统计数据分析 2007 期末论文
重点城市工业废气排放情况调查

林荟 05231113
陈鹏举 05231007

2008 年 1 月 4 日

目 录

§1 数据背景

§1.1 数据来源

随着生产力的发展和社会的不断前进，人类给自己创造了巨大的财富的同时，也给自己生存环境带来了很大的危害，人口爆炸、环境污染、传染疾患、生态战争等等，时时都在影响着人类的健康，威胁着人们的生命，与此同时，无论是专家学者还是普通人群都在探讨人类健康发展之路，环境保护已成为全球人类共同关心的热门话题。北京申办奥运时响亮地提出了绿色奥运的口号。在此背景之下，我们选取 06 年重点城市工业废气排放情况调查的数据作为分析对象，旨在从一定程度上揭示国家废气排放的现状，为国家废气治理提供一些意见，贡献一份力量。

§1.2 数据说明

原始数据有 113 个，分别描述了燃料煤消费量，原料煤消费量，燃料油消费量以及废气排放总量。除去一些不完整的数据，最后剩下 106 个数据。我们只对其中最主要的两项：废气排放总量以及燃料煤消费量进行研究分析。(见附录一)

§2 研究目的

1、分析废气排放的分布，揭示废气排放的现状。
2、基于之前的分析，给出用抽样的方式估计废气排放情况的方法。
3、结合燃料煤的消费量，分析二者的潜在关系。

§3 数据分析

§3.1 综述

我们整个数据分析的过程有 5 个主题：基本分布情况，耐抗性，残差，重新表达以及启示。它们常常组合起来。

1、基本分布情况：对一批数据分布的特点进行初步分析，如密集度，对称性，离群值等。

2、耐抗性：对于局部不良行为的非敏感性分析。人们知道：“好”数据也难免有那么百分之几的大错，因此要有防御大错的破坏性影响的措施，这促使人们关注耐抗性。

3、残差：从数据减去一个总括统计量或拟合模型以后的残余部分。当数据的大部分遵从一致的模式，这个模式决定一个耐抗拟合。耐抗残差包含对于这个模式的剧烈偏离及机遇起伏。有反常的残差就要求检查相应观测值产生过程与处理的详情。

4、重新表达：原始表达不合适时对数据做一个变换也许有足于促进对称性、变异恒定性、关系直线性或效应的加性。

5、启示：通过对数据的分析，抓住数据特点得出一些有指导意义的结论。

§3.2 对废气总排放量的分析

通过做出该批数据的统计图表分析数据的分布情况。

§3.2.1 茎叶图

$stem(total)\ Removed$

$N = 106\ Median = 1308.5\ Quartiles = 603, 1988$

$Decimal\ point\ is\ 3\ places\ to\ the\ right\ of\ the\ colon$

0 : 012233333344444

0 : 55555666666667788888999

1 : 000111111122333334444

1 : 55556666677777888889

2 : 00124

2 : 556677

3 : 0134

3 : 89

4 : 0

4 : 6

$High$: 5858 6512 6656 6757 6984 8523 9428 13511

茎叶图可以使我们看到整个一批数据：对称程度，展布，是否有些数远离其余数，是否有数据集中，数据是否有间隙。这是一种基本而通用的方法。于是我们决定先做出这批数的茎叶图对数据分布特征初步分析。以上是用 splus 做出的废气排放量的茎叶图。由茎叶图可以看出，这批数据有 8 个数据明显高于其余数据。数据主体部分显示出非对称性，小的值偏多，数据集中在 $0 \sim 2000$ 范围内。

§3.2.2 箱线图

图 1 废气总排放量箱线图

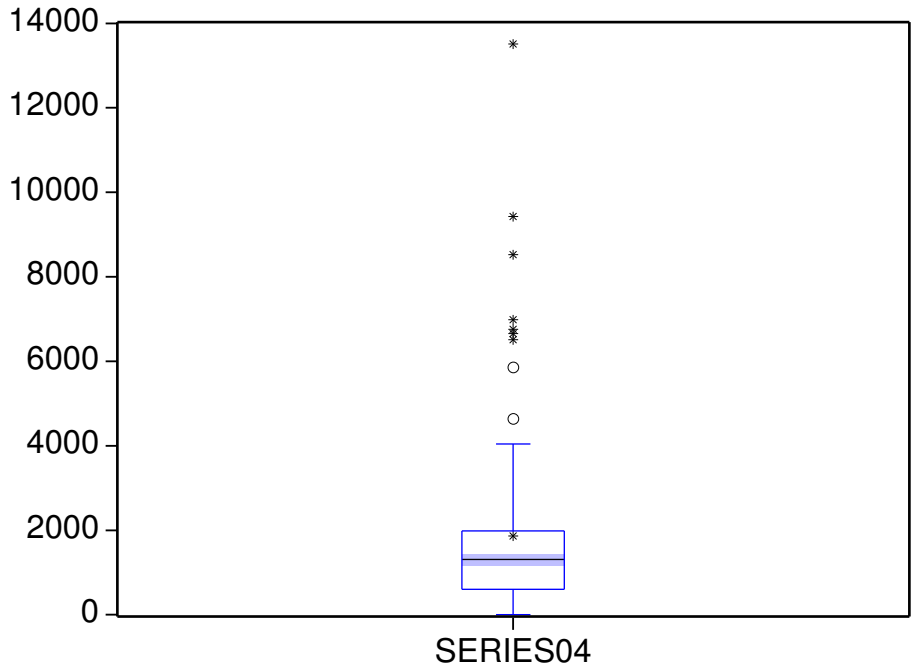
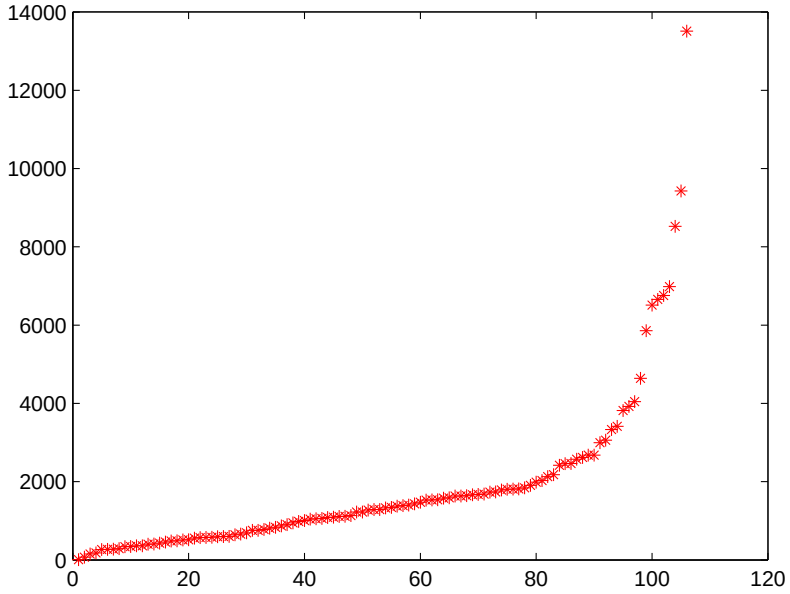


