一、所用额外过程步及统计方法原理阐释

1. 多重填补：proc mi

（1）原理简介

数据集中观测的某个解释变量缺失会使我们在统计分析中无法使用这条观测中其他解释变量所包含的信息，从而降低统计分析的质量。我们考虑使用多重填补法填补数据集中出现的缺失。

Xobs:存在观测值的变量；

Xmis：缺失观测值得变量；

MAR假设：在Xobs为条件的基础上，Xmis的缺失是随机的；

在MAR假设下，我们可以通过条件分布f(Xmis|Xobs)产生相应的填补值。因为数中

存在复杂的数据缺失模式，mi过程自动使用MCMC方法进行缺失数据的填补。

（2）过程步选项

PROC MI;

BY variables;

Class Variables;

EM <options>;

FCS <options>;

FREQ variable;

MCMC <options>;

MONOTONE <options>;

VAR variables;

2. LASSO REGRESSION

（1）原理简介

在传统意义中的最小二乘回归中，我们所得的回归系数是求解这样一个问题：

为了实现回归分析中变量选择的目标，我们使用LASSO REGRESSION，解决的是如下

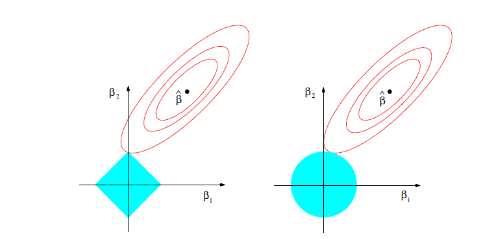
这样一个问题：

我们称为惩罚因子。当时LASSO REGRESSION即为最小二乘回归；而当时，的所有分量都会趋于0。并且LASSO REGRESSION具有一个极为优良的性质，当从0开始增加时，回归系数的各个分量会相继被压缩成0。这为我们进行变量选择提供了极大的方便。

（2）LASSO REGRESSION变量选择算法

|  |  |
| --- | --- |
|  | LASSO REGRESSION FEATURE SELECTION |
|  | STEP1对于已知的Y与X进行最小二乘回归得到回归系数。  STEP2逐步提升直至回归系数的某个分量被压缩为0，并记录此时评价模型优劣的某个指标，如AIC，SBC。  STEP3继续提升直至的又一个分量被压缩为0，如果评价模型优劣的指标相比之前的结果下降了，则重新从2开始执行；否则最优模型为上一次执行STEP2  中产生的模型，算法中止。 |

（3）图解LASSO REGRESSION中为什么回归系数的分量会被压缩为0



3. 响应变量具有多个有序值的logistic regression

（1）原理简介

在响应变量只有两个水平，即0,1的logistic regression中，我们有回归方程：

而当响应变量有多个水平，如k个时，我们采用如下的模型：

首先用p1表示响应变量为最小的类别的概率，有回归方程：

然后用p2表示响应变量为第二小的类别的概率，有回归方程

以此类推直到,即对于m=1,2，…，k-1

由前面k-1个概率决定，即

（2）似然函数的计算

由原理简介中的公式可知

假设各观测X1，X2，…,Xn独立同分布，且响应变量为y1，y2，…,yn,则样本的似然函数：

原模型中共计k+p-2个参数由求上式的极大似然估计可得。

二、问题背景与数据集介绍

1.问题背景

PM2.5是指空气中空气动力学当量直径小于等于2.5微米的颗粒物。它能较长时间悬浮在空气中，难以沉降。PM2.5相比较粗的大气颗粒物，粒径小，面积大，活性强，易吸附有毒有害物质，且容易进入人体呼吸道，因而对人体健康具有不可忽略的危害。PM2.5的浓度是衡量空气质量的最重要的指标。2012年2月29日，国务院要求各地向社会公布PM2.5的检测情况。2013年我国中东部分地区频繁陷入雾霾天气中，PM2.5的数值真正引起了社会各界 的广泛关注。

PM2.5空气质量检测网对PM2.5浓度及污染等级划分如下。

|  |  |
| --- | --- |
| 空气质量等级 | PM2.5浓度 |
| 优 | 0-35 μg/m³ |
| 良 | 35-75 μg/m³ |
| 轻度污染 | 75-115 μg/m³ |
| 中度污染 | 115-150 μg/m³ |
| 重度污染 | 150-250 μg/m³ |
| 严重污染 | >250 μg/m³ |

2. PM2.5来源简介

PM2.5的来源主要分为自然来源以及人为来源。自然来源主要包括火山爆发以及森林火灾，但并不是主要来源。人为来源又分为两类，固定源主要是各种工业过程中的发电，供热对燃煤燃气燃油的燃烧，而流动源主要源于机动车排放的汽车尾气。

3. 数据集介绍

我们所使用的数据集来自UCI Machine Learning Repository:

