基于回归模型的
城市 PM2.5 浓度预测问题

2022 年统计基础课程期末报告

学号：2022103795

姓名：吕明倩

日期：2022 年 1 月 2 日

目　录

第一章　背景介绍

1.1　问题陈述

伴随着现代城市化的飞速发展，雾霾逐渐由局部地区的环境问题升级为全国范围内的环境灾害，而 $PM_{2.5}$ 则被认为是带来大规模雾霾天气的“元凶”。$PM_{2.5}$ 来源广泛且成因复杂，其浓度水平既与其他空气污染物有关，又受到气象因素的影响，本文聚焦影响 PM2.5 浓度的外部因素，通过搭建回归模型实现对 PM2.5 浓度的精准预测，能够为制定 $PM_{2.5}$ 浓度的治理举措提供可靠的数据支撑。

1.2　数据简介

本文选择西安市 2020 年 12 月 1 日至 2021 年 12 月 31 日每日的空气污染物和气象因素的数据信息作为研究对象，以 2021 年 12 月 1 日为时间节点，将前一年（2020 年 12 月 1 日至 2021 年 11 月 30 日）的数据信息作为训练集进行模型拟合，之后的数据用作测试集。并将训练集按照每三个月一组，对数据依次赋予冬季、春季、夏季、秋季的季节标签。以 $PM_{2.5}$ 浓度为响应变量，其他空气污染物与气象因素的解释变量名称、符号和单位信息汇总如表 1 所示：

空气污染物			气象因素		
名称	符号	单位	名称	符号	单位
可吸入颗粒物	PM10	$\mu g/m^3$	最高气温	MAXT	℃
二氧化硫	SO2	$\mu g/m^3$	平均气温	AT	℃
二氧化氮	NO2	$\mu g/m^3$	最低气温	MINT	℃
一氧化碳	CO	mg/m^3	平均相对湿度	AH	%
臭氧	O3	$\mu g/m^3$	最高气压	MAXP	hPa
			平均气压	AP	hPa
			最低气压	MINP	hPa
			最大地面风速	MAXW	m/s
			总降水量	RAIN	mm/h
			最大地表辐射	MAXR	W/m^2

表 1：影响 $PM_{2.5}$ 的空气污染物和气象因素

1.3　描述性统计

首先根据中国空气质量等级依据 $PM_{2.5}$ 日均浓度给出的划分标准，对西安市每日的空气质量等级进行判定，$PM_{2.5}$ 污染等级标准以及各等级所占天数、比例如表 2 所示：统计结果显示，在 2020 年底至 2021 年底的一年时间内，西安市仅有 56.4% 的天数空气质量为优，有 50 天出现不同程度的环境污染。

空气质量标准		西安市	
污染等级	PM2.5 标准值 ($\mu g/m^3$)	天数	占比
优	0~35	206	56.4%
良	35~75	109	29.9%
轻度污染	75~115	28	7.7%
中度污染	115~150	12	3.3%
重度污染	150~250	10	2.7%

表 2：空气质量等级统计

其次通过散点图的分布来初步判断 15 个解释变量与 PM2.5 浓度之间的相关性，图 3 给出了与 $PM_{2.5}$ 浓度分别正相关、负相关以及不相关的三个代表性解释变量，完整散点图详见附录。通过散点图可以发现，PM_{10}、SO_2、NO_2、CO 等空气污染物浓度以及最大地面风速与 $PM_{2.5}$ 浓度存在明显的正相关关系，O_3 浓度、平均相对湿度以及三个气温因素均与 $PM_{2.5}$ 浓度存在轻微的负相关性，而总降水量、最大地表辐射以及三个气压因素与 $PM_{2.5}$ 浓度的相关性并不明显。

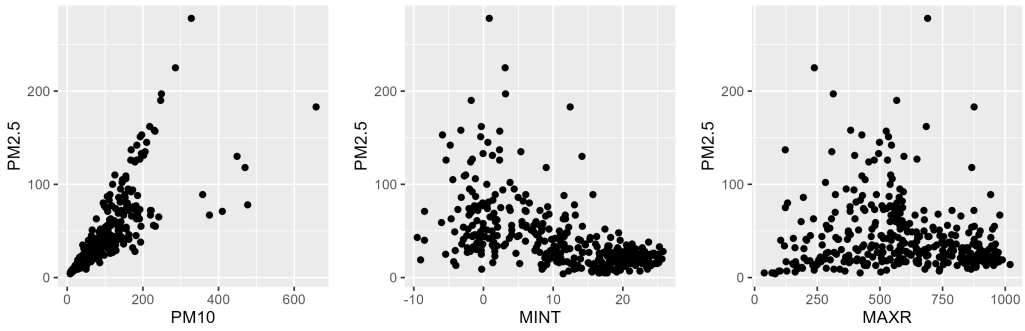


图 3：部分关系散点图

此外考虑到本数据集中各解释变量之间的相关性，不难发现三个气温数据高度正相关，三个气压数据高度正相关，同时气温和气压之间也存在相当大的负相关性，因此在建模过程中需要关注解释变量之间的多重共线性，以提升模型的稳定性。