表 1 重点城市工业废气排放量 12 个离群值表

唐山	上海	石家庄	苏州	重庆	邯郸	天津	本溪	北京	鞍山	南京	包头
13511	9428	8523	6984	6757	6656	6512	5858	4641	4044	3921	3818

箱线图可以将一批数据的经验分布的一些重要方面给人视觉印象。因此我们接下来用箱线图进一步刻画数据分布情况。

图 2 原尺度下废气排放量图



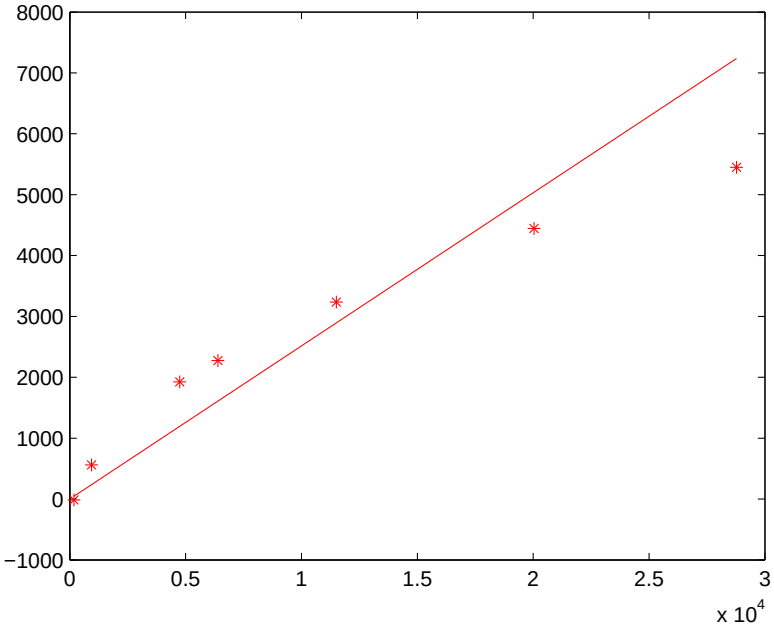
由箱线图（图 1）上四分数与下四分数相对于中位数的位置可以看出，这个批是几乎不偏斜的，但有 12 个离群值，且都是较大离群值。如表 1 所示. 这 12 个城市废气排放量明显高于其余的城市，应在废气治理过程中予以特别考虑。与之前的茎叶图相比，箱线图判断出的离群值多了 4 个，但同样都是较大离群值。说明了我国有几所城市废气排放量特别高，但大部分还是积聚在一个范围之内。基于人力财力有限，国家对每年对重点城市废气排放量进行普查是不现实的，通过以上分析，我们可以做出如下建议：采取分层抽样的方式，将 12 个较大离群值作为一层，全部抽查，对于其余的数据再分层，然后在各层内抽取一些作为代表估计全国废气排放量的情况。

§3.3 数据变换

一个批的对称性常常是想望的性质；位置的许多估计量，当数据来自对称分布时工作的最好，最能被理解。因此，为了更好的反映废气排放量的分布情况，我们决定对该批数据进行对称性变换。

§3.3.1 对称性变换图

图 3 废气排放量对称性变换图

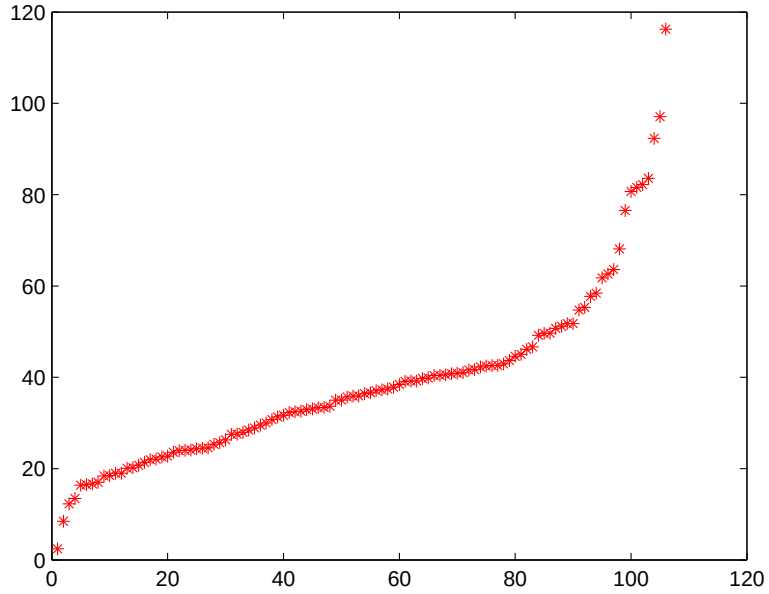


检查对称性的最简单的办法是定义一组中总括，对于每一对字母值定义一个。每个中总括是这一对相应的字母值的平均值。一旦我们算出所有各对字母值的中总括，我们就可以考查它们，寻求系统性偏斜的证据。(如果明显的偏斜由一两个离群值造成，那么只有最极端的字母值及它们的中总括才受影响。因此，用整个一组中总括提供更多的耐抗性。) 在一个完美的对称的批中，所有中总括应该等于中位数。如果数据右偏，那么中总括将随着相应的那对字母值向外移动而增加。而对于左偏的数据，中总括将随之减少。表是废气排放量样本带总括的字母值显示，由表 2 中原尺度下废气排放量的中总括单调递减的趋势知：数据有某种右偏的趋势。现在我们要找到把这个批变得更加对称的变换。我们决定采用作对称性变换图的方法。