<http://archive.ics.uci.edu/ml/>。

数据集包含了中国五个大城市从2010年1月1日0时到2015年12月31日每小时监测的PM2.5浓度及相关气象指标。2010年至2012年的大部分PM2.5数据缺失。我们使用上海市在静安，美国大使馆，徐汇区共计三个观测点从2013到2015年的数据进行统计分析。数据集的变量介绍如下。

|  |  |  |
| --- | --- | --- |
| Variable name | Type | Values |
| Year | 分类 | 2013 2014 2015 |
| Month | 分类 | 1,2,…,12 |
| Day | 分类 | 1,2,…,31 |
| Hour | 分类 | 0,1,2,…,23 |
| Season | 分类 | 1,2,3,4 |
| DEWP | 数值 | -17到28之间的整数 |
| HUMI | 数值 | 13到100之间的浮点数 |
| PRES | 数值 | 990到1040之间的浮点数 |
| TEMP | 数值 | -3到41之间的整数 |
| Iws | 数值 | 0到615之间的整数 |
| CBWD | 分类 | cv SE NE SW NW |
| Precipitation | 数值 | 0到62之间的浮点数 |
| Iprec | 数值 | 0到226之间的浮点数 |
| PM\_Jingan | 数值 | 0到700之间的浮点数 |
| PM\_USpost | 数值 | 0到700之间的浮点数 |
| PM\_Xuhui | 数值 | 0到700之间的浮点数 |

DEWP：Dew Point(Celsius Degree) 露点，一般由湿度和气压决定。

HUMI：Humidity(%) 湿度

PRES：Pressure(hpa) 大气压

TEMP：Temperature(Celsius Degree) 温度

IWS：Cumulated Wind Speed(m/s) 累积风速

CBWD：Combined Wind Direction 组合风向

PRECIPITATION：Hourly Precipitation(mm) 每小时降雨量

IPREC：Cumulated Precipitation(mm) 累计降雨

PM： PM2.5 Concentration(μg/m³) PM2.5浓度

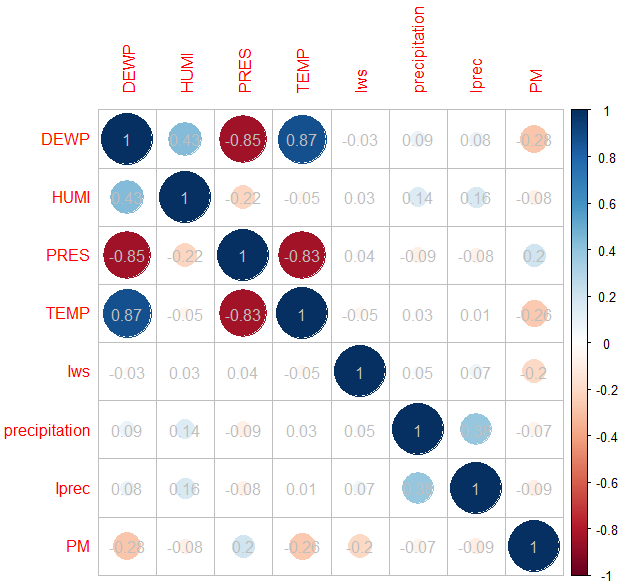
4.变量之间的相关性探索

(1)PM\_Jingan, PM\_USpost PM\_Xuhui分别为静安区，美国大使馆，徐汇区三个观测点观测到的PM2.5浓度，我们列出他们的相关系数。

|  |  |  |  |
| --- | --- | --- | --- |
|  | PM\_Jingan | PM\_Post | PM\_Xuhui |
| PM\_Jingan | 1.00 | 0.96 | 0.96 |
| PM\_Uspost | 0.96 | 1.00 | 0.95 |
| PM\_Xuhui | 0.96 | 0.95 | 1.00 |

三个变量之间存在高度的相关性，所得PM2.5浓度数据结果基本一致，我们考虑使用三个观测点的平均数作为新的变量PM参与剩余的相关分析。

（2）PM与连续解释变量的相关分析



露点在气象学中是指在固定气压之下，空气中所含的[气态](https://baike.baidu.com/item/%E6%B0%94%E6%80%81)[水](https://baike.baidu.com/item/%E6%B0%B4)达到饱和而[凝结](https://baike.baidu.com/item/%E5%87%9D%E7%BB%93) 成液态水所需要降至的温度。在这温度时，凝结的水飘浮在空中称为雾、而沾在固体表面上时则称为露，因而得名露点。据此我们得知，露点由气压以及湿度以及环境温度决定。在相关系数表中，可以看出露点与气压，湿度高度相关。且湿度越高，气压越低，环境温度越高，露点越高。

气压与温度之间的强烈相关关系是物理常识。

累计降雨量即从下雨开始每小时降雨量的累加，故它们之间0.38的相关系数是意料之中的，湿度与降雨量，累计降雨量的轻微相关性也是意料之中的。

在所有的连续变量中，PM2.5与露点，气压，温度，累计风速有0.15-0.3的相关系数。

可能的解释是温度越低，气压越高，露点越低，燃料燃烧更多，从而PM2.5的浓度水平越高。而风速越高，空气的流动性越好，则PM2.5的浓度越小。湿度，降雨量，累计降雨量也与PM2.5的浓度水平有小于0.1的负相关系数，可能的解释是雨水对于PM2.5具有吸附作用。