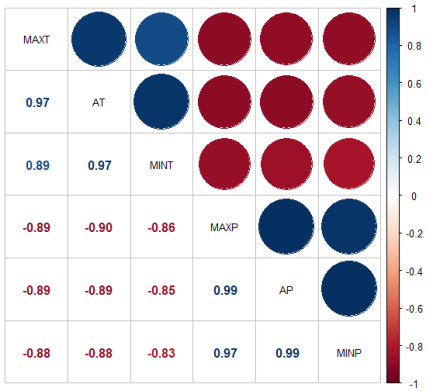


图 4：气温和气压相关性检验图

第二章　模型搭建

2.1　多元线性回归模型

首先将全部解释变量引入模型，构建全变量多元线性回归模型。

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	-2.85E+02	1.66E+02	-1.717	0.08678	.
PM10	2.21E-01	1.07E-02	20.59	<2E-6	***
SO2	1.44E+00	4.42E-01	3.253	0.00126	**
NO2	-1.16E-01	7.41E-02	-1.56	0.11976	
CO	7.49E+01	3.84E+00	19.512	<2E-6	***
O3	1.91E-01	4.41E-02	4.334	1.91E-05	***
MAXT	9.38E-02	6.49E-01	0.145	0.88513	
AT	5.74E-01	1.23E+00	0.468	0.64003	
MINT	-6.46E-01	7.06E-01	-0.916	0.36042	
AH	5.26E+00	7.25E+00	0.725	0.46919	
MAXP	-7.50E-01	6.73E-01	-1.115	0.26545	
AP	1.05E+00	1.34E+00	0.781	0.43556	
MINP	-5.22E-02	7.92E-01	-0.066	0.94742	
MAXW	8.39E-01	7.09E-01	1.183	0.23747	
RAIN	-3.38E-01	1.11E-01	-3.037	0.00257	**
MAXR	-9.19E-03	5.57E-03	-1.649	0.09997	.
Adjusted R2:	0.9142		R2:	0.9177	
F-statistic:	259.5		p-value:	<2.2E-16	

表 5：多元线性回归模型结果

根据 F 检验结果显示，多元线性回归模型可以通过显著性检验，各偏回归系数都不全为 0，根据 t 检验结果显示，在 95% 的置信水平下，解释变量 PM_{10}、SO_2、CO、O_3、$RAIN$ 是显著的，此外模型的 R 方和调整 R 方分别为 0.9142 和 0.9177，这说明全变量的多元线性回归模型对 $PM_{2.5}$ 浓度具备良好的解释能力。

在具有全变量的多元线性模型的基础上用逐步回归法进行变量选择，依次从全模型中删除解释变量 $MINP$、$MAXT$、AH、$MAXW$、$MAXP$、AT、$MINT$，得到逐步回归多元线性模型（记为 $Model1$），通过逐步回归使模型的调整 R 方获得提升，模型的拟合效果得到改善。

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	-2.12E+02	1.07E+02	-1.974	0.049143	*
PM10	2.19E-01	9.76E-03	22.392	<2E-6	***
SO2	1.36E+00	3.69E-01	3.696	0.000254	***
NO2	-1.09E-01	6.74E-02	-1.61	0.108228	
CO	7.56E+01	3.24E+00	23.321	<2E-6	***
O3	1.88E-01	3.96E-02	4.748	2.98E-06	***
AP	1.75E-01	1.10E-01	1.583	0.114301	
RAIN	-3.28E-01	1.09E-01	-2.997	0.002917	**
MAXR	-6.47E-03	3.88E-03	-1.665	0.096823	.
Adjusted R2:	0.9148		R2:	0.9166	
F-statistic:	489.4		p-value:	<2.2E-16	

表 6：逐步回归多元线性回归模型结果

2.2 模型诊断

多元线性回归模型的构建是需要一定的假设前提的，如果这些前提条件得不到满足，在一定程度上会影响模型的有效性、稳定性和准确性，接下来将利用正态性检验、多重共线性检验以及方差齐性检验三种检验方法对逐步回归多元线性回归模型进行诊断。

一、正态性检验

分别绘制直方图、PP 图和 QQ 图（图 6），通过可视化的方法发现响应变量 $PM_{2.5}$ 浓度并不服从正态分布假设，考虑采用 $BOX\text{-}COX$ 变换，由于转化参数的估计值为 0.48，参考 $COX\text{-}BOX$ 变换表，对响应变量进行开平方运算，得到变换后的逐步回归多元线性回归模型（记为 $Model2$）。

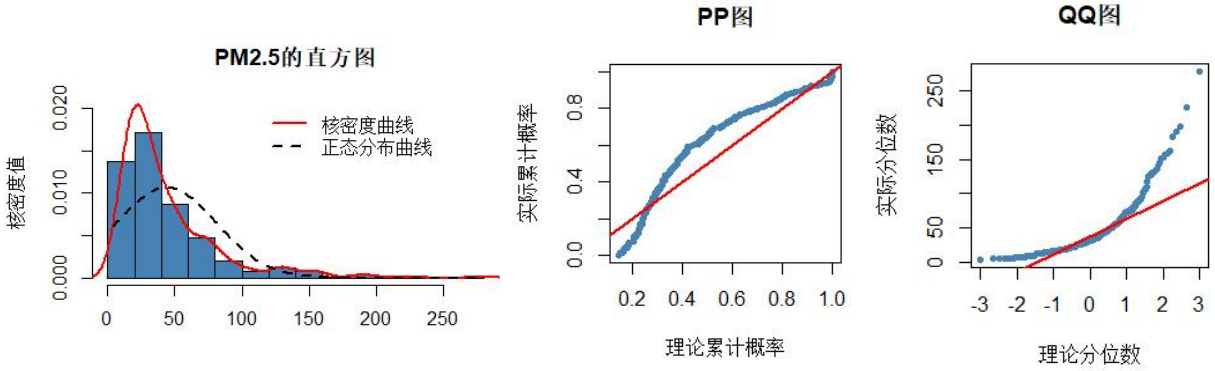


图 7：正态性检验图

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	0.8554877	0.2391446	3.577	0.000395	***
PM10	0.0132964	0.0005949	22.35	<2E-16	***
SO2	0.0480948	0.0236715	2.032	0.042922	*
NO2	0.0166128	0.0040074	4.146	4.24E-05	***
CO	3.9529	0.1939436	20.382	<2E-16	***
O3	0.0091899	0.0024472	3.755	0.000202	***
MINT	-0.01615	0.0078425	-2.059	0.040197	*
MAXW	0.0701553	0.0381125	1.841	0.066494	.
RAIN	-0.0366629	0.0063951	-5.733	2.11E-08	***
MAXR	-0.0005539	0.0002275	-2.434	0.015414	*
Adjusted R2:	0.9285		R2:	0.9303	
F-statistic:	526.1		p-value:	<2.2E-16	

