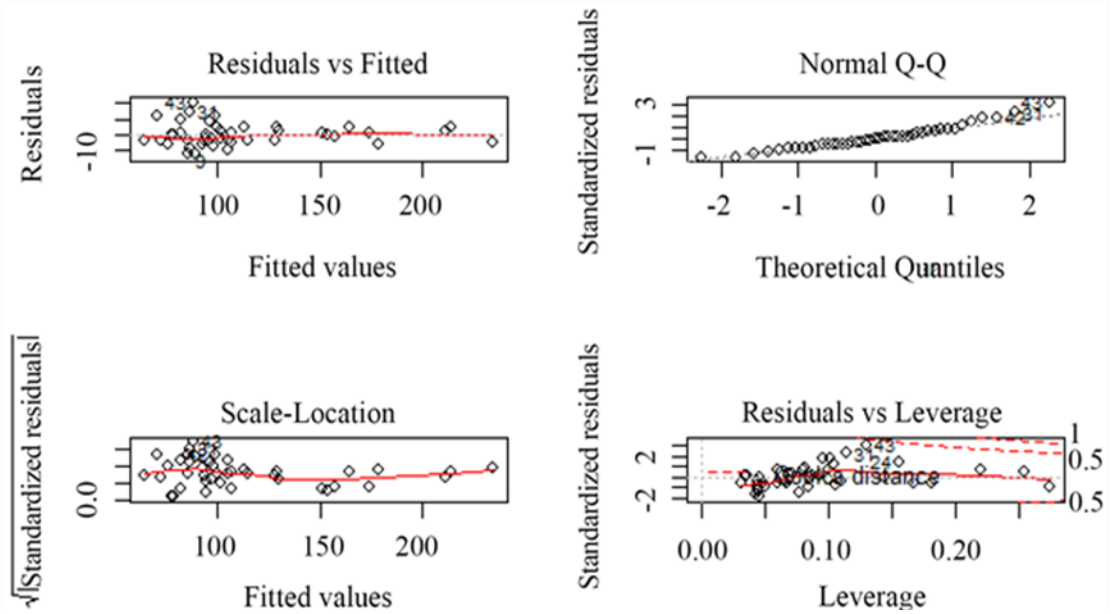

图 3　回归诊断图

（二）修正的多元线性回归模型

首先，我们从验证变量为正态分布的假设入手，得到各变量箱线图和 QQ 图。从图 4-10 中观察可得各变量数据存在偏态的情况，所以考虑对存在偏态的预测变量和响应应变量同时进行 BOX-COX 变换，得到$tAQI, tSO_2, tCO, tPM2.5$。