我们注意到这些连续的解释变量有些变量之间存在高度的相关性，提醒我们设计矩阵的复共线性可能会致使回归分析失去优良的性质。

三、描述分析

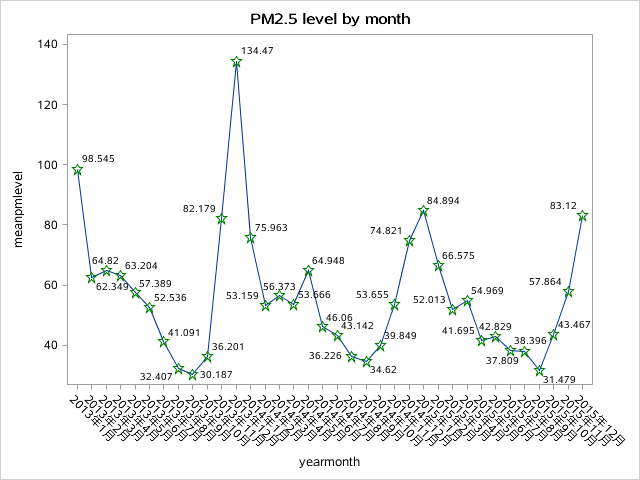
1.上海市PM2.5浓度的季度均值

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| season | observations | PM\_Jingan | PM\_USpost | PM\_Xuhui |
| 1 | 6624 | 57.37 | 54.70 | 57.28 |
| 2 | 6624 | 44.74 | 35.36 | 43.17 |
| 3 | 6552 | 46.40 | 43.11 | 46.84 |
| 4 | 6480 | 82.86 | 80.79 | 83.95 |

A close up of a map

Description generated with high confidence我们看到四个季节中，夏季的PM2.5均值最低，秋季与夏季的差别不大，春季比夏季平均多约15μg/m³,而冬季则明显的比其他三个季节要高，并且平均高了30μg/m³，说明冬季是PM2.5污染的重灾区。冬季北方供暖对燃烧了更多煤等化石燃料可能是这一现象的主要原因。

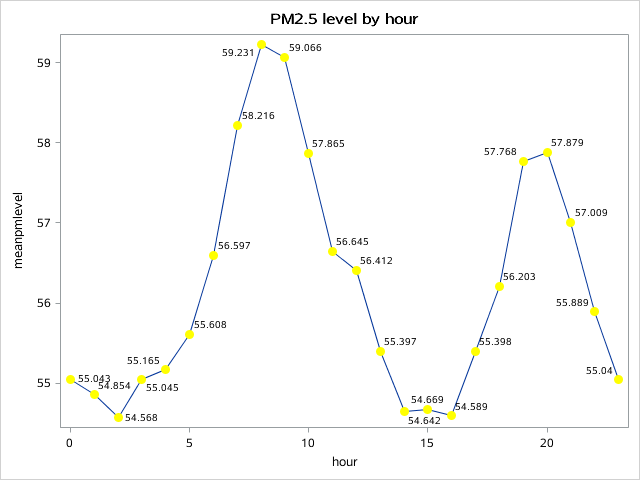
2.上海市PM2.5浓度2013-2015年月度变化趋势

我们计算了上海市2013年到2015年每月的平均浓度，来观察三年来上海市PM2.5浓度的变化以及所在月份对PM2.5的影响。

三年来上海市PM2.5浓度呈轻微的下降趋势，同时注意到2013年12月PM2.5的值异常的高，与当时新闻报道的全国各地频发雾霾天气一致。

3.上海市PM2.5浓度每日变化趋势

我们同时计算了三年来每日的每个小时PM2.5的平均浓度以探索PM2.5的浓度水平在一天的时间跨度内的分布和变化。观察到每天PM2.5浓度的变化幅度不大，极差只有大约5μg/m³, 但能看出较明显的变化趋势,每天的PM2.5浓度从6点以及下午三点开始显著提高，在8点以及20点达到2个峰值。可能的解释是6点以及下午三点开始行驶的机动车排放的尾气造成的PM2.5的累积。



4.上海市PM2.5水平与不同风向的关系

|  |  |  |
| --- | --- | --- |
| cbwd | observations | Mean pm2.5 level |
| . | 4 | 59.11 |
| NE | 9545 | 43.69 |
| NW | 4948 | 82.17 |
| SE | 7585 | 46.42 |
| SW | 3235 | 70.97 |
| cv | 963 | 80.20 |

