青岛大学数学与统计学院回归分析成果

论文题目：《青岛市空气质量影响因素分析》

小组成员：李海豹 马志航 侯同帅 唐超 周斌杰 翟泓智 辛悦

所在学院：数学与统计学院

专业班级：应用统计学一班

时间：2018 年 12 月 29 日

目录

摘要：本文通过对青岛近年来空气质量及各项污染物数据的搜集，分析近年来青岛空气质量变化趋势及各影响因素对空气质量的影响程度。利用回归分析的手段建立回归分析模型，进行最小二乘估计，复共线性分析，主成分分析，岭估计，回归方程的显著检验，回归系数的显著性检验，异常点检验，逐步回归分析；使用 MATLAB 等软件定量计算，给出分析结果，并对青岛市空气质量的监控提出意见和建议。

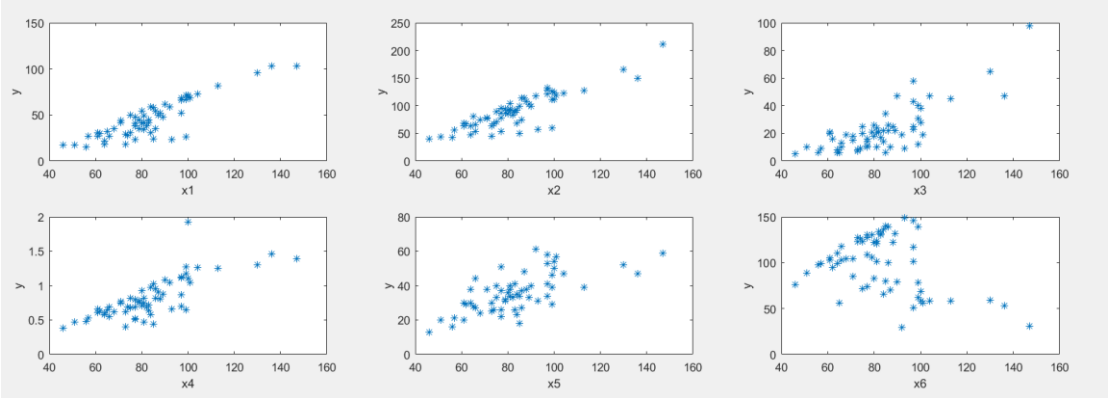
关键字：AQI 指数 线性回归 显著性检验 MATLAB

背景：2017 年青岛市区空气中主要指标年均值方面，PM2.5 为 37 微克/立方米、PM10 为 76 微克/立方米、SO_2为 14 微克/立方米、NO_2为 33 微克/立方米。与 2016 年相比，PM2.5、PM10 分别改善 17.8%、10.6%，均为 2013 年以来最好水平；SO_2、NO_2连续两年稳定达到国家一级标准。

一. AQI 指数及影响因素相关性分析

青岛市空气质量数据网站显示，对 AQI 指数影响最大的几个因素为 PM2.5、PM10、SO_2、CO、NO_2、O_3。

通过观察 AQI 的边际模型图来判断是否存在线性关系，其散点图（scatter diagram）分布如下：



AQI 与其各影响变量间拟合效果良好。

为进一步对其相关性进行分析，通过计算它们之间的判定系数 R^2 来判断相关关系的强度，这里

$$R^2=[\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})]^2/[\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(x_i-\bar{x})^2]$$

