颜色与物质浓度的辨识问题

摘要

本文是对颜色与物质浓度的辨识问题的研究，通过对溶液色度值与待测物浓度的实验数据进行多元回归分析，建立了线性和非线性回归方程模型，给出了数据的评价准则和模型的误差分析。

问题一：首先依据对数据初步分析，发现物质浓度与颜色读数存在着一定的关系。利用 MATLAB 统计工具箱中的 Regress 函数求出回归系数和置信区间，并进行残差分析，最终建立关于颜色读数和物质浓度的多元线性回归模型。基于对模型的检验分析的基础上，给出了判别数据优劣的五大准则，分别是评估模型是否成功的四个要素，F 检验、相关系数 R^2、P 值、估计误差方差 S^2；再加上数据完整性要素，即模型拟合过程中是否存在异常数据剔除。根据判别准则，数据优劣的排序为：组胺>溴酸钾>奶中尿素>硫酸铝钾>工业碱。

问题二：首先建立二氧化硫浓度与颜色读数之间的线性回归模型，模型的残差较大，拟合效果不佳。考虑建立非线性二次回归模型，利用 MATLAB 统计工具箱中的 rstool 函数建模，通过剩余标准差和残差评估模型优劣。最终建立的非线性二次回归模型中，剩余标准差很小，预测模型非常好，模型的残差相比五元线性回归模型降低了一个数量级，因此线性二次回归模型比线性回归模型更优。

问题三：首先降低多元线性回归模型中颜色的维度来分析颜色维度对模型的影响；然后再通过减少数据量来分析数据量对模型的影响。通过分析发现：数据量不能低于 6,一般在 10-15 之间；颜色纬度可以降低，二纬和三纬都可以，一纬模型就不太优甚至不成立了，而且颜色维度的大小比数据量的多少对模型的影响更大；于是最后使用层次分析法对数据量的多少和颜色维度的大小对模型的影响因子进行分析求解，得出了影响因子分别为 0.414 和 0.586。

关键词： 多元线性回归，多元非线性二次回归，MATLAB，误差，层次分析法

问题重述

（一）问题的背景：

比色法是目前常用的一种检测物质浓度的方法，即把待测物质制备成溶液后滴在特定的白色试纸表面，等其充分反应以后获得一张有颜色的试纸，再把该颜色试纸与一个标准比色卡进行对比，就可以确定待测物质的浓度档位了。由于每个人对颜色的敏感差异和观测误差，使得这一方法在精度上受到很大影响。随着照相技术和颜色分辨率的提高，希望建立颜色读数和物质浓度的数量关系，即只要输入照片中的颜色读数就能够获得待测物质的浓度。试根据附件所提供的有关颜色读数和物质浓度数据完成下列问题：

（二）问题的提出：

1.附件 Data1.xls 中分别给出了 5 种物质在不同浓度下的颜色读数，讨论从这 5 组数据中能否确定颜色读数和物质浓度之间的关系，并给出一些准则来评价这 5 组数据的优劣。

2.对附件 Data2.xls 中的数据，建立颜色读数和物质浓度的数学模型，并给出模型的误差分析。

3.探讨数据量和颜色维度对模型的影响。

二、模型假设与符号说明

1．模型假设

1）反应应具有较高的灵敏度和选择性；

2）反应生成的有色化合物的组成恒定且较稳定；

3）选择适当的显色反应和控制好适宜的反应条件。

2．符号说明：见表 1

符号	含义	单位
Y_i	各待测物理论浓度 (i=1,2,3,4,5,6)	$ppm(mg/L)$
$\hat{Y}_i$	各待测物实际浓度 (i=1,2,3,4,5,6)	$ppm(mg/L)$
C	回归方程回归系数	
r	残差	
R^2	相关系数	
F	F 值	
P	与 F 对应的概率	
S^2	估计误差方差	

B	蓝色颜色值	
G	绿色颜色值	
R	红色颜色值	
H	色调	
S	饱和度	
m	数据量	

表 1 符号说明

三、问题的分析

首先对 Data1.xls 和 Data2.xls 提供的数据利用 MATLAB 进行相关性分析发现：颜色读数（五个维度：B,G,R,H,S）对物质浓度呈现一定线性相关性，而这一结论与文献[1]使用朗博-比尔吸收定律得到的结论一致。即物质浓度和颜色读数之间存在一定的关系。

其次利用统计学中的多元回归[2]对给出的六组数据进行回归分析，从而得出物质浓度与颜色读数（五维）之间的相互关系，确定它们之间合适的数学表达式（或数学模型）即经验公式或回归方程。

针对问题一

基于对 Data1.xls 的数据分析，我们可以利用 MATLAB 统计工具箱中的 Regress 函数求出回归系数和置信区间，绘出残差图并进行残差分析，剔除置信区间不包含零点的异常点数据，重新进行多元线性回归，能够更好地建立关于颜色读数和物质浓度的多元线性回归模型。