A screenshot of a video game

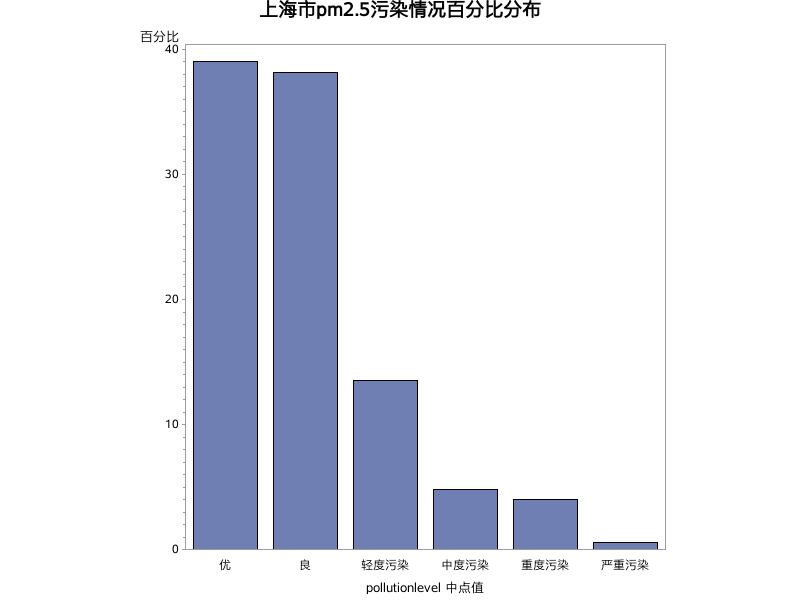
Description generated with high confidence

可以明显的看出刮西南和西北风，或者不刮风时PM2.5的水平非常高，而刮东北和东南风时PM2.5的浓度水平较低。

5.上海市空气质量分析

我们在数据集介绍中给出了PM2.5污染水平的分类标准，统计得2013-2015年上海市在不同污染水平的时间占比。

|  |  |  |  |
| --- | --- | --- | --- |
| Pollutionlevel | Observations | Propotion | Cumulated proportaion |
| 优 | 51275 | 39.02 | 39.02 |
| 良 | 50156 | 38.17 | 77.19 |
| 轻度污染 | 17721 | 13.49 | 90.68 |
| 中度污染 | 6286 | 4.78 | 95.46 |
| 重度污染 | 5243 | 3.99 | 99.45 |
| 严重污染 | 719 | 0.55 | 100.00 |

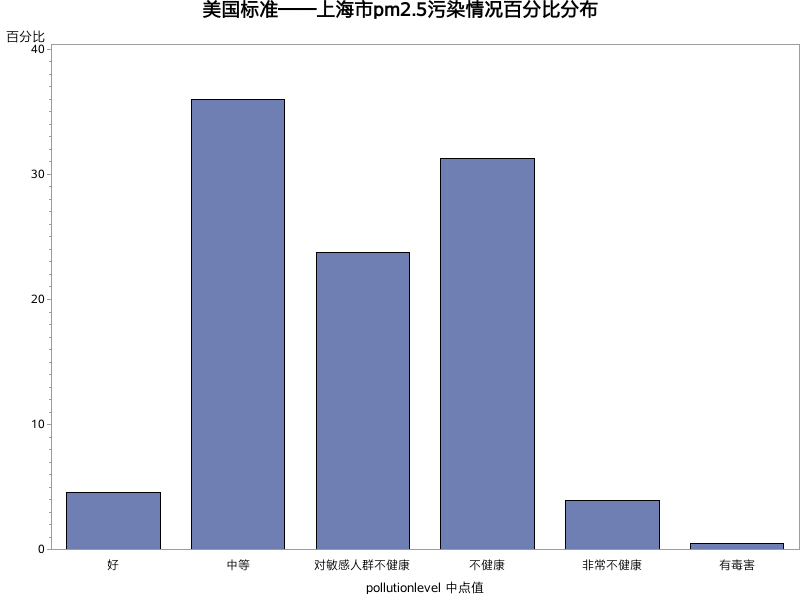


出乎意料的，上海市空气质量为优和良的时间占了77.19%。据悉，我国使用的PM2.5标准与世界卫生组织的PM2.5标准不同，下面使用美国标准重新做比例统计。



中美pm2.5日均浓度对应指数等级

来源 中国环保部官网、美国环保署（EPA）官网



可以发现按美国标准上海的PM2.5浓度水平在好和中等的时间只占50%。

四、协方差分析与回归分析

1.协方差分析

（1）我们首先将缺失值进行多重填补，填补的结果放入一个新的数据集来进行协方差分析以及回归分析。

（2）协方差分析中引入虚拟变量的SAS默认的基水平如下表。

|  |  |
| --- | --- |
| 分类变量 | 基水平 |
| year | 2015 |
| month | 12 |
| hour | 24 |
| day | 31 |
| season | 4 |
| cbwd | cv |

需注意的是sean的虚拟变量与month的虚拟变量之间存在完全的复共线性，故我们在全模型中不包括season，即使包括SAS也会检测到复共线性从而不使用season变量。不仅引入分类变量，我们还增加了DEWP，HUMI，PRES，TEMP之间以及cbwd，iws之间可能的交互效应。

**方差分析总表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 源 | 自由度 | 平方和 | 均方 | F值 | P值 |
| 模型 | 94 | 95566142.5 | 1016661.1 | 734.36 | <.0001 |
| 误差 | 131305 | 181780835.6 | 1384.4 |  |  |
| 校正合计 | 131399 | 277346978.1 |  |  |  |