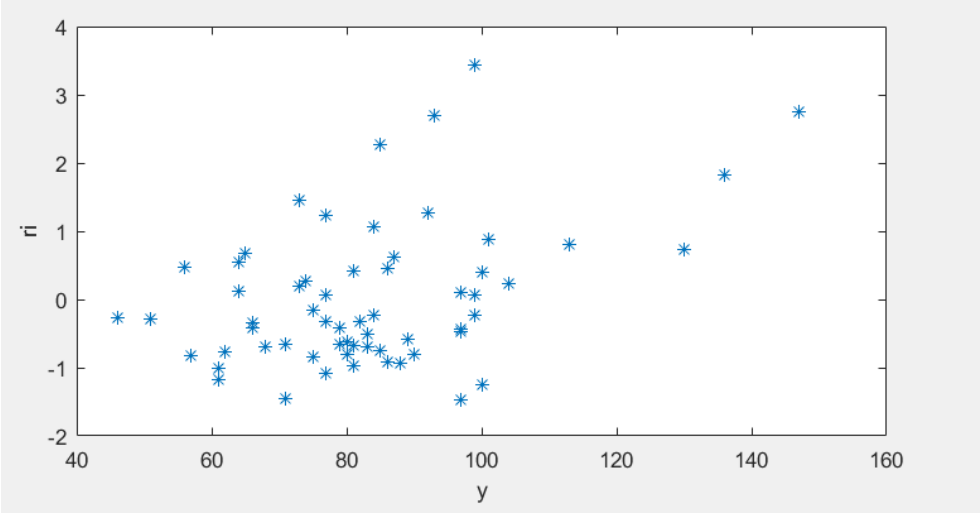
在 MATLAB 中计算得出 R^2=0.8472。AQI 与 PM2.5、PM10 的相关系数分别为 0.8732、0.8703，说明 AQI 与各影响因素的拟合效果良好，尤其 AQI 与 PM2.5，PM10 的相关性极强。F 检验的 p-value < 3.79e-5，说明存在过拟合现象，需要对某些变量进行处理。

二．回归模型的建立及求解

我们假设各变量间满足以下关系：$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\varepsilon$。若令 $Y=(y_1,y_2,...,y_n)^T$，$\beta=(\beta_0,\beta_1,...,\beta_6)^T$，则上述关系式可表示成矩阵形式：$Y=X\beta$。在最小二乘回归估计下，求得 $\hat{\beta}=(12.2681,0.6410,0.0964,0.1389,4.6614,0.1375,0.2171)^T$，即原模型可以表示为：$y=12.2681+0.641x_1+0.0964x_2+0.1389x_3+4.6614x_4+0.1375x_5+0.2171x_6$。$E(\varepsilon)=-1.4151e-12$，$D(\varepsilon)=53.99$，$\hat{\sigma}^2=59.99$，则 ε 满足高斯--马尔科夫假定 $\varepsilon\sim N(0,\sigma^2)$。

三．检验模型拟合优良程度

因变量 y 的预测 $\hat{y}=X\hat{\beta}$，残差 $\hat{e}=y-\hat{y}$，对残差 $\hat{e}$ 进行学生化，学生化结果记为 r，作出 y 关于 r 的散点图，如图所示：

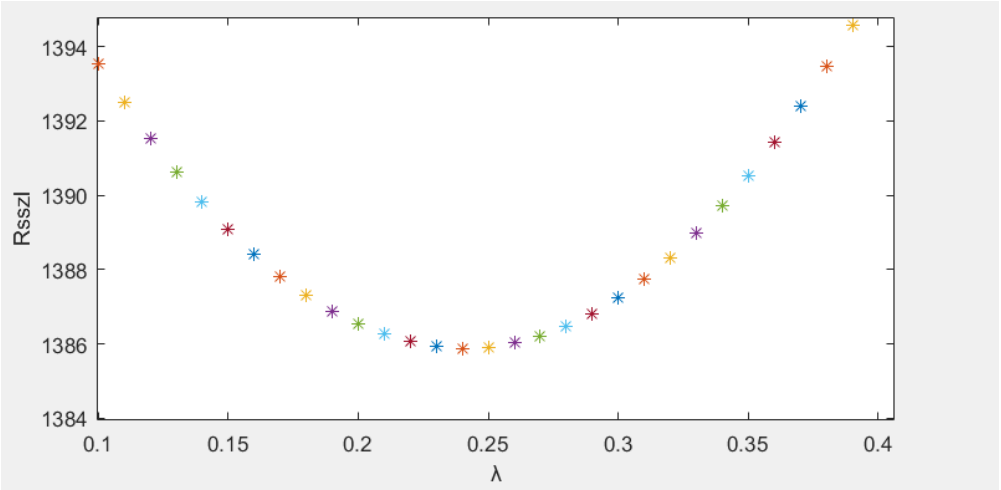


学生化残差 r 大致在-2~2 之间呈散乱无规律分布(仍有有几个点分布比较突出，后面做异常点处理和 box-cox 变换)，需要进一步处理，才能转入复共线性检验。