基于对模型的检验分析的基础上，可以考虑评估模型是否成功的四个要素，F 检验、相关系数 R^2、P 值、估计误差方差 S^2，相关系数 R^2 越接近 1，说明回归方程越显著；$F > F_{1-\alpha}(k, n-k-1)$ 时拒绝 H_0，F 越大，说明回归方程越显著；与 F 对应的概率 $P < \alpha$ 时拒绝 H_0，回归模型成立。估计误差方差越小，回归方程越显著。还可以再考虑数据完整性要素，即模型拟合过程中是否存在异常数据剔除。给出对应评价 Data1.xls 中 5 组数据优劣的五大准则，并根据 5 组数据是否同时满足五大准则对其优劣进行判别。

针对问题二

问题二是在问题一的基础上，进一步确定颜色读数和物质浓度的数学模型-线性回归方程。首先建立二氧化硫浓度与颜色读数之间的线性回归模型，模型的残差较大，拟合效果不佳。

考虑建立非线性二次回归模型，利用 MATLAB 统计工具箱中的 rstool 函数建

模，通过剩余标准差和残差评估模型优劣。最终建立的非线性二次回归模型中，剩余标准差很小，预测模型非常好，模型的残差相比多元线性回归模型降低了一个数量级，因此线性二次回归模型比线性回归模型更优。通过两种模型的误差的对比发现：非线性回归二次方程的精度更高。

针对问题三

问题三是讨论数据量和颜色维度对模型的影响。根据问题一和问题二的求解结果发现：数据量的大小会影响模型的优劣；以及通过枚举法调整线性回归中变量的数量即颜色维度发现：颜色维度的多少也会影响模型的优劣。而且数据量对模型优劣的影响度大于颜色维度对模型优劣的影响度。因此本文提出采用层次分析法对两者的影响因子进行分析，最终得出了数据量和颜色维度对模型优劣的影响因子。

四、模型的建立与求解

问题一：

基于对数据的分析，本文认为有 Data1.xls 提供的 5 组数据能确定颜色读数与物质浓度之间的关系，并建立了多元线性回归模型：

$$Y=\varepsilon+C_1R+C_2G+C_3B+C_4H+C_5S \qquad (\text{Ⅰ})$$

（Ⅰ）式中C_1, C_2, C_3, C_4, C_5表示方程的回归系数。

利用 matlab 统计工具箱建立多元线性回归方程：

[b, bint,r,rint,stats]=regress(Y,X,alpha)（Ⅱ）

式（Ⅱ）中 b 为回归系数，bint 为回归系数的置信区间，r 为残差，rint 为残差的置信区间，alpha 为显著性水平。stats 包含四个统计量，相关系数R^2、F 值、与 F 对应的概率 p，估计误差方差。相关系数R^2越接近 1，说明回归方程越显著；$F > F_{1-\alpha}(k, n-k-1)$时拒绝$H_0$，F 越大，说明回归方程越显著；与 F 对应的概率$p < \alpha$时拒绝H_0，回归模型成立。估计误差方差越小，回归方程越显著。

1.组胺浓度与颜色读数之间的关系函数：

根据组胺的实验数据（见表 2），其中 0 表示待测物质浓度为零的情形，即水溶液，使用 matlab 对数据进行多元线性回归（代码见附录中程序 1），画出残差图（图 1）并给出具体的残差值（表 3）和其置信区间（表 4）。

浓度（ppm）	B	G		R	H	S
0	68	110		121	23	111
100	37	66		110	12	169
50	46	87		117	16	155

25	62	99		120	19	122
12.5	66	102		118	20	112
0	65	110		120	24	115
100	35	64		109	11	172
50	46	87		118	16	153
25	60	99		120	19	126
12.5	64	101		118	20	115

表 2 组胺的实验数据

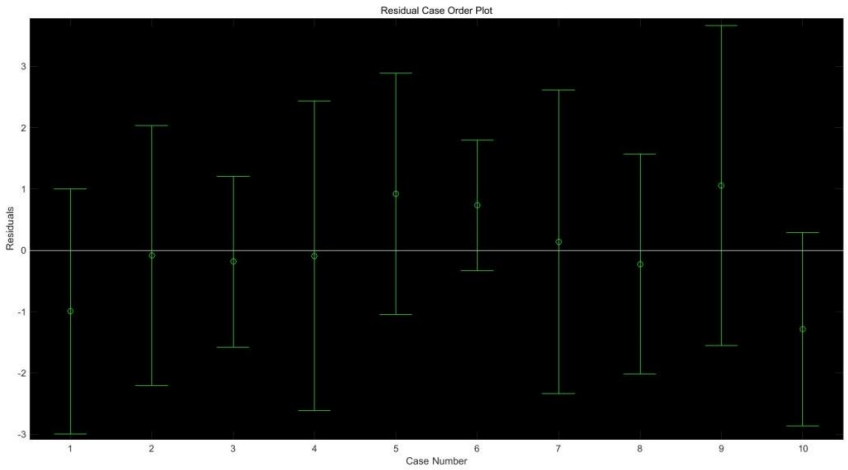

图 1 组胺浓度与颜色读数线性回归残差图

浓度（ppm）	残差值 r
0	-0.993129343227508
100	-0.083240562564029
50	-0.184054282892987
25	-0.087070619285910
12.5	0.920198815901770
0	0.733366635833562
100	0.141799543614084
50	-0.222352920091282
25	1.056506552856305
12.5	-1.282023820143920