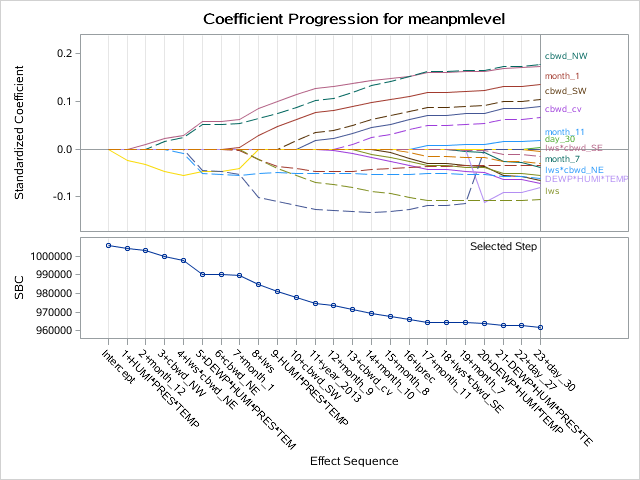
在方差分析总表中，我们看到F值为734.36，P值<0.0001,证明全模型对因变量具有较为明显的解释效应。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 源 | 自由度 | I型平方和 | 均方 | F 值 | P值 |
| year | 2 | 2987760.87 | 1493880.44 | 1079.07 | <.0001 |
| month | 11 | 47939565.65 | 4358142.33 | 3148.00 | <.0001 |
| day | 30 | 2213180.11 | 73772.67 | 53.29 | <.0001 |
| hour | 23 | 273543.86 | 11893.21 | 8.59 | <.0001 |
| DEWP | 1 | 1873608.54 | 1873608.54 | 1353.36 | <.0001 |
| HUMI | 1 | 3038644.38 | 3038644.38 | 2194.89 | <.0001 |
| DEWP\*HUMI | 1 | 412505.23 | 412505.23 | 297.96 | <.0001 |
| PRES | 1 | 3734562.35 | 3734562.35 | 2697.57 | <.0001 |
| DEWP\*PRES | 1 | 970758.38 | 970758.38 | 701.20 | <.0001 |
| HUMI\*PRES | 1 | 602.90 | 602.90 | 0.44 | 0.5093 |
| DEWP\*HUMI\*PRES | 1 | 66979.86 | 66979.86 | 48.38 | <.0001 |
| TEMP | 1 | 316166.18 | 316166.18 | 228.38 | <.0001 |
| DEWP\*TEMP | 1 | 29644.97 | 29644.97 | 21.41 | <.0001 |
| HUMI\*TEMP | 1 | 43136.50 | 43136.50 | 31.16 | <.0001 |
| DEWP\*HUMI\*TEMP | 1 | 282538.06 | 282538.06 | 204.08 | <.0001 |
| PRES\*TEMP | 1 | 30913.13 | 30913.13 | 22.33 | <.0001 |
| DEWP\*PRES\*TEMP | 1 | 500028.00 | 500028.00 | 361.18 | <.0001 |
| HUMI\*PRES\*TEMP | 1 | 13863.25 | 13863.25 | 10.01 | 0.0016 |
| DEWP\*HUMI\*PRES\*TEMP | 1 | 101043.90 | 101043.90 | 72.99 | <.0001 |
| cbwd | 5 | 22385788.31 | 4477157.66 | 3233.97 | <.0001 |
| Iws | 1 | 6336691.54 | 6336691.54 | 4577.16 | <.0001 |
| Iws\*cbwd | 5 | 1586920.74 | 317384.15 | 229.25 | <.0001 |
| precipitation | 1 | 107220.99 | 107220.99 | 77.45 | <.0001 |
| Iprec | 1 | 320474.98 | 320474.98 | 231.49 | <.0001 |

从上述全模型的方差分析表中我们可以看到除了HUMI与PRES的交互作用之外，其余所有解释变量均对模型有显著的解释作用。但全模型使用的变量太多，需进行一定的选择。

2.LASSO REGRESSION以及glmselect过程

我们使用SAS的glmselect过程对模型进行变量选择，选择的算法设定为LASSO，因为LASSO自然地对变量的重要性进行了排序。判断到达模型最优子集的原则设定为SBC。LASSO变量选择过程的系数变化情况以及变量入选的进程如下图。



最终入选的变量如下表：

|  |  |
| --- | --- |
| Parameter | Estimate |
| Intercept | 55.719879 |
| year\_2013 | 8.695799 |
| month\_1 | 22.297001 |
| month\_7 | -6.204367 |
| month\_8 | -10.978182 |
| month\_9 | -12.008698 |
| month\_10 | -9.230074 |
| month\_11 | 3.037157 |
| month\_12 | 28.526608 |
| day\_27 | -1.088090 |
| day\_30 | 0.952805 |
| DEWP\*HUMI\*TEMP | -0.000193 |
| cbwd\_NE | -3.255272 |
| cbwd\_NW | 20.684571 |
| cbwd\_SW | 14.422757 |
| cbwd\_cv | 16.093599 |
| Iws | -0.072961 |
| Iws\*cbwd\_NE | -0.053198 |
| Iws\*cbwd\_SE | -0.017756 |
| Iprec | -0.186161 |