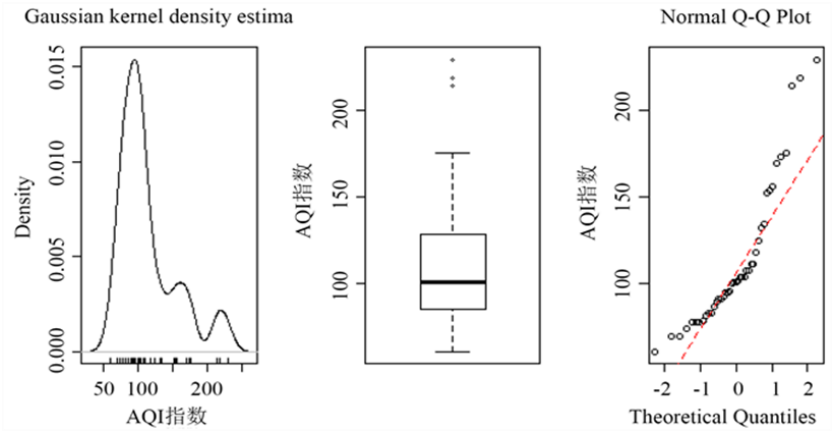

图 4　AQI 正态分布检验

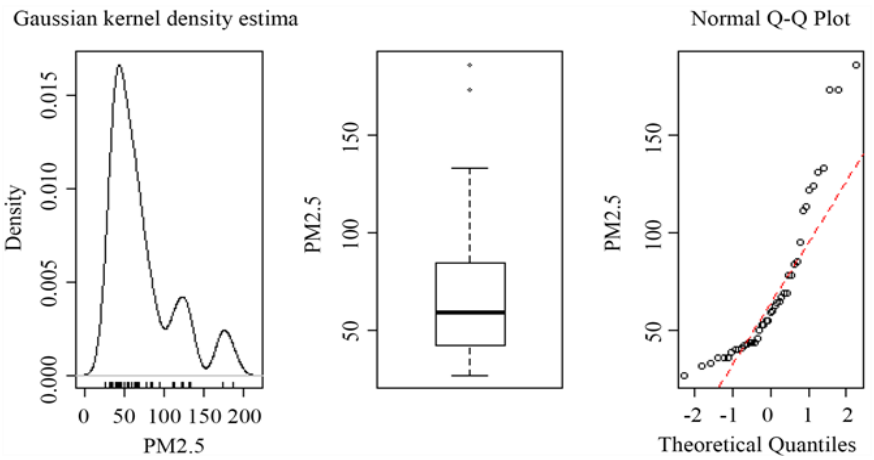

图 5　PM2.5 正态分布检验

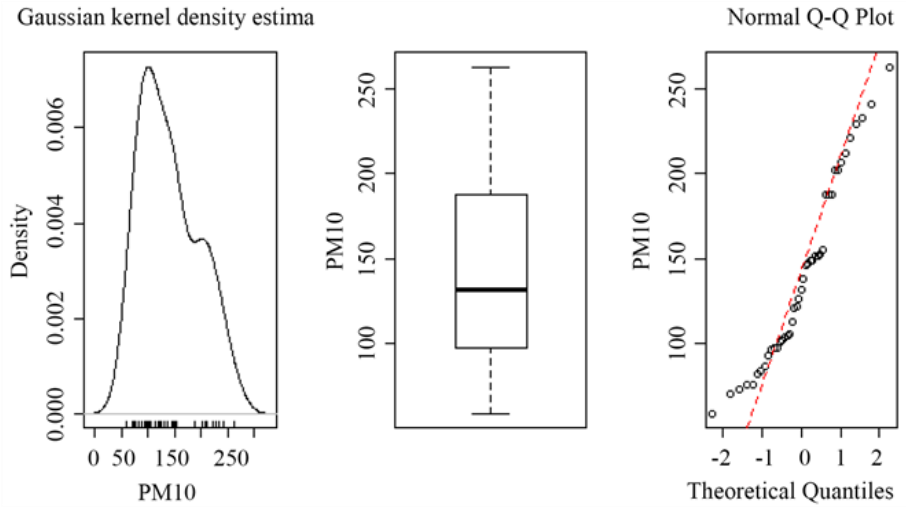

图 6　PM10 正态分布检验

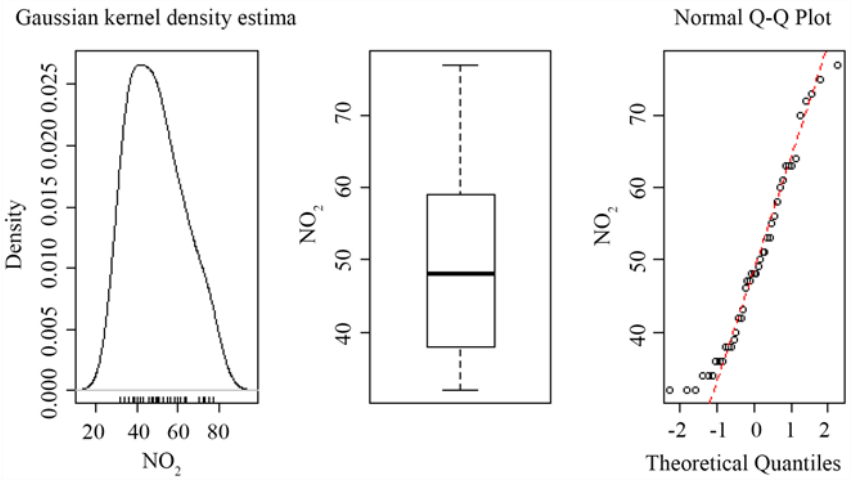

图 7　NO_2 正态分布检验

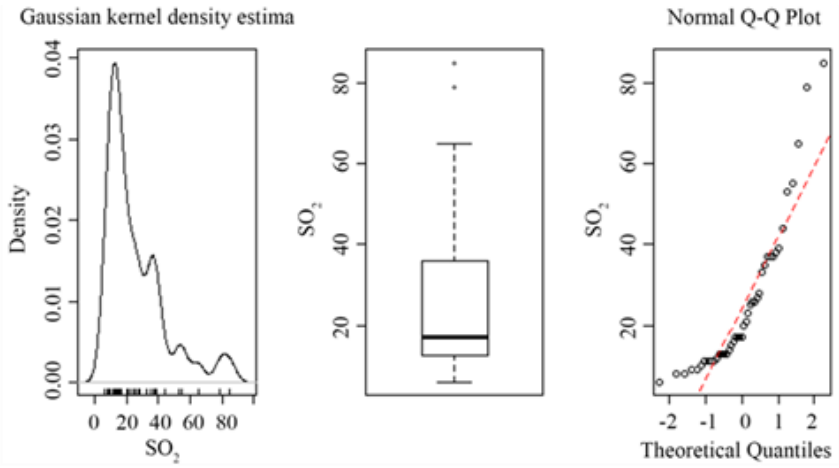

图 8　SO_2 正态分布检验

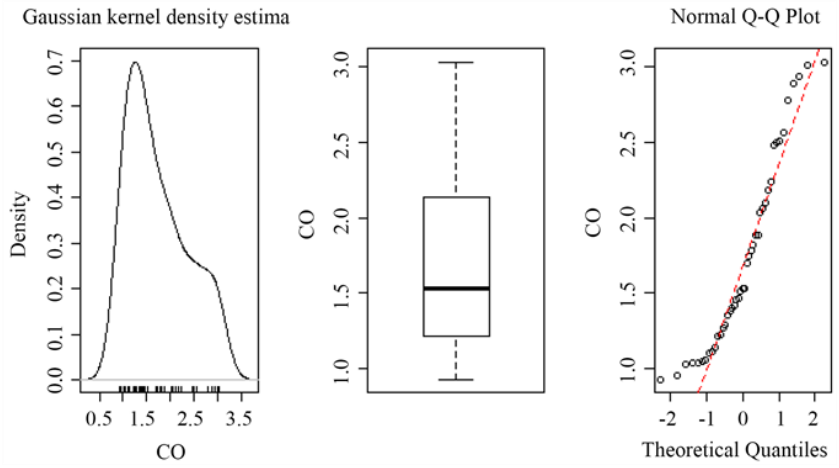

图 9　CO 正态分布检验

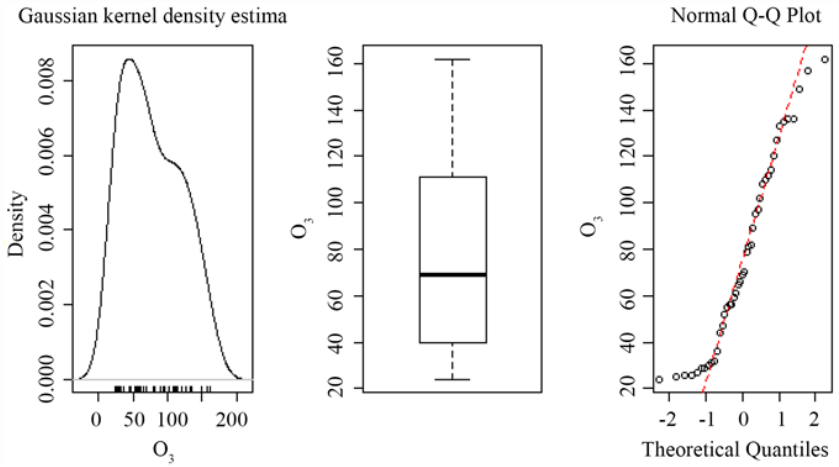

图 10　O_3 正态分布检验

从表 3 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，因而进行多重共线性的判别，结果如表 4 所示。

表 3　模型拟合结果

Coefficients:	Estimate Std. Error t value Pr(>|t|)
(Intercept)	3.99e−05 ***
PM2.5	0.01660 *
PM10	0.00165 **
NO_2	0.94995
tSO_2	0.75565
tCO	0.49459
O_3	0.00102 **
	$R^2 = 0.9318$
	调整 $R^2 = 0.9205$
	$p < 2.2\times10^{-16}$

由表 4 可以看出，方差膨胀因子存在大 10 的值，因此数据存在多重共线性，利用向后剔除法进行变量选择。

表 4　多重共线性诊断

变量	PM10	NO_2	SO_2	CO	O_3
VIF	24.069357	3.452873	10.288414	14.752069	6.674461

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。

表 5　模型拟合结果

Coefficients:	Estimate Std. Error t value Pr(>|t|)
(Intercept)	4.22e−06 ***
PM2.5	0.00354 **
PM10	0.00104 **
O_3	1.21e−07 ***
	$R^2 = 0.9308$
	调整正 $R^2 = 0.9225$
	$p < 2.2\times10^{-16}$

绘制上述模型的诊断图（图 11），由图 11 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。该模型为有效模型。