表 8：BOX-COX 变换后的逐步回归多元线性回归模型结果

模型结果如表 7 所示，$Model2$ 仍然可以通过 F 检验，但与 $Model1$ 相比，显著变量增加了 NO_2、$MINT$ 和 $MAXR$，模型的调整方差也由 0.9148 提高至 0.9285。

二、多重共线性检验

对 $Model2$ 进行多重共线性检验，根据方差膨胀因子判断解释变量间是否存在多重共线性，由于 VIF 值均落在区间 0 到 10 之间，因此 $Model2$ 可以通过多重共线性检验。

变量	PM10	SO2	NO2	CO	O3	MINT	MAXW	RAIN	MAXR
VIF	1.806147	5.707714	4.219109	3.137132	4.138361	4.479404	1.174675	1.90997	2.619628

表 9：多重共线性检验结果 (BOX-COX 变换后的逐步回归多元线性回归模型)

三、方差齐性检验

采用 BP 检验法对 $Model2$ 的方差齐性进行定量检验，该检验的 p 值为 2.3512e-06，远小于 0.005，所以 $Model1$ 并不满足方差齐性假设，考虑采用残差绝对值的倒数作为权重，构建加权多元线性回归模型进行改进。

2.3 加权多元线性回归模型

构建以回归残差绝对值的倒数为对角元的加权矩阵，对原数据加权，并利用逐步回归进行变量选择。与 $Model2$ 相比，该模型的调整 R 方显著提高至 0.9987，可以通过显著性检验的变量增加了 $MAXT$、AT、AH、$MAXP$、AP、$MINP$，但 SO_2 反而无法通过显著性检验，并且在逐步回归的过程中将 $Model2$ 中的变量 $RAIN$ 剔除。

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	5.664581	4.060093	1.395	0.16384	
PM10	0.251221	0.001669	150.49	<2E-16	***
SO2	-0.364991	0.262301	-1.391	0.16496	
NO2	-0.184095	0.044095	-4.175	3.77E-05	***
CO	80.328871	2.597463	30.926	<2E-16	***
O3	0.281556	0.025299	11.129	<2E-16	***
MAXT	1.367639	0.317844	4.303	2.19E-05	***
AT	-3.647147	0.549548	-6.637	1.22E-10	***
MINT	1.753116	0.300541	5.833	1.24E-08	***
AH	-22.107305	4.093303	-5.401	1.23E-07	***
MAXP	-0.943269	0.449088	-2.1	0.03641	*
AP	2.188308	0.795548	2.751	0.00626	**
MINP	-1.254727	0.409494	-3.064	0.00235	**
MAXW	-2.131407	0.361585	-5.895	8.84E-09	***
MAXR	-0.012326	0.002522	-4.887	1.56E-06	***
Adjusted R2:	0.9987		R2:	0.9988	
F-statistic:	2.021E+04		p-value:	<2.2E-16	

表 10：加权多元线性回归模型结果

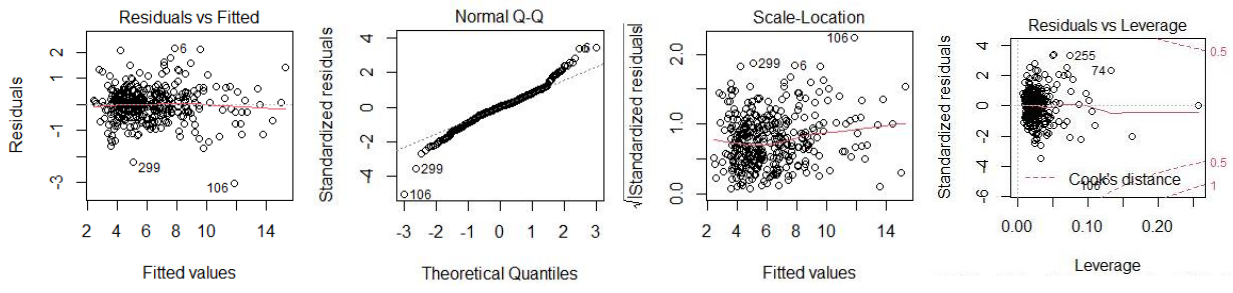


图 11：加权多元线性回归模型诊断图

根据模型诊断结果显示，残差基本服从白噪声假设和正态性假定，但存在 3 个异常点，分别出现在 2020 年 12 月 6 日、2021 年 3 月 16 日和 2021 年 9 月 25 日，COOK 检验结果显示异常点分别出现在 2021 年 2 月 12 日、2021 年 3 月 16 日和 2021 年 8 月 12 日，其中 2021 年 3 月 16 日是两种异常值检验的公共异常值点。

注意到表 9 中显示，此时模型中同时存在三个气温变量和三个气压变量，根据前文关于解释变量之间的相关性分析，推测模型极有可能存在多重共线性，方差膨胀因子值也证明了这一点。为了防止模型通过最小二乘法得到的偏回归系数失效，这里通过逐步从上述模型中删除 VIF 值最大的变量和不显著的变量，获得不存在显著多重共线性的加权多元线性回归模型（记为 $Model3$），该模型的四个解释变量均可以通过 t 检验，此时模型的调整 R 方为 0.9914，具备解释 $PM_{2.5}$ 浓度变化情况的良好能力。

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	16.114441	9.664993	1.667	0.096324	.
PM10	0.263887	0.002266	116.476	<2E-16	***
NO2	0.250023	0.027342	9.144	<2E-16	***
AH	6.939268	1.851332	3.748	0.000207	***
MAXR	-0.00491	0.001209	-4.06	6.01E-05	***
Adjusted R2:	0.9914		R2:	0.9915	
F-statistic:	1.055E+04		p-value:	<2.2E-16	