令 M 是一个批的中位数，并且令 X_L 和 X_U 表示各个字母值的下字母值和上字母值。对称变换图在水平坐标轴上放

$$\frac{(x_U - M)^2 + (M - x_L)^2}{4M} \tag{1}$$

图 4 平方根下废气排放量图



并且在垂直坐标上放

$$\frac{x_U + x_L}{2} - M \tag{2}$$

如果得到的图象近似线性，那么,1 减斜率是为对称而变换的指示幂指数，这个变换有以下形式:

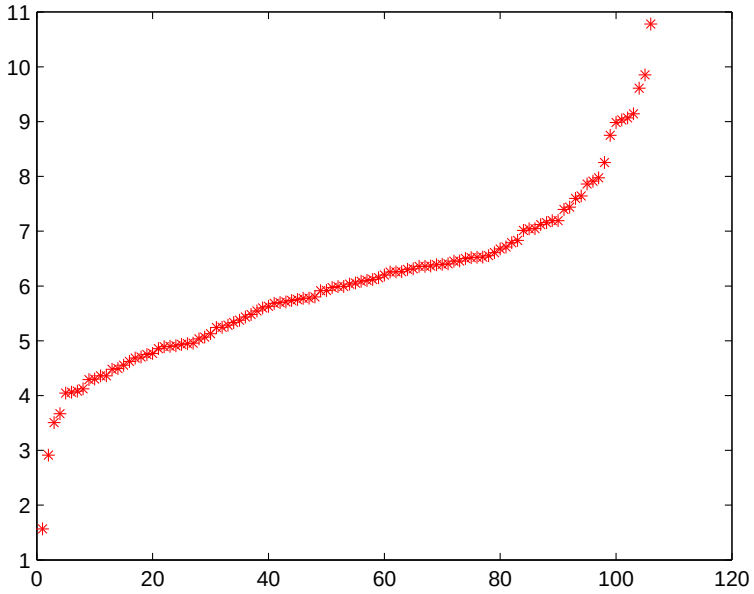
$$T(x) = kx^p \tag{3}$$

就像其它诊断图，对称性变换图是要给我们一个初步近似选择一个好的变换。它告诉我们接着要试什么，但不保证我们要试的就是最好的。

§3.3.2 尺度变换

图 3 给出了废气排放量的对称性变换图。有多种方法给上图中的 7 个点拟合一条直线。我们用耐抗的方法（这种方法不受少数点很大地偏离其余点的模式造成的严重影响）拟合一条过原点的直线。一种快速的方法是考虑 7 条直线，每条通过原点和 7 个点中的一个点，然后取这 7 个斜率的中位数，以它作为斜率并通过原点的直线，就是结果。这个方法给出中位数斜

图 5 四次方根下废气排放量图



率是 0.2516, 在图 3 中给出了结果直线。关系式：幂指数 =1- 斜率指示幂变换重新表达的幂指数是 0.7484, 我们将这个指数舍入到最接近的半整数。这里 p=0.5 是最接近的。因此我们对它进行平方根变换。

由图 4 看出，平方根变换后还不够对称，于是我们用四次方根变换，变换后图像如图 5 所示对称效果相当好。事实上，由表 2 和表 3 中不同尺度下中总括的变化对比也可以很容易看出四次方根尺度下的中总括变化最小意味着该变换后数据最对称。通过以上分析，我们决定对四次方根尺度下的数据进行进一步分析。

§3.4 数据变换后的线性拟合

对废气排放单批数据进行变换分析之后，我们接下来考虑废气排放量与燃料煤消费量这二个批数据间的关系。由图 6 可以看出，未经变换的燃料煤消费两同废气排放量二者之间的关系不明显，基于以上的分析，我们决定对这两批数据作四次方根变换，对尺度变换后的数据进行分析。由图 7 可以看出，变换后的两个批数据具有明显的线性性，于是我们决定对尺度变换后的数据

表 2 废气排放量样本的带中总括的字母值显示（一）

#	106	原尺度下废气排放量			平方根尺度下废气排放量		
M	53.5		1308.5			36.1723	
F	27	603	1295.5	1988	24.5561	34.57155	44.587
E	14	408	1869.5	3331	20.199	38.9569	57.7148
D	7.5	283	3234	6185	16.8217	47.7195	78.6173
C	4	181	3582.5	6984	13.4536	48.51195	83.5703
B	2.5	111.5	4543.5	8975.5	10.3867	52.54785	94.709
A	1.5	39	5754.5	11470	5.4674	56.0674	106.6674
Z	1	6	6758.5	13511	2.4495	59.34315	116.2368

表 3 废气排放量样本的带中总括的字母值显示（二）

#	106	四次方根尺度下废气排放量			对数尺度下废气排放量		
M	53.5		6.0143			7.1765	
F	27	4.9554	5.81635	6.6773	6.4019	6.9984	7.5949
E	14	4.4943	6.04565	7.597	6.0113	7.06115	8.111
D	7.5	4.1014	6.48365	8.8659	5.6452	7.18685	8.7285
C	4	3.6679	6.4048	9.1417	5.1985	7.02495	8.8514
B	2.5	3.2092	6.47015	9.7311	4.647	6.874	9.101
A	1.5	2.239	6.2783	10.3176	3.0342	6.18275	9.3313
Z	1	1.5651	6.1732	10.7813	1.7918	5.65155	9.5113

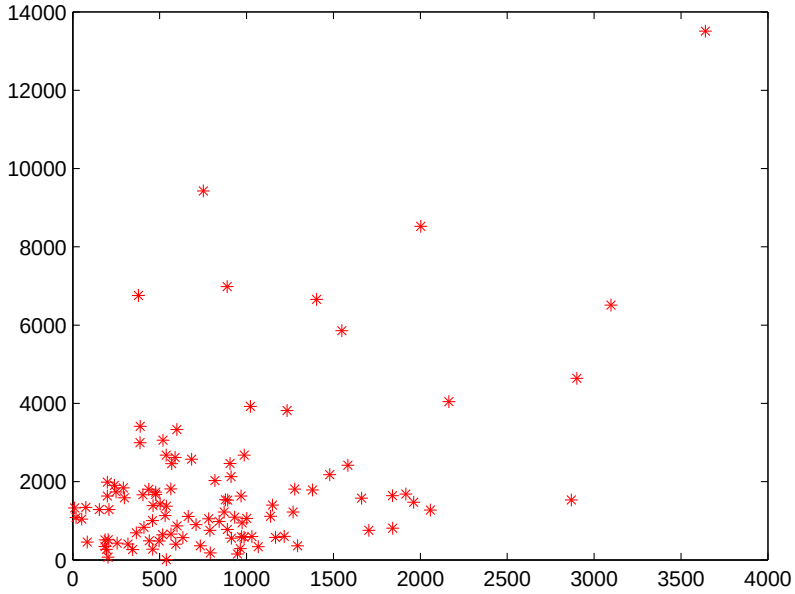
作线性回归。

注：之后所说的燃料煤消费量和废气排放总量若无特别说明，都是对于四次方根变换数据后而言的。

§3.4.1 最小二乘拟合

对于数据 $(x_i, y_i),\ i = 1, ..., n,$ 拟合一条直线已经发展了各种方法。其中最著名且最广泛使用的是最小二乘回归。于是我们先用最小二乘法对四次方