3.变量选择的结论

在最优模型中，2013年的虚拟变量入选，同时具有8.695799的估计值，这说明2013年这一年相较2014,2015年PM2.5的污染情况相对严重，平均要高接近9 μg/m³。

1,7,8,9,10，11,12月份的虚拟变量入选，均为夏秋冬的月份，说明季节因素在PM2.5水平的预测中有重要影响。同时我们注意到夏季PM2.5的水平显著的低，同时冬季PM2.5的水平显著的高。

入选的每个月的第27或30天的系数入选但系数不大并没有过多的现实意义。

入选的DEWP，HUMI，TEMP的交互效应虽然系数较小，但考虑到这一变量值非常大，对预测有非常明显的影响。结合之前的分析，露点越低，湿度越低，温度越低，则PM2.5的浓度越高。所以DEWP\*HUMI\*TEMP的值越低，PM2.5的浓度越高。

注意到五种风向中，东北风，西北风，西南风，无风向入选。东北风从海上吹来，故对PM2.5的浓度有一定的减轻，而从内陆吹来的西北风，西南风则显著提高了PM2.5的浓度，分别到达了20μg/m³与14μg/m³，西北风来自河北，北京等地区，西南风来自内陆，来自河北，北京的风裹挟的PM2.5比来自内陆的风更多也是可以理解的。而cv无显著风向则会使PM2.5的水平明显提高16μg/m³则说明空气流动性对于减少PM2.5也是至关重要的。风速，以及风速与东北风，东南风的交互效应入选则进一步说明了空气流动性的重要性。累计降雨量对PM2.5的浓度有显著的负效应则印证了我们的印象，雨水能够有效吸附PM2.5从而减少PM2.5的浓度。

4.模型评价

|  |  |
| --- | --- |
| Root MSE | 38.82887 |
| Dependent Mean | 56.10326 |
| R-Square | 0.2857 |
| Adj R-Sq | 0.2856 |
| AIC | 1093050 |
| AICC | 1093050 |
| SBC | 961844 |

最终的R2只有0.2857，对于回归模型而言不太理想，可能的原因是因变量存在许多大于300的异常值，与平均值差距非常大。为了克服这一困难，下面我们考虑使用响应变量为多个有序水平的logistic regression来拟合模型获得更好的拟合结果。

五、响应变量有多个有序水平的logistic regression

1.入选变量及系数

我们使用逐步回归法在全模型中进行变量选择。最终入选的变量及其效应如下

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Step | Variable | DF | Chi2 | Pr>Chi2 |
| 1 | **month** | 11 | 19674.8514 | <.0001 |
| 2 | **cbwd** | 5 | 16003.1849 | <.0001 |
| 3 | **Iws** | 1 | 6306.9249 | <.0001 |
| 4 | **TEMP** | 1 | 2334.9496 | <.0001 |
| 5 | **year** | 2 | 1585.3825 | <.0001 |
| 6 | **hour** | 23 | 1491.1501 | <.0001 |
| 7 | **day** | 30 | 1335.4508 | <.0001 |
| 8 | **Iprec** | 1 | 1020.6479 | <.0001 |
| 9 | **PRES** | 1 | 919.7053 | <.0001 |
| 10 | **precipitation** | 1 | 89.6816 | <.0001 |
| 11 | **DEWP** | 1 | 27.8228 | <.0001 |
| 12 | **HUMI** | 1 | 635.0961 | <.0001 |