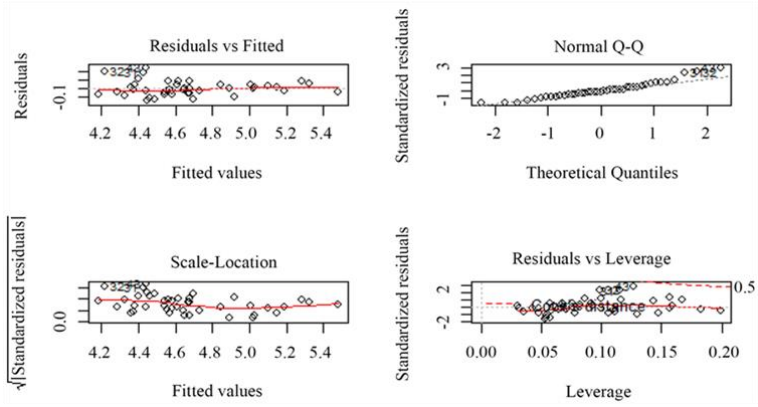

图 11　回归诊断图

综上，最终预测模型为：

$$AQI = PM2.5^{0.399506} \times 10^{0.003997PM10+0.003511O_3+2.187334} \quad (5)$$

五、预测

得到的预测结果如表 6 所示 ，预测结果显示因变量 AQI 的各个真实值全在相对应的预测区间内，这表明所选模型的预测效果很好，由此进一步证明了模型的有效性。

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	4.886	是
4.828314	4.839	4.643	5.035	是
4.875197	5.001	4.81	5.192	是
4.624973	4.624973	4.523	4.911	是

六、结论

本文基于国内众多学者对我国空气质量指数的分析方法，对邯郸市的空气质量数据，首先建立一般多元线性回归模型，但通过模型回归诊断发现一般多元线性回归模型的拟合效果不好，不能很好的反映因变量与各自变量之间的关系。然后通过排除显著性检验结果，排除没有通过检验的自变量；以及对变量进行 BOX-COX 变换和变量选择，在原模型上进行了修正，建立了变换后的模型。通过对变换后的模型回归诊断发现，该模型的诊断拟合比较好，因而将变换后的模型确定为最终预测模型，并用其进行模型的预测。最终，将模型预测的因变量值与因变量真实值进行比较，发现预测效果很好，进一步证明了变换后的模型的有效性。

参考文献

[1] 杨新兴,冯丽华,尉鹏.大气颗粒物 PM2.5 及其危害[J].前沿科学,2012,6(01): 22-31.

[2] 白洋,刘晓源.“雾霾”成因的深层法律思考及防治对策[J].中国地质大学学报(社会科学版),2013,13(06): 27-33.

[3] 王红磊,沈利娟,施双双等.武汉市三类不同大气污染过程下大气污染物特征及潜在源区分析[J].三峡生态环境监测,2019,4(02): 27-39.

[4] 王华丽.多元线性回归分析实例[J].科技资讯,2014,29(001):1672-3791.

[5] 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.

[6] 李丹.基于聚类分析和多元回归的空气质量的分析[D].天津:《南开大学博硕士学位论文全文数据库》,2015.

[7] 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.

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

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

[10] 何晓群, 刘文卿.应用回归分析[M].北京: 中国人民大学出版社,2015.

附录：本案例所使用的 R 软件程序命令(部分)

```
#导入数据
z<-read.table("C:\Users\Administrator\Desktop\datas.txt",header=TURE)
#对各个自变量与因变量间的线性相关性显著性检验
cor.test(Y=AQI,X1=PM2.5,data=z)
cor.test(Y=AQI,X2=PM10,data=z)
cor.test(Y=AQI,X3=SO2,data=z)
cor.test(Y=AQI,X4=CO,data=z)
cor.test(Y=AQI,X5=NO2,data=z)
#通过各自变量间的相关系数大小检验各自变量间是否存在多重共线性
cor(X1=PM2.5,X2=PM10,use="everything",method="pearson",data=z)
cor(X1=PM2.5,X3=SO2,use="everything",method="pearson",data=z)
cor(X1=PM2.5,X4=CO,use="everything",method="pearson",data=z)
cor(X1=PM2.5,X5=NO2,use="everything",method="pearson",data=z)
cor(X2=PM10,X3=SO2,use="everything",method="pearson",data=z)
cor(X2=PM10,X4=CO,use="everything",method="pearson",data=z)
cor(X2=PM10,X5=NO2,use="everything",method="pearson",data=z)
cor(X3=SO2,X4=CO,use="everything",method="pearson",data=z)
cor(X3=SO2,X5=NO2,use="everything",method="pearson",data=z)
cor(X4=CO,X5=NO2,use="everything",method="pearson",data=z)
#将相关系数矩阵可视化
states<-z[,2:6]
chart.Correlation(states,method="pearson")
#将数据标准化
scale(data=z[,2:6],center=T,scale=T)
#对相关数据拟合多元线性回归方程
lma<-lm(AQI~PM2.5+PM10+SO2+CO+NO2,data=z)
#查看回归模型的各种详细内容，如回归系数及其显著性等
summary(lam)
```