图 6 原料煤消费量对废气排放量图



根尺度下的数据进行拟合。得到：

$$y = 1.1228 * x + 0.2963 \tag{4}$$

拟合后的图象如图 7 所示。

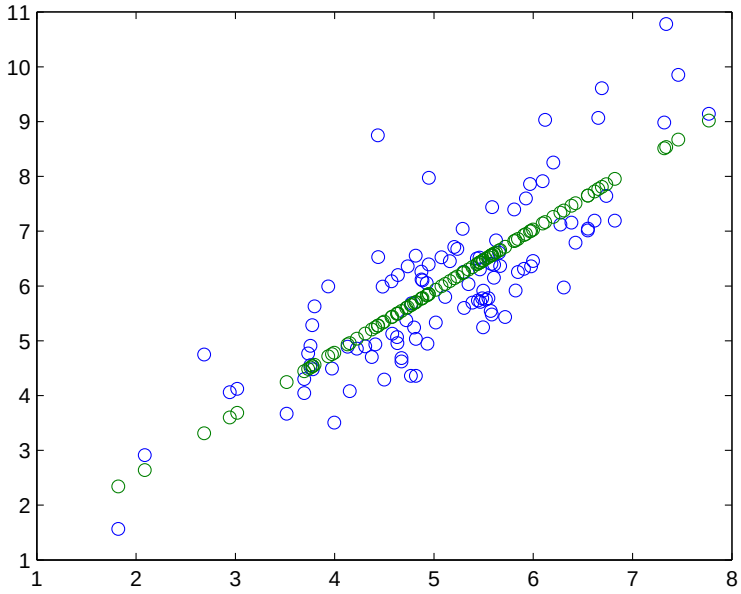
§3.4.2 最小二乘残差分析

残差分析在数据分析过程中尤为重要。拟合体现数据中的某些主要模式，要更详细的查勘数据，我们考察残差。最小二乘拟合后的残差如图 8 所示

§3.4.3 三组耐抗线拟合

由于最小二乘回归线不提供耐抗性。一个也数据点就可以支配控制拟合线，使它给出一个完全引人歧途的 y 和 x 之间关系的概括公式。三组耐抗线避免了这个困难，在探索 y 对 x 数据时更有用。于是我们用迭代的方法得到尺度变换后废气排放量对燃料煤消费量的三组耐抗线。拟合的初始曲线

图 7 四次方根变换后的燃料煤消费量对废气排放量最小二乘拟合



为：

$$y = 6.0110 * x + 1.1980 \tag{5}$$

拟合结果如图 9

经过近一步迭代的拟合

$$y = -0.0751 * x - 0.1367 \tag{6}$$

$$y = -0.0050 * x + 0.0428 \tag{7}$$

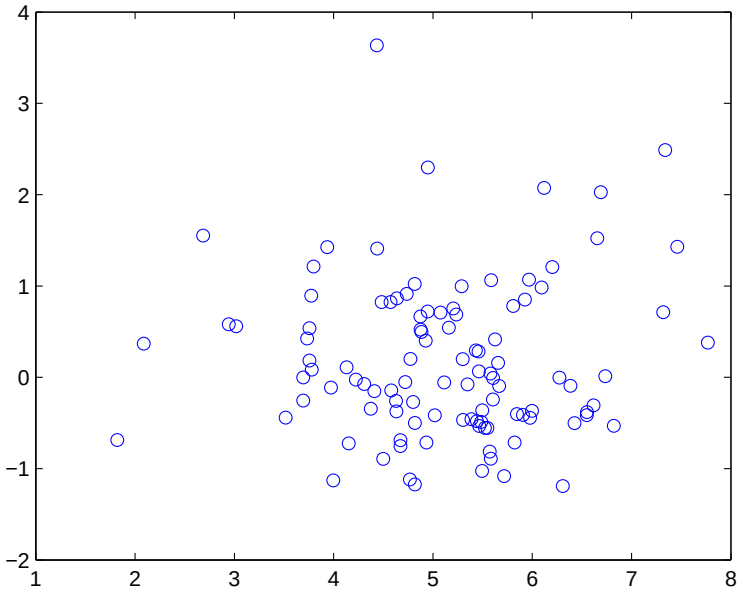
$$y = 0.0025 * x - 0.0120 \tag{8}$$

$$y = -0.0014 * x + 0.0024 \tag{9}$$

第五次拟合的斜率小于 b_0 的 1%，迭代结束。迭代过程中没有出现剧烈的震荡情况。图 10 为四次方根变换后的废弃排放量对燃料煤消费量的耐抗线迭代拟合的第五步残差图。最终的拟合直线为图 11

$$\hat{y} = 5.9320 + (x - 5.1826) * 1.0945 \tag{10}$$

图 8 四次方根变换后的燃料煤消费量对废气排放量最小二乘拟合残差图



§3.4.4 杆杠率

我们通过计算杠杠率的方法考察两种拟合方式。对于简单的线性回归，第 i 个点的杠杠率是：

$$h_i = \frac{1}{n} + \frac{(x_i - barx)^2}{\sum_{j=1}^{n}(x_j - \overline{(x)})^2} \tag{11}$$