表 12：逐步选择的加权多元线性回归模型结果

2.4 正则化回归模型

考虑到全模型存在严重的多重共线性，除了直接根据方差膨胀因子和显著性检验结果对变量进行筛选外，也可以采用在目标函数上添加正则项的方法实现对偏回归系数的缩减，此外如果添加 $l1$ 正则项构造 LASSO 回归模型还可以降低模型的复杂度。但对于本数据集而言，LASSO 回归效果欠佳，所以我们添加 $l2$ 正则项构建岭回归模型。

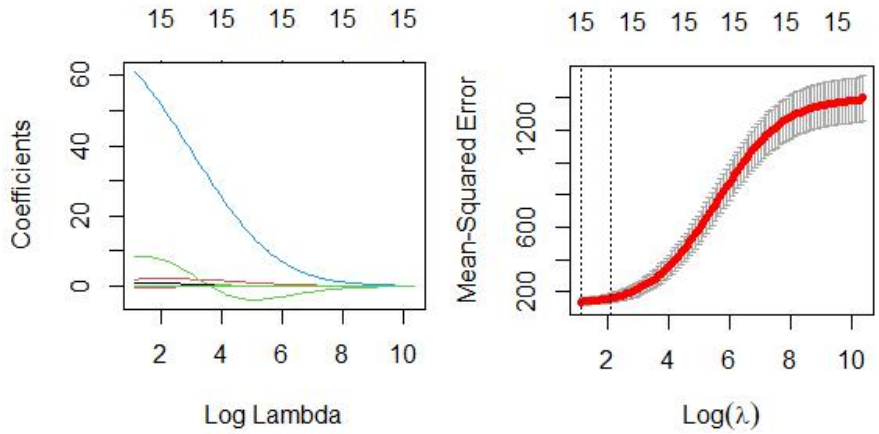


图 13：岭回归系数关系图（左）和交叉验证结果（右）

关于岭回归中参数 λ 的选择问题，首先通过展示不同 λ 值与岭回归系数之间的折线图，发现折线族存在明显的喇叭形，即模型存在明显的多重共线性，而后采用交叉验证法确定模型参数的最优 λ 值为 3.18，以此建立岭回归模型（记为 $Model4$）。该模型的参数个数并没有减少，但对回归系数实现了缩减，回归系数结果如表 14 所示。

	Estimate		Estimate		Estimate		Estimate
(Intercept)	-62.51	CO	61.02	MINT	-0.05383	MINP	0.08270
PM10	0.1936	O3	0.1402	AH	8.285	MAXW	0.8013
SO2	1.955	MAXT	-0.02565	MAXP	-0.08925	RAIN	-0.2852
NO2	0.04205	AT	-0.04341	AP	0.02348	MAXR	-0.004228

表 14：岭回归模型系数结果

2.5 广义可加模型

以上模型均基于解释变量与响应变量之间的确定性关系建立，但考虑到变量之间可能存在非参数关系，所以在多元线性回归模型中引入非参数项，将解释变量推广至样条函数，建立广义可加模型（GAM）。GAM 的一般表达式如下所示，其中 $g(\cdot)$ 为模型的连接函数，对于 $PM_{2.5}$ 浓度的预测问题，选择采用对数连接函数，模型中的 $s_i(\cdot)$ 为解释变量的样条函数，采用约束性最大似然法 (REML) 建立广义可加模型。

$$g(Y) = s_0 + \sum_{i=1}^{p} s_i(X_i)$$

第一步，建立 $PM_{2.5}$ 浓度与各解释变量之间的单因子 GAM，完成多因子 GAM 的初始变量筛选。发现只有变量 $MAXW$ 不显著且对 $PM2.5$ 浓度的解释能力不足 10%。

第二步，利用 14 个显著变量建立多因子 GAM。与一般多元线性回归相似，GAM 的解释变量之间也可能存在高度相关性，因此需要对模型进行共曲线性检验，通过从多因子 GAM 中移除 $MAXT$、AT、$MAXP$、AP、$MINP$，使剩余变量可以通过共曲线性检验。

第三步，调整估计模型。对拟合平滑函数阶数不足 2 次的变量 NO_2、$MINT$、AH、$RAIN$、$MAXR$ 进行线性参数估计，并剔除调整后仍不显著变量 NO_2、$MINT$、$RAIN$、$MAXR$，此时所有变量均可以通过平滑度检验。

将广义可加模型（记为 $Model5$）的结果汇总如下（表 15）：模型显示，调整后多因子 GAM 的调整 R 方可以达到 0.956，其中模型保留了线性变量 AH，对解释变量 PM_{10}、SO_2、CO 和 O_3 构建了样条平滑函数，与 $PM_{2.5}$ 浓度的关系如图 16 所示，PM_{10} 与 CO 均对 $PM_{2.5}$ 浓度存在正向效应，其中在 PM_{10} 浓度低于 $100\mu g/m^3$ 时，对 $PM_{2.5}$ 浓度变化的正向效应显著高于 PM_{10} 浓度超过 $100\mu g/m^3$。

	自由度	参考自由度	F 值	p 值	k_{index}
s(PM10)	6.835	6.985	234.0	<2e-16	1.09
s(SO2)	3.877	4.812	4.591	0.00064	0.99
s(CO)	2.917	3.726	69.96	<2e-16	0.94
s(O3)	3.419	4.287	2.937	0.01842	0.98
	估计值	标准误	t 值	p 值	
常数项	3.292	0.05805	56.71	<2e-6	
AH	0.29533	0.08669	3.407	0.000735	
模型解释度：95.8%			Adjusted R2：0.956		