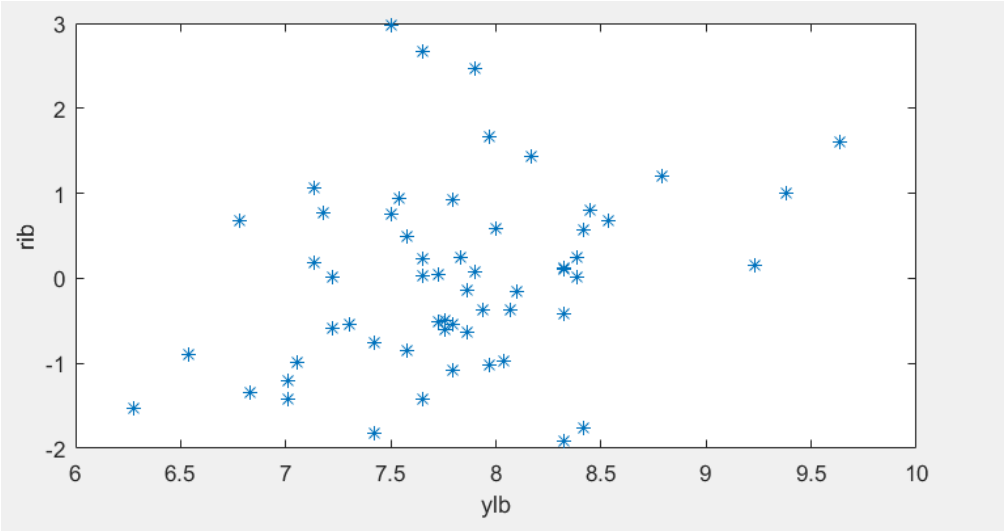
四．Box-Cox 变换

通过上面的分析，我们已经知道学生化残差存在不太明显的规律性活数据存在异常点，故先而进行 box-cox 变换来尝试减弱这种规律性。

对给定的$\lambda \in (-0.5,1)$，进行 box-cox 变换，变换下的$RSS_{z^{\lambda}}$变化如下图所示：



得到使 $RSS_{z^{\lambda}}$ 达到最小时的λ=0.240，选取这个值进一步计算学生化残差，得到学生化残差分布图如下：



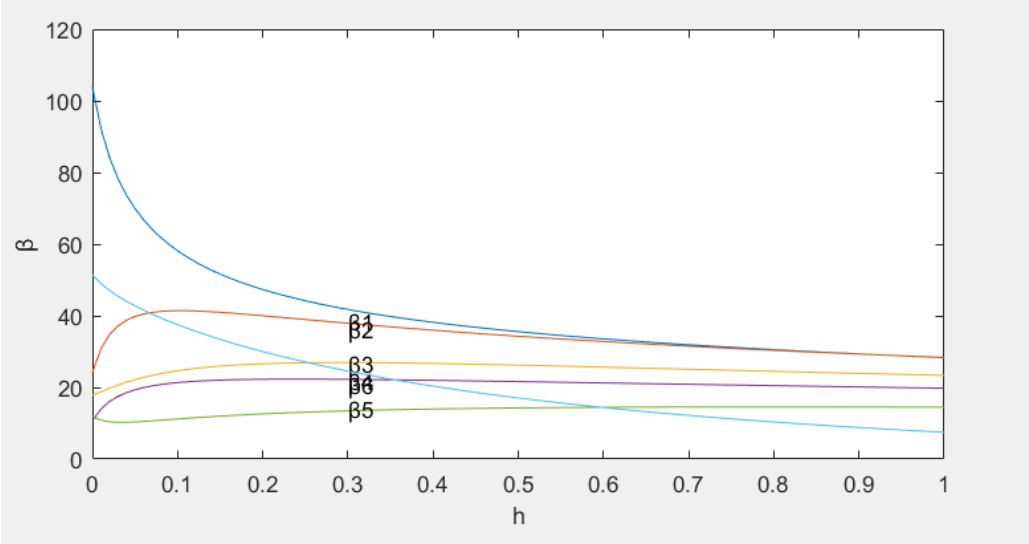
已经无规律分布，可以进行下一步工作。

五．复共线性强度检验

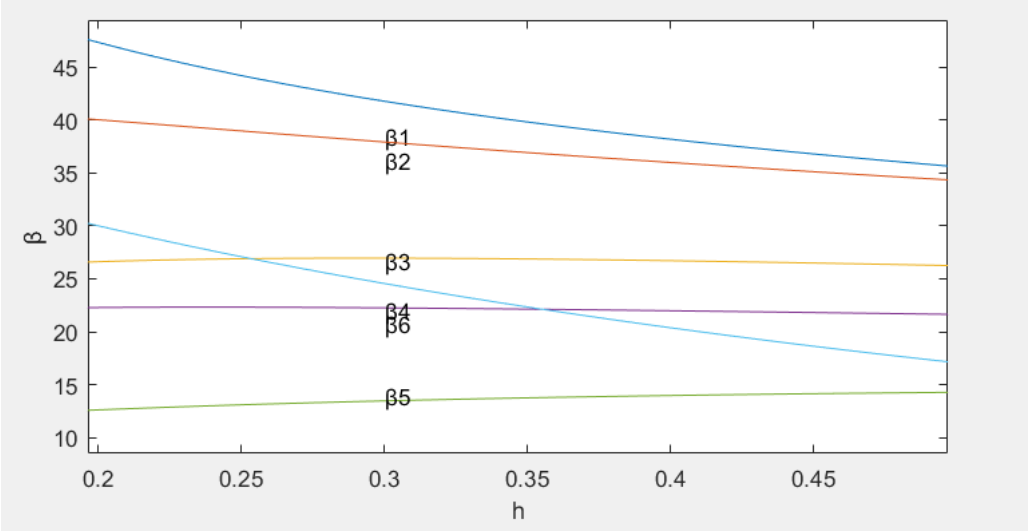
对矩阵 X 做中心化和标准化，并记中心化和标准化之后的结果为 Z,求得 Z’ Z 的特征值分别为λ=[4.7293，0.5594，0.35745，0.19189，0.1240，0.0379]，条件数 k=124.88,故可以判定存在中等强度的复共线性。

六．岭估计

记岭参数为 h，$0 \le h \le 1$,分割精度取 0.01，此时回归系数 β 的岭估计为：$\hat{\beta}(h) = (Z'Z + hE)^{-1}Z'y$，作出 $\hat{\beta}(h)$ 的各个分量随 h 的变化而变化的图示，如图所示：



其在 0.2-0.5 上的局部放大图如图所示：



由上图可知，在 h=0.35 时，$\hat{\beta}(h)$ 的各个分量已基本趋于稳定，故我们选取 h=0.35 的估计，此时 $\hat{\beta}(0.35)=(39.81,36.92,26.87,22.13,13.75,22.34)^{T}$，岭回归方程为：

$$\hat{Y}=11.13X_1+8.38X_2+16.48X_3+549.18X_4+9.05X_5+5.37X_6+2880.73$$