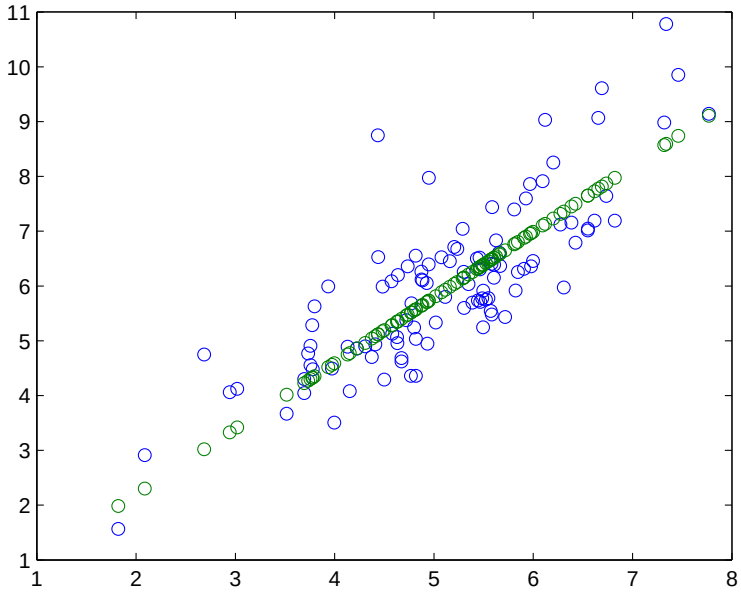
其中我们把 h_{ii} 缩写成 h_i。第 i 个残差的方差是：

$$var(r_i) = \sigma^2(1 - h_i) \tag{12}$$

因此，一个模式通过数据点的各自的杠杠率引进到残差中。当 x_i 远离 $\bar{x}$，杆杠率就大，残差 r_i 的方差就小。因此，当模型成立时，远离 $\bar{x}$ 的点倾向于被拟合得更靠近回归线。我们作出杆杠率对燃料煤的图（图 12）

由图 12 可见极小的四个点和极大的四个点的杠杆率很大，用这样的点对数据进行的最小二乘拟合会有较大的误差，对整体拟合产生较大的扰乱。因此，我们用三组耐抗线迭代的方法拟合的结果对残差进行分析效果更好。

图 9 四次方根变换后的废弃排放量对燃料煤消费量的耐抗线初步拟合图



§3.5 三组耐抗线残差分析

§3.5.1 各种残差说明

由数据点 x 值所引起的残差模式，当我们把每个残差除以它的标准差的一个倍数，就被去掉。然后，这后一种残差就有零均值和等方差。第 i 个调整残差是：

$$r_{ai} = \frac{r_i}{\sqrt{1-h_i}} \tag{13}$$

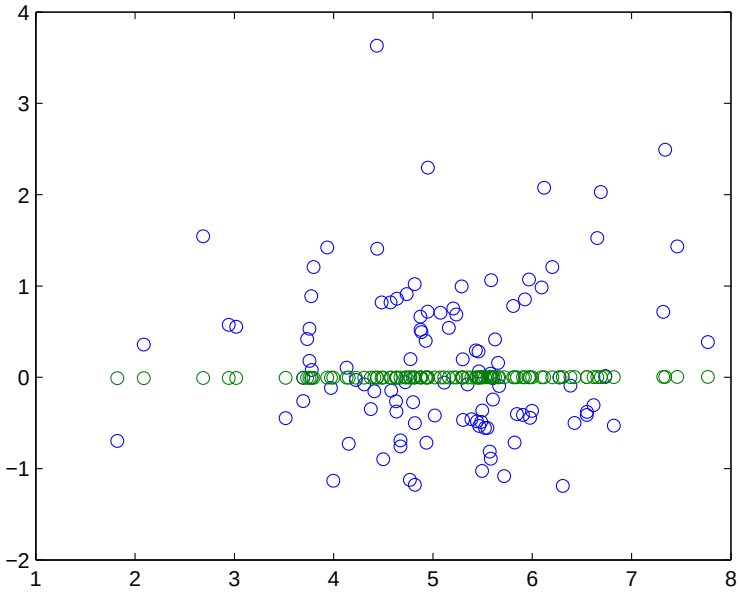
我们可以用残差平方和的适当倍数估计 σ^2，

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} r_i^2 \tag{14}$$

第 i 个标准化残差是：

$$r_{si} = \frac{r_i}{s\sqrt{1-h_i}} \tag{15}$$

图 10 迭代拟合第五步残差图



换种办法，我们可以用去掉第 i 个点后其余所有的点的拟合回归线的残差方差，估计第 i 个残差的方差。我们以 $s^2(i)$ 记这个残差的方差。第 i 个学生化残差是：

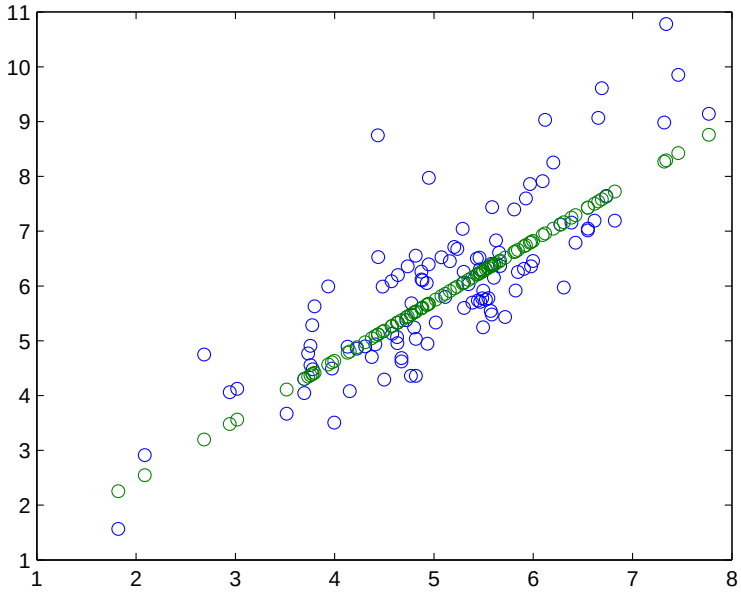
$$r_i^* = \frac{r_i}{s(i)\sqrt{1-h_i}} \tag{16}$$

避免了当 σ^2 在第 i 点，跟 σ^2 在其它点，很不一样的困难。

§3.5.2 学生化残差

图 13 为三组耐抗线拟合及残差对比图。在图 14 中显示了四次方根尺度下废气排放总量对燃料煤消费量的三组耐抗回归残差的散点图，其中横坐标是燃料煤消费量。图 12(杠杆率图) 表明左边的点杠杆率大，这也是为什么图 14 中左边的点的方差比其余的小（x 的值是左偏的）。在图 15（学生化残差图）中，我们按点的杆杠率调整残差。可以看出，图 14 和图 15 几乎相同。