表 15：广义可加模型结果

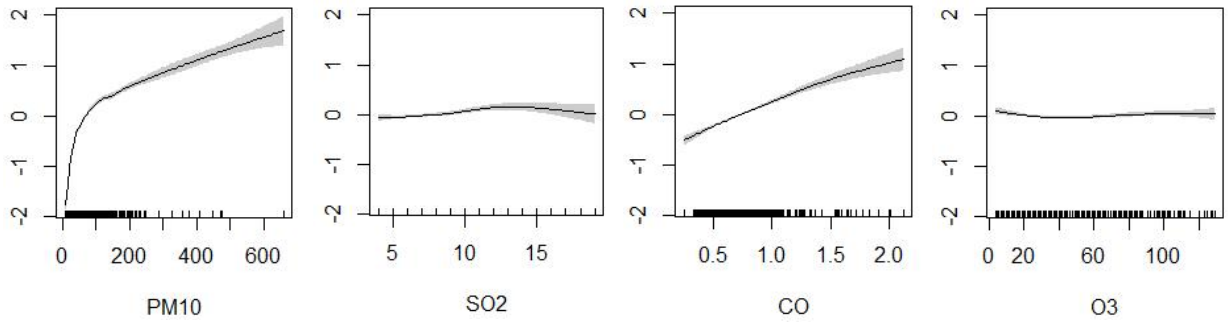


图 16：非线性变量与响应变量之间的关系

第三章　评价总结

3.1　预测结果

依次采用上述五个回归模型对测试集 $PM_{2.5}$ 浓度进行预测，绘制预测结果的二元散点图，其中红点的横纵坐标分别代表预测值和真实值，黑线表示预测值与真实值相等。

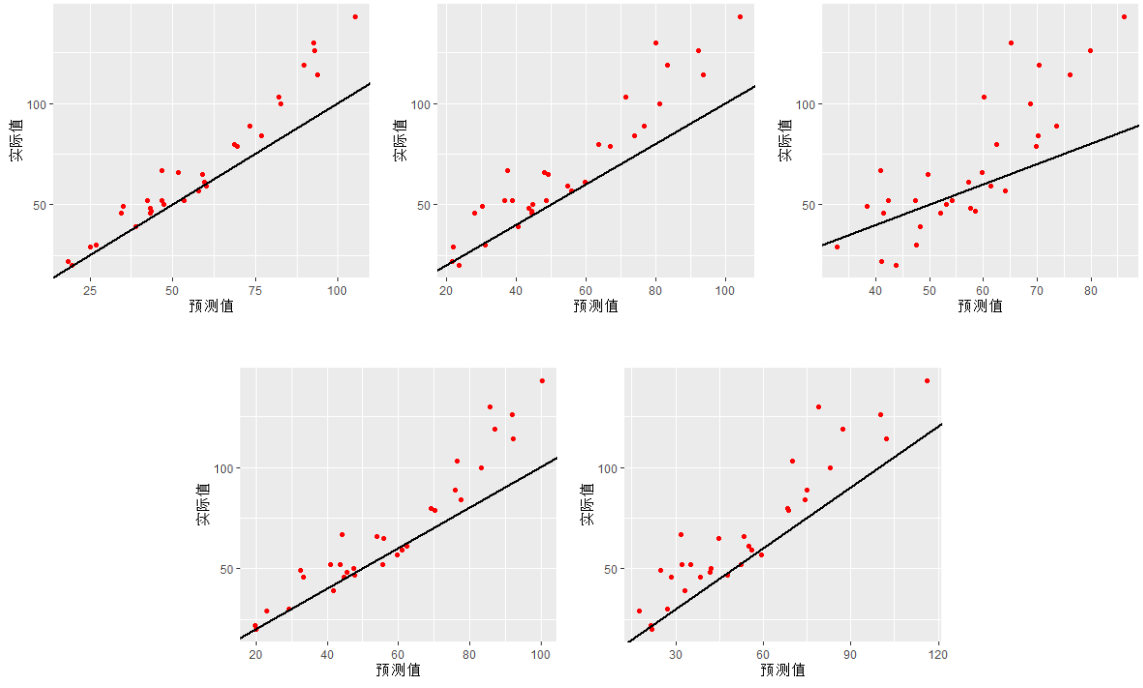


图 17：模型预测结果图

图像显示 $Model1$、$Model2$、$Model4$ 和 $Model5$ 的预测表现效果类似，对低浓度的 $PM_{2.5}$ 预测效果更好，随着 $PM_{2.5}$ 浓度的增加，预测效果逐渐降低，并且预测数值普遍低于真实值。而对于预测效果相对较差的 $Model3$ 而言，当预测较小值时，预测结果偏高，当预测较大值时，预测结果偏低，推测可能存在过拟合现象。

汇总五个回归模型的预测均方误差 RMSE 以及拟合模型的调整 R 方结果如表 18 所示。对于 $Model1$、$Model2$ 和 $Model3$，拟合回归模型的调整 R 方逐渐增大，这意味着模型的拟合效果越来越好，但预测均方误差反而逐渐增大，导致模型在预测新数据时的表现反而变差，佐证了过拟合现象的出现。所以对于本文数据而言，当响应变量 $PM_{2.5}$ 浓度未能通过多重共线性检验时，直接根据方差膨胀因子和显著性进行变量选择可能会使模型出现过拟合现象。

此外注意到，添加正则项建立岭回归模型和添加非线性样条函数建立广义可加模型，也可以达到降低解释变量共线性的目的。$Model4$ 和 $Model5$ 的调整 R 方高于 $Model1$，且均方误差小于 $Model2$ 和 $Model3$，这意味着岭回归和 GAM 在提升模型拟合度的情况下，又避免了预测均方误差的加大，两模型的表现效果更好。最后根据 $Model4$ 的预测均方误差略低于 $Model5$，判断岭回归模型最优。

Model	Method	RMSE	Adjusted R2
1	基础模型	15.71	0.9148
2	添加 BOX-COX 变换	19.31	0.9285
3	用残差绝对值的倒数加权	25.32	0.9914
4	添加正则化函数	17.31	0.9486
5	添加非线性样条函数	18.69	0.9570