七．主成分估计

取中心化和标准化之后的矩阵 Z，上面已求得 Z’Z 的特征值分别为λ=[4.7293，0.5594，0.35745，0.19189，0.1240，0.0379]，并计算得各步累计贡献率分别为[0.788,0.881,0.941,0.973,0.994,1]；

当取到第四个值的时候，累计贡献率已经达到 97.3% > 95%,因此剔除后两个主成分，保留前四个，计算前四个主成分分别为：

$Z_1 = -0.44X_1 - 0.43X_2 - 0.41X_3 - 0.42X_4 - 0.39X_5 + 0.35X_6$

$Z_2 = 0.17X_1 + 0.27X_2 + 0.44X_3 + 0.04X_4 - 0.33X_5 + 0.76X_6$

$Z_3 = 0.03X_1 - 0.31X_2 - 0.34X_3 + 0.85X_4 - 0.05X_5 + 0.23X_6$

$Z_4 = 0.07X_1 - 0.22X_2 + 0.35X_3 + 0.14X_4 - 0.76X_5 - 0.47X_6$

还原后的经验方程为：

$$\hat{Y} = 81.31 + 3.86X_1 + 2.47X_2 + 4.68X_3 + 256.77X_4 + 5.84X_5 - 1.6X_6$$

八．回归方程的显著性检验

在最小二乘估计回归方程下，对回归方程的显著性进行检验。作如下假设：

H_0：$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$，计算得 $\beta_0^* = 83.2951$，$RSS_H = 21198.6$，$RSS = 3239.4$，在原假设成立的情况下 $F = [(RSS_H - RSS)/(p-1)]/[RSS/(n-p)]$ 服从 F 分布，计算 F 统计量的值为 F=49.8959，相应的 p-value=3.79e-5，显然拒绝原假设，即最小二乘估计下的各变量系数不全为零。