图 11 第五次对残差迭代拟合图



§3.5.3 残差的批的显示

$stem(shujufenxi)$ *Removed*

$N = 106$ $Median = -0.0020325$ $Quartiles = -0.4411,\ 0.7132$

Decimal point is at the colon

-1 : 221110

-0 : 9988777776655555555

-0 : 44444444443333322111111110000

0 : 0011122223344444

0 : 5555667777788889999

1 : 0001122444

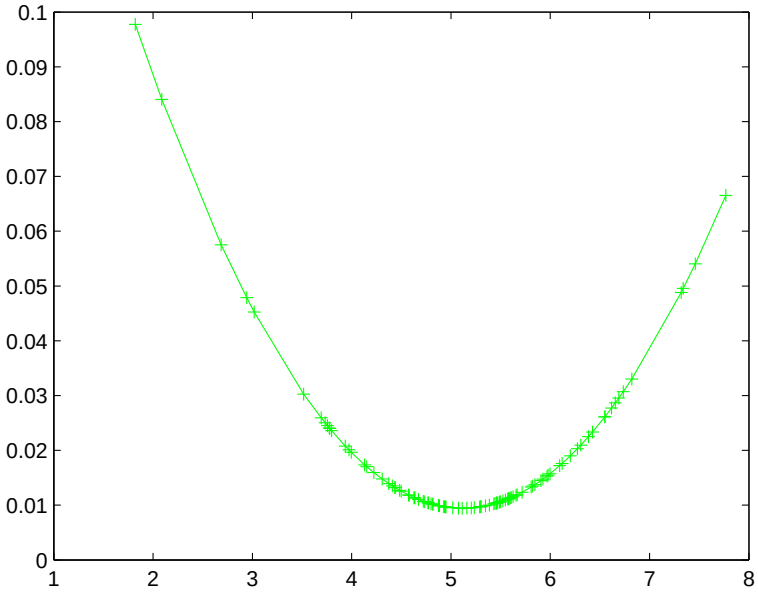
1 : 56

2 : 013

2 : 5

$High$: 3.6344

图 12 杠杠率对燃料煤消费量图



考察一批残差时，我们希望看见分布的偏度、极端的值对中心部分的关系、离群值以及值得注意的成组。有两种显示技术揭示整个分布，它们是箱线图 (图 16) 和茎叶图。这两个图都显示有较大正残差，且由这两个图都可以看出这个图是右偏的，残差集中在小于零的部分。

§3.5.4 分位数图

我们用分位数——分位数图看残差和某个理想分布的靠近程度。通常我们把残差跟高斯分布作比较。这个图就叫做正态概率图。我们把残差排序，然后把从批容量为 n 的批中的第 i 个排过序残差，对相应的高斯分位数，绘出点来. 以 Φ 记标准高斯分布函数，第 i 个分位数方便地取作

$$\phi(\frac{i-\frac{1}{3}}{n+\frac{1}{3}}) \tag{17}$$

残差分布的尾部形状由这个图的尾部来指示。图 §3.5.4 是由学生化残差得到的正态概率图。图 §3.5.4 略微有一点轻尾，当总的来说是合理地直线性的，启示残差的分布是近似高斯的。

表 4 各百分位切尾均值表

估计量	切尾均值
均值	1.8667
T(5%)	1.6387
T(10%)	1.6311
T(20%)	1.6406
中均值	1.5903
T(30%)	1.3449
BMED(37.5%)	1.4951
T(40%)	1.5021
中位数 (45%)	1.3085
三均值	1.302

§3.5.5 切尾均值

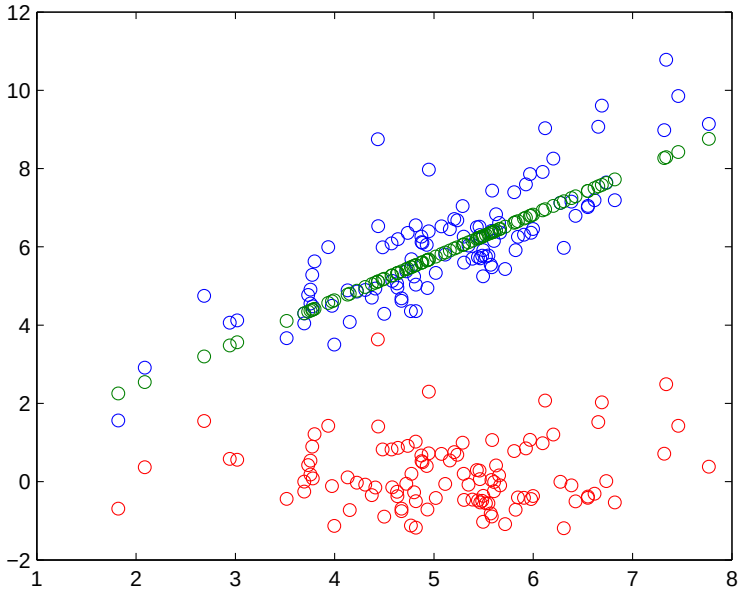
估计我国废气排放的总体状况是很必要的. 我们从稳健性的角度出发, 基于数据的分布特征, 寻求最优切尾效率时的位置估计量, 来预测我国废气排放的平均状况. 各百分位的切尾均值如表 §3.4.4 所示。图 18 为切尾均值对百分率点图。

切尾均值对百分位的点图: 由于数据为非正态分布组, 故最优切尾效率为 0.23, 可以求得此时的切尾均值为 1353.4. 此值即为我国各大城市平均排放废气量的最优位置估计量

§4 总结

通过对分析废气排放的分布的分析, 我们在一定程度上揭示了我国重点城市废气排放的现状：除了那 12 所城市废气排放量特别高外, 大部分还是积聚分布在一个范围之内。基于人力财力有限, 国家对每年对重点城市废气排放量进行普查是不现实的, 通过以上分析, 我们可以做出如下建议：采取分层抽样的方式, 将 12 个较大离群值作为一层, 全部抽查, 对于其余的数