从卡方统计量可以看出，月份，风向，以及风速对模型的影响是最大的。

2.最大似然估计分析

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | | Estimate | Wald Chi2 | Pvalue |
| Intercept | **1** | -52.59 | 919.3548 | <.0001 |
| Intercept | **2** | -50.3 | 841.8632 | <.0001 |
| Intercept | **3** | -48.97 | 798.2346 | <.0001 |
| Intercept | **4** | -48.08 | 769.4987 | <.0001 |
| Intercept | **5** | -45.79 | 698.0788 | <.0001 |
| year | **2013** | -0.306 | 1556.0427 | <.0001 |
| year | **2014** | 0.145 | 348.2292 | <.0001 |
| month | **1** | -2.364 | 8164.2616 | <.0001 |
| month | **2** | -1.318 | 2682.4965 | <.0001 |
| month | **3** | -0.841 | 1685.0263 | <.0001 |
| month | **4** | -0.452 | 593.4883 | <.0001 |
| month | **5** | 0.146 | 56.323 | <.0001 |
| month | **6** | 1.038 | 2020.9705 | <.0001 |
| month | **7** | 2.068 | 5732.2928 | <.0001 |
| month | **8** | 2.275 | 6863.5581 | <.0001 |
| month | **9** | 1.985 | 6190.4987 | <.0001 |
| month | **10** | 0.816 | 1564.9649 | <.0001 |
| month | **11** | -0.917 | 2198.938 | <.0001 |
| day | **1** | -0.088 | 9.0732 | 0.0026 |
| day | **2** | -0.213 | 55.4243 | <.0001 |
| day | **3** | -0.17 | 34.7638 | <.0001 |
| day | **4** | -0.175 | 36.171 | <.0001 |
| day | **5** | 0.172 | 32.9575 | <.0001 |
| day | **6** | 0.126 | 17.7721 | <.0001 |
| day | **7** | 0.01 | 0.1176 | 0.7317 |
| day | **8** | 0.329 | 113.3324 | <.0001 |
| day | **9** | 0.121 | 16.8627 | <.0001 |
| day | **10** | -0.03 | 1.0505 | 0.3054 |
| day | **11** | -0.148 | 26.0887 | <.0001 |
| day | **12** | 0.042 | 2.0867 | 0.1486 |
| day | **13** | 0.105 | 12.8532 | 0.0003 |
| day | **14** | -0.329 | 127.5691 | <.0001 |
| day | **15** | 0.037 | 1.62 | 0.2031 |
| day | **16** | 0.068 | 5.1685 | 0.023 |
| day | **17** | 0.272 | 80.2193 | <.0001 |
| day | **18** | 0.3 | 97.8516 | <.0001 |
| day | **19** | 0.273 | 82.0238 | <.0001 |
| day | **20** | -8E-04 | 0.0008 | 0.9778 |
| day | **21** | -0.14 | 22.8153 | <.0001 |
| day | **22** | -0.153 | 27.1127 | <.0001 |
| day | **23** | -0.01 | 0.1133 | 0.7364 |
| day | **24** | -0.013 | 0.2002 | 0.6546 |
| day | **25** | 0.239 | 64.6221 | <.0001 |
| day | **26** | -0.391 | 177.0296 | <.0001 |
| day | **27** | 0.23 | 57.1631 | <.0001 |
| day | **28** | 0.076 | 6.4617 | 0.011 |
| day | **29** | -0.307 | 101.1159 | <.0001 |
| day | **30** | -0.284 | 88.2083 | <.0001 |
| hour | **0** | -0.034 | 1.6622 | 0.1973 |
| hour | **1** | 0.013 | 0.2457 | 0.6201 |
| hour | **2** | 0.033 | 1.4807 | 0.2237 |
| hour | **3** | 0.027 | 1.021 | 0.3123 |
| hour | **4** | 0.009 | 0.1208 | 0.7281 |
| hour | **5** | -0.032 | 1.3933 | 0.2378 |
| hour | **6** | -0.11 | 16.8383 | <.0001 |
| hour | **7** | -0.184 | 48.8718 | <.0001 |
| hour | **8** | -0.145 | 31.3063 | <.0001 |
| hour | **9** | -0.062 | 5.7574 | 0.0164 |
| hour | **10** | 0.052 | 3.8401 | 0.05 |
| hour | **11** | 0.193 | 52.2488 | <.0001 |
| hour | **12** | 0.259 | 92.9246 | <.0001 |
| hour | **13** | 0.316 | 135.7497 | <.0001 |
| hour | **14** | 0.341 | 157.0752 | <.0001 |
| hour | **15** | 0.266 | 95.6495 | <.0001 |
| hour | **16** | 0.211 | 61.0387 | <.0001 |
| hour | **17** | 0.111 | 17.373 | <.0001 |
| hour | **18** | -0.096 | 13.5282 | 0.0002 |
| hour | **19** | -0.25 | 93.3322 | <.0001 |
| hour | **20** | -0.319 | 152.5387 | <.0001 |
| hour | **21** | -0.261 | 100.3722 | <.0001 |
| hour | **22** | -0.192 | 53.7887 | <.0001 |
| DEWP |  | -0.164 | 619.1288 | <.0001 |
| HUMI |  | 0.042 | 606.3495 | <.0001 |
| PRES |  | 0.048 | 828.517 | <.0001 |
| TEMP |  | 0.098 | 235.2108 | <.0001 |
| cbwd | **.** | 0.78 | 3.3901 | 0.0656 |
| cbwd | **NE** | 0.701 | 67.492 | <.0001 |
| cbwd | **NW** | -0.927 | 117.2635 | <.0001 |
| cbwd | **SE** | 0.587 | 47.2689 | <.0001 |
| cbwd | **SW** | -0.432 | 25.3384 | <.0001 |
| Iws |  | 0.008 | 5862.2915 | <.0001 |
| precipitation |  | -0.092 | 88.9048 | <.0001 |
| Iprec |  | 0.121 | 1068.8658 | <.0001 |

在上述模型中，除了日期，时间中的部分变量，以及cbwd的缺失水平，其余变量均在Wald 卡方的意义下显著。logistic回归的系数所表现出的对污染水平的正负效应与回归分析中的结论基本一致。