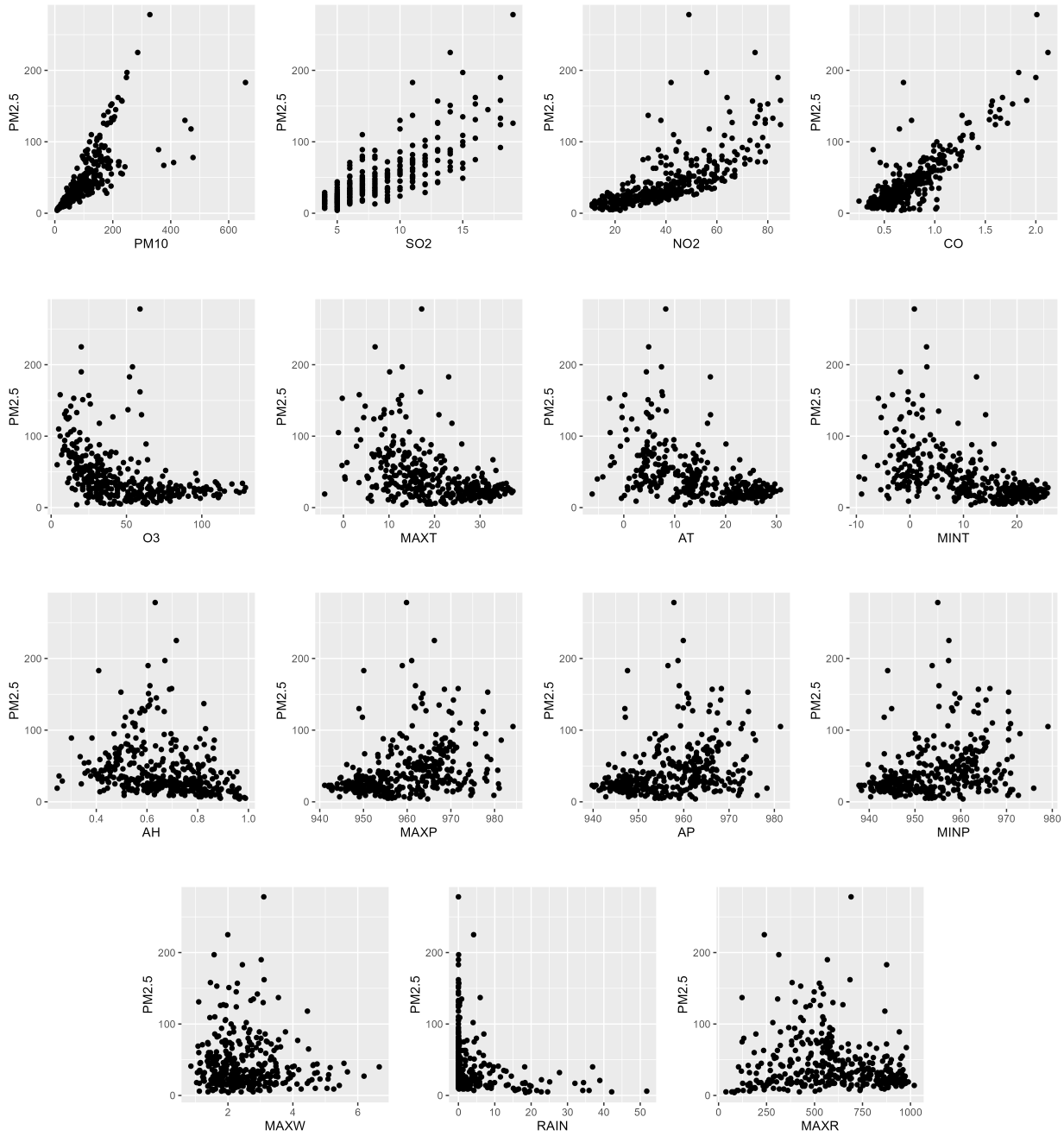
表 18：模型拟合（调整 R 方）和预测（RMSE）结果对比

3.2 总结拓展

本文以预测城市 $PM_{2.5}$ 浓度为出发点，尝试建立适合的回归模型，首先建立了多元线性回归模型，并在此基础上通过 $BOX\text{-}COX$ 变换改善响应变量的正态性，通过构建以残差绝对值的倒数为对角元的加权矩阵，建立可以通过方差齐性检验的加权多元线性回归模型。此外，考虑到模型解释变量之间存在的多重共线性，通过在目标函数中添加 $l2$ 正则项搭建岭回归模型来解决这一问题。最后，在保留连接函数的同时，引入非参数平滑函数构建广义可加模型，更细致地发掘解释变量与响应变量之间的非线性关系。模型的拟合和预测效果达到理想预期。

通过查阅资料可知，对于 $PM_{2.5}$ 浓度的预测问题，除了搭建回归模型外，还可以尝试时间序列分析中 SARIMA 模型的和神经网络中的 BP-ANN 等方法实现预测。