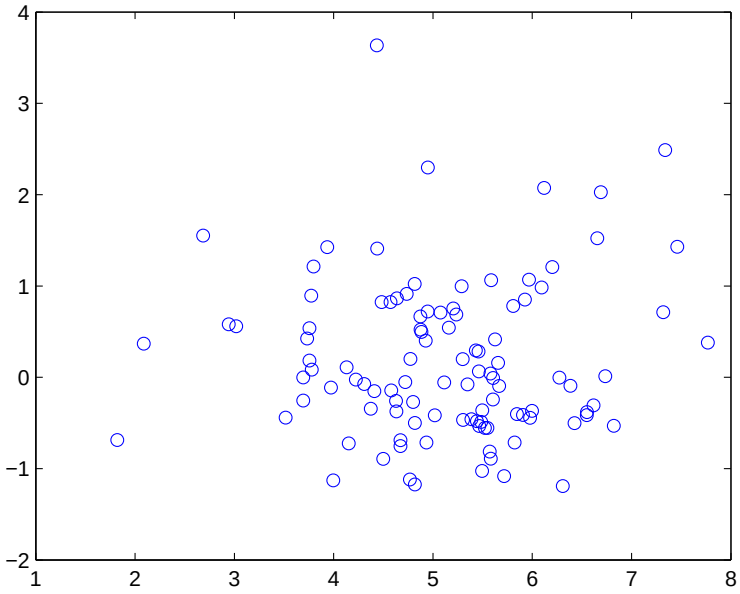
图 13 三组耐抗拟合结果及残差图



据再分层，然后在各层内抽取一些作为代表估计全国废气排放量的情况。

在结合燃料煤的消费量，对其潜在关系进行分析的过程中我们发现，废气排放量和燃料煤消费量在四次方根尺度变换下呈现较为明显的先性性。于是我们用最小二乘和三组耐抗线两种主要线性拟合的方法对尺度变换后的两批新数据进行了拟合，并且做了残差分析。最后综合考虑耐抗性，我们决定采用三组耐抗线的方法拟合。考虑这二者的关系是具有重要意义的，它可以提供了一种通过煤的消费量判断该城市废气排放总量的数据是否可信。若是煤的消费量很高而上报的废气排放量很低，则可能在数据统计的某个环节出现了错误。

图 14 三组耐抗拟合残差对燃料煤消费量图



§5 程序附录

MATLAB 程序

% 求取分位数

$function\ fenwei = f(x, n)$

$stotal1 = sort(x);$

$f1 = stotal1;$

$stotal2 = sort(x,' descend');$

$l = length(stotal1);$

$m = median(stotal1);$

$a = [m, m, (l + 1)/2];$

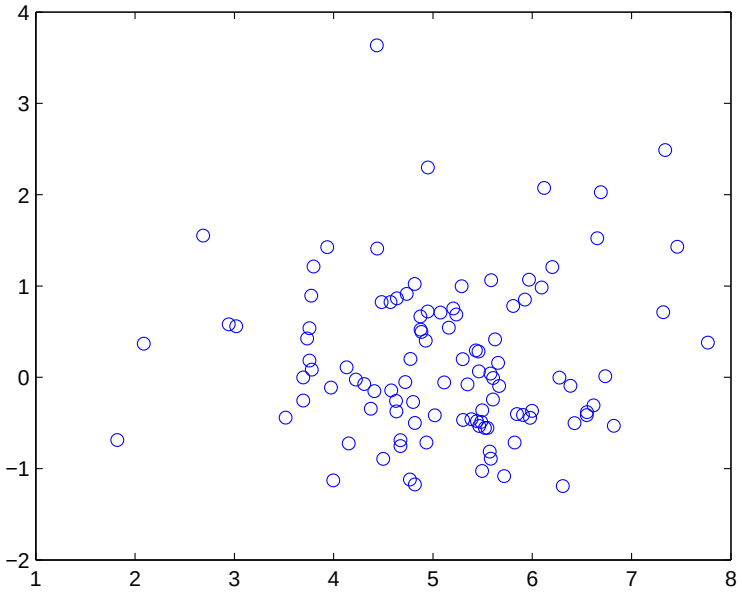
$fori = 2 : n;$

$f = find(f1 <= m);$

$f1 = stotal1(f);$

$f2 = stotal2(f);$

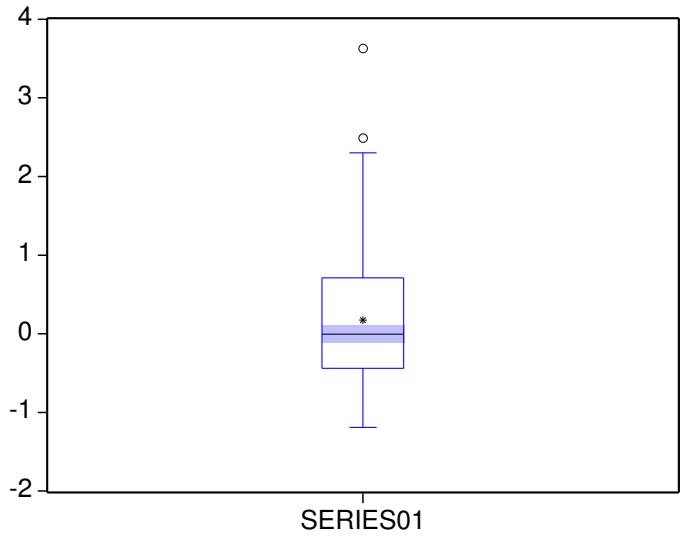
图 15 三组耐抗拟合学生化残差对燃料煤消费量图



$l1 = length(f1);$
$m = median(f1);$
$m2 = median(f2);$
$a = [a; m, m2, (l1 + 1)/2];$
$endfenwei = a;$

$total = [4641, 6512, 8523, 13511, 1476, 6656, 1535, 2994, 2178, 872,$
$1786, 1226, 1273, 3818, 945, 1115, 1683, 4044, 1578, 5858, 810, 1053,$
$1643, 1062, 758, 361, 9428, 3921, 2678, 3414, 1112, 6984, 1223, 361, 901,$
$1635, 3331, 2678, 599, 1812, 1431, 603, 780, 1816, 756, 576, 1387, 580, 659,$
$2462, 1804, 2621, 2028, 1907, 1665, 2421, 1401, 556, 2573, 151, 2466, 1094,$
$1536, 1530, 982, 3060, 1290, 181, 289, 593, 339, 482, 572, 72, 2126, 1132, 489,$
$1004, 1373, 277, 403, 6757, 1633, 268, 1844, 408, 272, 1044, 692, 1673, 1732, 1988,$
$1738, 517, 6, 642, 509, 835, 456, 1587, 1342, 342, 1287, 430, 1083, 1327];$