3.优比分析

|  |  |  |  |
| --- | --- | --- | --- |
| Odds Ratio | | | |
| Variable | **Estimate** | **95% Wald** | |
| **CF Interval** | |
| year 2015-2013 | 0.627 | 0.61 | 0.643 |
| year 2015-2014 | 0.984 | 0.96 | 1.01 |
| month 12-1 | 1.072 | 1.02 | 1.126 |
| month 12-2 | 3.053 | 2.9 | 3.22 |
| month 12-3 | 4.921 | 4.66 | 5.193 |
| month 12-4 | 7.258 | 6.83 | 7.711 |
| month 12-5 | 13.198 | 12.3 | 14.18 |
| month 12-6 | 32.212 | 29.7 | 34.95 |
| month 12-7 | 90.194 | 82.3 | 98.86 |
| month 12-8 | 110.921 | 101 | 121.6 |
| month 12-9 | 83.013 | 76.3 | 90.28 |
| month 12-10 | 25.797 | 24.1 | 27.64 |
| month 12-11 | 4.558 | 4.31 | 4.821 |
| hour 23-0 | 1.116 | 1.04 | 1.202 |
| hour 23-1 | 1.17 | 1.09 | 1.26 |
| hour 23-2 | 1.193 | 1.11 | 1.285 |
| hour 23-3 | 1.187 | 1.1 | 1.278 |
| hour 23-4 | 1.166 | 1.08 | 1.256 |
| hour 23-5 | 1.119 | 1.04 | 1.205 |
| hour 23-6 | 1.035 | 0.96 | 1.115 |
| hour 23-7 | 0.961 | 0.89 | 1.034 |
| hour 23-8 | 0.999 | 0.93 | 1.075 |
| hour 23-9 | 1.085 | 1.01 | 1.169 |
| hour 23-10 | 1.216 | 1.13 | 1.311 |
| hour 23-11 | 1.401 | 1.3 | 1.511 |
| hour 23-12 | 1.497 | 1.39 | 1.616 |
| hour 23-13 | 1.584 | 1.47 | 1.71 |
| hour 23-14 | 1.625 | 1.5 | 1.754 |
| hour 23-15 | 1.506 | 1.4 | 1.626 |
| hour 23-16 | 1.426 | 1.32 | 1.539 |
| hour 23-17 | 1.29 | 1.2 | 1.392 |
| hour 23-18 | 1.049 | 0.97 | 1.13 |
| hour 23-19 | 0.899 | 0.84 | 0.968 |
| hour 23-20 | 0.839 | 0.78 | 0.903 |
| hour 23-21 | 0.89 | 0.83 | 0.958 |
| hour 23-22 | 0.953 | 0.89 | 1.026 |
| DEWP | 0.849 | 0.84 | 0.86 |
| HUMI | 1.043 | 1.04 | 1.047 |
| PRES | 1.049 | 1.05 | 1.052 |
| TEMP | 1.103 | 1.09 | 1.117 |
| cbwd cv-. | 4.43 | 1.63 | 12.01 |
| cbwd cv-NE | 4.096 | 3.87 | 4.34 |
| cbwd cv-NW | 0.804 | 0.76 | 0.852 |
| cbwd cv-SE | 3.653 | 3.45 | 3.874 |
| cbwd cv-SW | 1.319 | 1.24 | 1.403 |
| Iws | 1.008 | 1.01 | 1.008 |
| precipitation | 0.912 | 0.9 | 0.93 |
| Iprec | 1.128 | 1.12 | 1.136 |

由上表优比估计值可以得出结论：

一、从年份来看，2015年与2013年相比，pm2.5值显著降低；2014年与2013相比，pm2.5值也显著降低；而2014年与2015年差异不明显。可知说明2014年与2015年治理雾霾有一定效果。

二、从月份来看，12月份与其他月份优比值均大于1，说明12月份pm2.5值比较高，其中12月份与8月份优比估计值高达110.921，说明冬季与夏季相比，pm2.5值在冬季较高。也可得知pm2.5受季节影响明显。

三、从每天时间来看，从总体上来说，23点与其他时间的优比估计值先降低后增加再降低后增加，所以一天中的pm2.5值有两个峰值。23点与15点的优比值在增加阶段达到最大，说明一天中大约15点的pm2.5值达到最小；而23点与7点的优比估计值在第一降低阶段达到最低，说明一天中大约7点为pm2.5值出现第一个峰值；而第二个降低阶段中23点与20点优比值最小，说明一天中大约20点出现第二个峰值。

四、露点优比值为0.849，说明露点越低，pm2.5值越高；而湿度、气压、温度、累计风速和每小时降雨量的优比值接近1，说明其影响不明显。

五、组合风向优比值为4.43，说明风向对pm2.5的影响显著，其中风向为东北和东南时，pm2.5值显著较高；风向为西南时，pm2.5值较高；而风向为西北时，pm2.5值较低。

六、累计降雨优比值为1.128，说明累计降雨量越大，pm2.5值越高。

4.模型评价

|  |  |
| --- | --- |
| Predicted Response and Actual Response | |
| Same | 77.8 |
| Diffrent | 21.8 |
| Tie | 0.4 |

使用这一logistic回归模型，我们获得了77.8%的预测准确率。相比回归模型在预测方面显著地做的更好。