附录



影响因素与 PM2.5 浓度之间的散点图

```r
setwd("D:/研究生/统计基础/期末大作业")
# 导入训练集和测试集数据
data_train <- read.csv("D:/研究生/统计基础/期末大作业/西安pm2.5数据训练集.csv",header=TRUE)
data_test <- read.csv("D:/研究生/统计基础/期末大作业/西安pm2.5数据测试集.csv",header=TRUE)
data_train <- as.data.frame(data_train)
data_test <- as.data.frame(data_test)
# 查看数据结构
str(data_train)
# 查看训练集数据的概览信息
summary(data_train)

# PM2.5与15个解释变量之间的散点图
for (i in 1:15) {
  pg <- ggplot(data = data_train, mapping = aes(x = data_train[,i+2], y = data_train[,2])) +
        geom_point(fill = 'steelblue') + labs(x = names(data_train)[i+2], y = 'PM2.5')
  ggsave(paste0(i, "_", ".png"), pg, width = 3, height = 3)
}

# 多元线性回归模型(Model1)
liner <- lm(PM2.5 ~ PM10 + SO2 + NO2 + CO + O3 + MAXT + AT + MINT + AH + MAXP + AP + MINP + MAXW + RAIN +
              MAXR, data = data_train)
summary(liner)
# 用逐步回归方法进行多元线性回归模型的变量选择
liner_step <- step(liner)
summary(liner_step)

# 模型诊断
# 正态性检验
# 绘制直方图
hist(x = data_train$PM2.5, freq = FALSE, main = 'PM2.5的直方图',
     ylab = '核密度值',xlab = NULL, ylim = c(0,0.02), col = 'steelblue')
lines(density(data_train$PM2.5), col = 'red', lty = 1, lwd = 2)
x <- data_train$PM2.5[order(data_train$PM2.5)]
lines(x, dnorm(x, mean(x), sd(x)),
      col = 'black', lty = 2, lwd = 2.5)
legend('topright',legend = c('核密度曲线','正态分布曲线'),
       col = c('red','black'), lty = c(1,2),
       lwd = c(2,2.5), bty = 'n')

# PP图
real_dist <- ppoints(data_train$PM2.5)
theory_dist <- pnorm(data_train$PM2.5, mean = mean(data_train$PM2.5), sd = sd(data_train$PM2.5))
plot(sort(theory_dist), real_dist, col = 'steelblue',
     pch = 20, main = 'PP图', xlab = '理论累计概率', ylab = '实际累计概率')
abline(a = 0,b = 1, col = 'red', lwd = 2)

# QQ图
qqnorm(data_train$PM2.5, col = 'steelblue', pch = 20, main = 'QQ图', xlab = '理论分位数', ylab = '实际分位数')
qqline(data_train$PM2.5, col = 'red', lwd = 2)

# shapiro检验
shapiro.test(data_train$PM2.5)

# BOX-COX变换：开平方(Model2)
powerTransform(liner_step)
liner_sqrt <- lm(sqrt(PM2.5) ~ PM10 + SO2 + NO2 + CO + O3 + MAXT + AT + MINT + AH + MAXP + AP +
                   MINP + MAXW + RAIN + MAXR, data = data_train)
summary(liner_sqrt)
# 用逐步回归方法进行变量选择
liner_sqrt_step <- step(liner_sqrt)
summary(liner_sqrt_step)
```

```r
# 多重共线性检验：通过
vif(liner_sqrt_step)
plot(liner_sqrt_step)

# 方差齐性检验：不通过
ncvTest(liner_sqrt_step)

# 加权多元线性回归模型(Model3)
W <- diag(1/abs(liner_sqrt_step$residuals))
M <- as.matrix(data_train)
M <- M[,2:17]
M <- apply(M,2,as.numeric)
WM <- W %*% M
data_train_W <- as.data.frame(WM)
liner_weight <- lm(PM2.5 ~ PM10 + SO2 + NO2 + CO + O3 + MAXT + AT + MINT + AH + MAXP + AP + MINP + MAXW +
                     RAIN + MAXR, data = data_train_W)
summary(liner_weight)
# 用逐步回归方法进行变量选择
liner_weight_step <- step(liner_weight)
summary(liner_weight_step)
vif(liner_weight_step)    #存在显著的共线性
# 根据变量显著性和共曲线性进行变量筛选
liner_weight_vif <- lm(PM2.5 ~ PM10 + NO2 + AH + MAXR, data = data_train_W)
summary(liner_weight_vif)
vif(liner_weight_vif)
plot(liner_weight_vif)

# 岭回归(Model4)
# 数据转化为矩阵
M <- as.matrix(data_train)
X <- M[,3:17]
X <- apply(X,2,as.numeric)
y <- M[,2]
# 可视化确定lambda
ridge <- glmnet(X,y,alpha=0)
coef(ridge)
plot(ridge, xvar = 'lambda')
# 交叉验证
cv_ridge<-cv.glmnet(X,as.numeric(y), alpha=0)
cv_ridge$lambda.min
plot(cv_ridge)
# 模型拟合
ridge_coef = coef(object = cv_ridge, s = cv_ridge$lambda.min)
ridge_coef

# 广义可加模型(Model5)
# 单因子广义可加模型
summary(gam(log(PM2.5)~s(PM10,bs="cr"),data=data_train,method="REML"))
summary(gam(log(PM2.5)~s(SO2,bs="cr"),data=data_train,method="REML"))
summary(gam(log(PM2.5)~s(NO2,bs="cr"),data=data_train,method="REML"))
summary(gam(log(PM2.5)~s(CO,bs="cr"),data=data_train,method="REML"))
summary(gam(log(PM2.5)~s(O3,bs="cr"),data=data_train,method="REML"))
summary(gam(log(PM2.5)~s(MAXT,bs="cr"),data=data_train,method="REML"))
summary(gam(log(PM2.5)~s(AT,bs="cr"),data=data_train,method="REML"))
summary(gam(log(PM2.5)~s(MINT,bs="cr"),data=data_train,method="REML"))
summary(gam(log(PM2.5)~s(AH,bs="cr"),data=data_train,method="REML"))
summary(gam(log(PM2.5)~s(MAXP,bs="cr"),data=data_train,method="REML"))
summary(gam(log(PM2.5)~s(AP,bs="cr"),data=data_train,method="REML"))
summary(gam(log(PM2.5)~s(MINP,bs="cr"),data=data_train,method="REML"))
summary(gam(log(PM2.5)~s(MAXW,bs="cr"),data=data_train,method="REML"))
summary(gam(log(PM2.5)~s(RAIN,bs="cr"),data=data_train,method="REML"))