九．系数的显著性检验

作如下假设：H_i：$\beta_i = 0$，方阵 $C = (X'X)^{-1}$，在原假设成立的情况下，$t = \beta_i/(\hat{\sigma}\sqrt{C_{(i,i)}})$ 服从 t 分布，计算得各假设下的 t 检验统计量为：

t = [3.39, 0.88, 1.08, 0.58, 0.79, 4.57]，相应的 p-value 分别为：

p = [0.0006，0.19，0.14，0.28，0.22，1.43e-05]；自然，对于

X_1和X_6的假设我们有足够的理由拒绝，而对于其他各变量的系数的假设我们不能轻易拒绝，因此还需进一步进行变量分析。

十．异常点检验

由第四步的分析知，这组数据中存在差异较大的点，称之为异常点（outlier），本步骤的工作就是寻找出其中的异常点，并剔除。

将各组数据对应的残差列成表格如下，并将其中值比较大的点突出表示，这些点最有可能是数据中的异常点：

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
13.42	5.12	-5.81	-10.72	-6.959	-3.347	-2.40	-3.74	-2.39	-1.19	-5.15	-5.74	-2.80	1.73	-1.67
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
-6.75	-4.31	-3.13	3.30	-8.64	-5.26	-7.43	-5.04	-1.69	11.92	5.81	-5.55	-7.28	0.69	-4.65
31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
-2.50	-6.06	-2.06	7.95	-5.74	-6.29	2.93	0.49	-6.07	-10.77	-5.06	25.44	9.20	1.94	10.70
46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
19.57	-2.96	0.48	8.45	6.18	-4.96	-7.91	3.08	1.45	16.77	-1.79	3.41	4.01	0.94	4.28
61														
4.54														

通过计算得到这组残差的均值为$\mu_e = 0$，y 的方差为$\sigma_y = 7.75$；查阅资料了解到判断异常点的一个方法是 2σ 方法，即若$\mu_e - 2\sigma_y < e_i < \mu_e + 2\sigma_y$，则第 i 组数据正常，否则为异常。

最终得到异常点数据为第 42 组，第 46 组，第 55 组。从而，应该剔除这三组数据在进行下一步处理。

十一．逐步回归

记 B = X’ X,增广矩阵 A 可表示为 $A=\begin{pmatrix} B & X'y \\ y'X & y'y \end{pmatrix}$，检验的显著性水平取 α = 0.05。

增广矩阵 $A^{(0)}$如下：

60.00	56.30	51.68	53.12	46.04	-39.06	984.84
56.30	60.00	51.25	48.77	47.60	-33.11	981.62
51.68	51.25	60.00	46.72	37.53	-32.74	896.60
53.12	48.77	46.72	60.00	46.10	-38.79	868.97
46.04	47.60	37.53	46.10	60.00	-37.17	762.04
-39.06	-33.11	-32.74	-38.79	-37.17	60.00	-415.41
984.84	981.62	896.60	868.97	762.04	-415.41	444421

引入第一个变量，对于一元回归方程：$y=\beta_0+\beta_i X_i+e$，i=1,2,3,4,5,6.

以 $A^{(0)}(i,i)$为枢轴进行枢轴消去变换，并记 $A^{(1)}=T_1A^{(0)}=(a_{ij}^{(1)})$,则有：$\hat{\beta}_i=a_{i7}^{(1)}=a_{i7}^{(0)}/a_{ii}^{(0)}$，$RSS_1(i)=a_{77}^{(1)}=a_{55}^{(0)}-[a_{i7}^{(1)}]^2/a_{11}^{(0)}$,i = 1,2,3,4,5,6.

变量 X_i 的偏回归平方和 $p_i^{(1)}$为：$p_i^{(1)}=[a_{i7}^{(0)}]^2/a_{ii}^{(0)}$, i = 1,2,3,4,5,6.