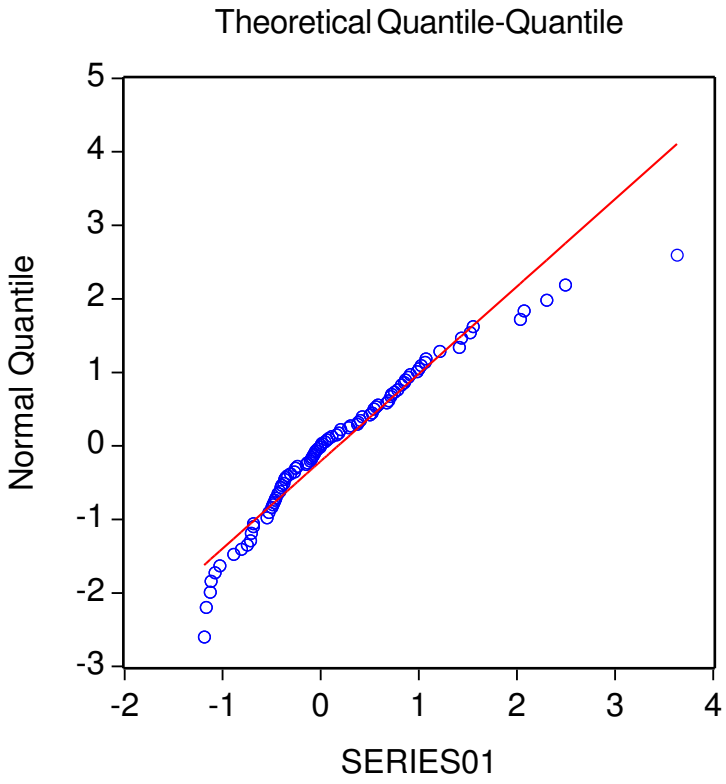
图 16 学生化残差箱线图




```
% 对称性变换图
f = fenwei(total, 9); % 自动生成分位数函数
m = median(total);
x = []; y = [];
for i = 2 : 9
y(i − 1) = (f(i, 1) + f(i, 2))/2 − m;
x(i − 1) = ((f(i, 1) − m)^2 + (f(i, 2) − m)^2)/4/m;
end
plot(x, y, ′r∗′)
holdon
k = y./x;
k = median(k);
plot(x, k ∗ x, ′r′)

% 对数据进行变换
stotal2 = sqrt(stotal);
plot(stotal2, ′r∗′)
```

图 17 学生化残差的正态概率图



$stotal3 = sqrt(stotal2);$

$plot(stotal3,'r*')$

$x = 1:106;$

$polyfit(x, stotal3, 1)$

$plot(x, 0.0456 * x + 3.6022,'r', x, stotal3,'r*')$

% 求杠杆率

$data = Sheet1(1:106);$

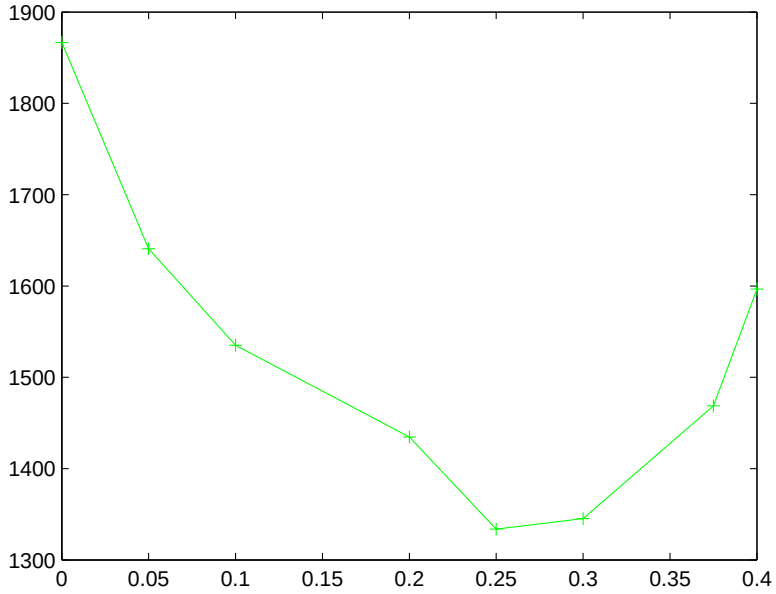
$x = data;$

$n = length(x);$

$m = mean(x);$

$for(i = 1:n)$

图 18 切尾均值对百分率点图



$for(j = 1 : n)$
$h(i, j) = 1/n + ((x(i) - m) * (x(j) - m))/sum((x - m).^2);$
end
end
$for(i = 1 : n)$
$g(i) = h(i, i);$
end
$plot(x, g,' g + -');$

% 最小二乘拟合
$a = polyfit(data1, data2, 1);$
$y1 = 1.1228 * x + 0.2963;$
$m = y - y1;$
$plot(x, y,' 0', x, y1,' o-')$

% 三组耐抗线拟合

$data1 = Sheet1(1:106);$

$data2 = Sheet1(107:212);$

$x = data1;$

$y = data2;$

$x1 = data1(1:35);$

$x2 = data1(36:71);$

$x3 = data1(72:106);$

$y1 = data2(1:35);$

$y2 = data2(36:71);$

$y3 = data2(72:106);$

$b = (median(y1) - median(y3))/(median(x1) - median(x3));$

$a = 1/3 * (median(y1) - b * (median(x1) - median(x2)) + (median(y2)) +$
$median(y3) - b * (median(x3) - median(x2)));$

$m = a + (data1 - median(data1)) * b; r = data2 - m;$

$a\ b\ plot(x, y,' o', x, m,' o')$

% 求切尾均值

$function\ qiewei = qw(data)$

$N = length(data)$

$alfa = [0, 0.05, 0.1, 0.2, 0.25, 0.3, 0.375, 0.4];$

$g = floor(alfa * N);$

$r = alfa * N - g;$

$for(i = 1:8)$

$c(i) = 1/(N * (1 - alfa(i) * 2)) * ((1 - r(i)) * (data(g(i) + 1) + data(N -$
$g(i))) + sum(data((g(i) + 2) : (N - g(i) - 1))))$

end

$plot(alfa, c,' o-')$