summary(gam(log(PM2.5)~s(MAXR,bs="cr"),data=data_train,method="REML"))
```

```r
# 单因子关系图
plot(gam(log(PM2.5)~s(PM10,bs="cr"),data=data_train,method="REML"),shade=TRUE)
plot(gam(log(PM2.5)~s(SO2,bs="cr"),data=data_train,method="REML"),shade=TRUE)
plot(gam(log(PM2.5)~s(NO2,bs="cr"),data=data_train,method="REML"),shade=TRUE)
plot(gam(log(PM2.5)~s(CO,bs="cr"),data=data_train,method="REML"),shade=TRUE)
plot(gam(log(PM2.5)~s(O3,bs="cr"),data=data_train,method="REML"),shade=TRUE)
plot(gam(log(PM2.5)~s(MAXT,bs="cr"),data=data_train,method="REML"),shade=TRUE)
plot(gam(log(PM2.5)~s(AT,bs="cr"),data=data_train,method="REML"),shade=TRUE)
plot(gam(log(PM2.5)~s(MINT,bs="cr"),data=data_train,method="REML"),shade=TRUE)
plot(gam(log(PM2.5)~s(AH,bs="cr"),data=data_train,method="REML"),shade=TRUE)
plot(gam(log(PM2.5)~s(MAXP,bs="cr"),data=data_train,method="REML"),shade=TRUE)
plot(gam(log(PM2.5)~s(AP,bs="cr"),data=data_train,method="REML"),shade=TRUE)
plot(gam(log(PM2.5)~s(MINP,bs="cr"),data=data_train,method="REML"),shade=TRUE)
plot(gam(log(PM2.5)~s(MAXW,bs="cr"),data=data_train,method="REML"),shade=TRUE)
plot(gam(log(PM2.5)~s(RAIN,bs="cr"),data=data_train,method="REML"),shade=TRUE)
plot(gam(log(PM2.5)~s(MAXR,bs="cr"),data=data_train,method="REML"),shade=TRUE)

# 多因素广义可加模型
gam_model=gam(log(PM2.5)~s(PM10,bs="cr")+s(SO2,bs="cr")+s(NO2,bs="cr")
           +s(CO,bs="cr")+s(O3,bs="cr")+s(MAXT,bs="cr")+s(AT,bs="cr")
           +s(MINT,bs="cr")+s(AH,bs="cr")+s(MAXP,bs="cr")+s(AP,bs="cr")
           +s(MINP,bs="cr")+s(RAIN,bs="cr")+s(MAXR,bs="cr"),data=data_train,method="REML")
summary(gam_model)
concurvity(gam_model,full=FALSE)
# 根据气温和气压的相关性，仅保留单因子方差解释度最高的变量MINT
gam_model=gam(log(PM2.5)~s(PM10,bs="cr")+s(SO2,bs="cr")+s(NO2,bs="cr")
           +s(CO,bs="cr")+s(O3,bs="cr")+s(MINT,bs="cr")+s(AH,bs="cr")
           +s(RAIN,bs="cr")+s(MAXR,bs="cr"),data=data_train,method="REML")
summary(gam_model)
concurvity(gam_model,full=FALSE)
```

```r
# 消除了模型的共曲线性，但需要对变量NO2、MINT、AH、RAIN、MAXR进行参数估计
gam_model=gam(log(PM2.5)~s(PM10,bs="cr")+s(SO2,bs="cr")+NO2+s(CO,bs="cr")+s(O3,bs="cr")
           +MINT+AH+RAIN+MAXR,data=data_train,method="REML")
summary(gam_model)
concurvity(gam_model)
# 剔除不显著变量NO2、MINT、RAIN、MAXR
gam_model=gam(log(PM2.5)~s(PM10,bs="cr")+s(SO2,bs="cr")+s(CO,bs="cr")+s(O3,bs="cr")+AH,
              data=data_train,method="REML")
summary(gam_model)
concurvity(gam_model)
# 确定最终的模型变量：PM10、SO2、CO、O3、AH
plot(gam_model,shade=TRUE)

# 预测结果比较
# 多元线性回归模型
pred1 <- predict(liner_step, newdata = data_test[,c('PM10','SO2','NO2','CO','O3','AP','RAIN','MAXR')])
pred1
ggplot(data = NULL, mapping = aes(pred1, data_test$PM2.5)) +
  geom_point(color = 'red', shape = 19) +
  geom_abline(slope = 1, intercept = 0, size = 1) +
  labs(x = '预测值', y = '实际值')
RMSE1 <- sqrt(mean((data_test$PM2.5-pred1)**2))
RMSE1
# BOX-COX多元线性回归模型
pred2 <- predict(liner_sqrt_step, newdata =
data_test[,c('PM10','SO2','NO2','CO','O3','MINT','MAXW','RAIN','MAXR')])
pred2^2
ggplot(data = NULL, mapping = aes(pred2^2, data_test$PM2.5)) +
  geom_point(color = 'red', shape = 19) +
  geom_abline(slope = 1, intercept = 0, size = 1) +
  labs(x = '预测值', y = '实际值')
RMSE2 <- sqrt(mean((sqrt(data_test$PM2.5)-pred2)**2))
RMSE2
```

```r
# 加权多元线性回归模型
pred3 <- predict(liner_weight_vif, newdata = data_test[,c('PM10','NO2','AH','MAXR')])
pred3
ggplot(data = NULL, mapping = aes(pred3, data_test$PM2.5)) +
  geom_point(color = 'red', shape = 19) +
  geom_abline(slope = 1, intercept = 0, size = 1) +
  labs(x = '预测值', y = '实际值')
RMSE3 <- sqrt(mean((data_test$PM2.5-pred3)**2))
RMSE3
# 岭回归模型
pred4 = predict(object = cv_ridge, s = cv_ridge$lambda.min,
                newx = as.matrix(data_test[,3:17]))
ggplot(data = NULL, mapping = aes(pred4, data_test$PM2.5)) +
  geom_point(color = 'red', shape = 19) +
  geom_abline(slope = 1, intercept = 0, size = 1) +
  labs(x = '预测值', y = '实际值')
RMSE4 <- sqrt(mean((data_test$PM2.5-pred4)**2))
RMSE4
# 广义可加模型
pred5 = predict(gam_model,newdata = data_test[,3:17])
exp(pred5)
ggplot(data = NULL, mapping = aes(exp(pred5), data_test$PM2.5)) +
  geom_point(color = 'red', shape = 19) +
  geom_abline(slope = 1, intercept = 0, size = 1) +
  labs(x = '预测值', y = '实际值')
RMSE5 <- sqrt(mean((log(data_test$PM2.5)-pred5)**2))
RMSE5
```