经计算，最大的为 $p_2^{(1)}$ = 49.5317,变量 X_2 对相应的检验统计量的值为：

$F_{引}^{(1)}=371.40>F_{1,56}(0.05)=4.0130$,引入变量 X_2. 对 $A^{(0)}$施行（2,2）为枢轴的消去变换，得到 $A^{(1)}$

7.16	-0.94	3.57	7.40	1.51	-8.23	3.36
0.93	0.02	0.85	0.81	0.79	-0.52	0.93
3.56	-0.85	15.95	4.95	-3.11	-4.85	2.53
7.40	-0.80	4.95	19.93	7.08	-12.53	3.23
1.51	-0.78	-3.11	7.08	21.78	-11.41	-0.92
-8.23	0.51	-4.85	-12.53	-11.41	41.87	2.71
3.36	-0.93	2.53	3.23	-0.92	2.71	7.47

经计算，最大的为 $p_1^{(2)}$ = 1.5810,变量 X_1 对相应的检验统计量的值

为：

$F_{引}^{(2)}$ = 14.77>$F_{1,56}$(0.05)=4.0130,引入变量 X_1.对 $A^{(1)}$施行（1,1）为枢轴的消去变换，得到 $A^{(2)}$

0.14	-0.13	0.49	1.03	0.21	-1.15	0.47
-0.13	-0.14	0.38	-0.16	0.59	0.56	0.49
-0.49	-0.38	14.16	1.26	-3.86	-0.75	0.85
-1.03	0.16	1.26	12.27	5.52	-4.03	-0.25
-0.21	-0.59	-3.87	5.52	21.46	-9.67	-1.63
1.15	-0.56	-0.75	-4.03	-9.67	32.42	6.57
-0.47	-0.49	0.85	-0.25	-1.64	6.57	5.89

经计算，最大的为 $p_6^{(3)}$ = 1.3327,变量 X_6 对相应的检验统计量的值为：

$F_{引}^{(3)}$ = 15.80>$F_{1,56}$(0.05)=4.0130,引入变量 X_6.对 $A^{(2)}$施行（6,6）为枢轴的消去变换，得到 $A^{(3)}$

0.18	-0.15	0.47	0.89	-0.13	0.04	0.70
-0.15	0.15	0.39	-0.09	0.75	-0.02	0.38
-0.47	-0.39	14.15	1.17	-4.10	0.02	1.00
-0.89	0.09	1.17	11.78	4.32	0.12	0.57
0.13	-0.75	-4.10	4.32	18.58	0.29	0.33
0.035	-0.02	-0.02	-0.12	-0.29	0.03	0.20
-0.70	-0.37	1.00	0.57	0.32	-0.20	4.55

下一步开始剔除变量，在已经引入的变量中，计算最小的偏回归平方和为 $p_2^{(4)}$=0.96,变量 X_2 对应的检验统计量为 $F_{剔}^{(4)}$ = 14.45<$F_{1,56}$(0.05) = 4.0130,故没有变量可以剔除。

接下来进一步判断是否还有变量可以引入，计算未引入的变量的偏回归平方和最大的为 $p_3^{(5)}$=0.07,变量 X_3 对应的检验统计统计量的值为 $F_{引}^{(5)}$=0.8389<$F_{1,56}$(0.05)=4.0130,变量 X_3 不可引入，即已经没有可以引入的变量，逐步回归结束。

分析各个变量间的相关系数，其相关系数矩阵如下：

1	0.94	0.86	0.88	0.76	-0.65
0.94	1	0.85	0.81	0.79	-0.55
0.86	0.85	1	0.77	0.62	-0.54
0.88	0.81	0.77	1	0.76	-0.64
0.76	0.79	0.62	0.76	1	-0.61
-0.65	-0.55	-0.54	-0.64	-0.61	1

发现各个变量间的线性关系均很强，因此变量的去除是很有意义的！

最终得到最优回归方程为：Y = 0.7032X1 + 0.3791X2 + 0.2028X6

结束语：

通过上述分析，发现对空气质量影响最大的三个因素为 PM2.5，PM10，O_3，因此我们对青岛市政府提出如下建议:

1. 加强工业排放监管，减少颗粒物污染；

2. 进一步加强环境绿化治理，形成人类生活自然防线；3

3. 减少氟化物的使用，为保护大气层做出更大的努力。

参考文献：

[1] 农林. 东营空气质量指数的相关性分析[J]. 财经界(学术版),2014(22):278-278.

[2] 张争辉,汪蜜,陈丽贞,朱家明. 蚌埠市空气质量影响因素计量分析[J]. 河北北方学院学报(自然科学版)，2016(03):50-60.