表 3 组胺浓度与颜色读数线性回归残差值

相关系数 R^2	F 值	与 F 对应的概率 P	估计误差方差
0.999580110101	1904.461358300401	0.000000771022	1.312155935514

表 4

由表 4：相关系数 R^2=0.999580110101，说明回归方程非常显著。F 对应的概率 $p < \alpha$，拒绝 H_0，根据 F 检验，回归模型（Ⅲ）成立。

$$y = -212.7650200904682 + 2.8548267848162B - 4.4873189075604G$$
$$+2.3213368359433R + 4.5932448138408H + 1.1415190993725S \quad (III)$$

2. 溴酸钾浓度与颜色读数之间的关系函数：

根据溴酸钾的实验数据（见表 5），首先利用 matlab 统计工具箱建立多元线性回归方程（代码见附录中程序 1），画出残差图（图 2）。从残差图可以看出，除第十个数据外，其余数据的残差离零点均较近，且残差的置信区间均包含零点，这说明回归模型能较好的符合原始数据，而这个数据可视为异常点(剔除)。去掉异常点之后再次进行多元线性回归，绘出残差图（图 3），并给出具体的残差值（表 6）和其置信区间（表 7）。

浓度（ppm）	B	G	R	H	S
0	129	141	145	22	27
100	7	133	145	27	241
50	60	133	141	27	145
25	69	136	145	26	133
12.5	85	139	145	26	106
0	128	141	144	23	28
100	7	133	145	27	242
50	57	133	141	27	151
25	70	137	146	26	132
12.5	87	138	146	26	102

表 5 溴酸钾的实验数据

由表 7：相关系数 R^2=0．9985210281929，说明回归方程非常显著。F 对应

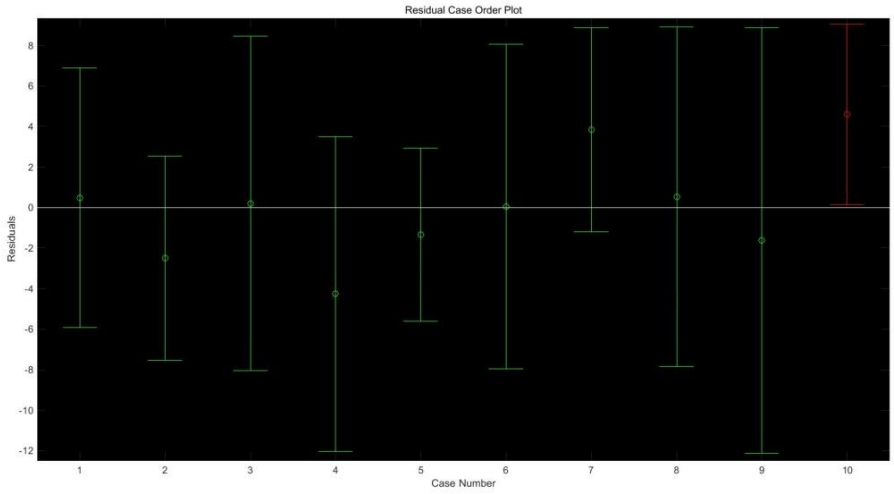

图 2 溴酸钾浓度与颜色读数线性回归残差图

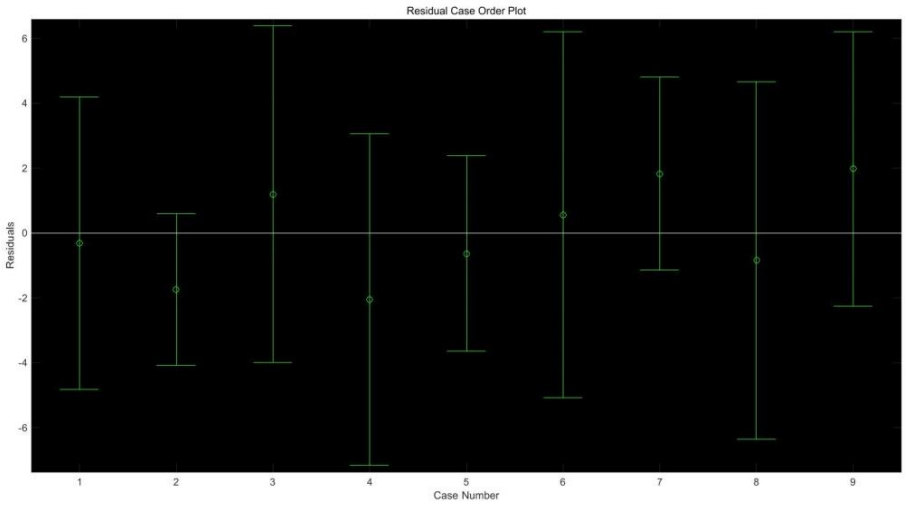

图 3 剔除异常点后溴酸钾浓度与颜色读数线性回归残差图

浓度（ppm）	残差值 r
0	-0.311054621479073
100	-1.739298695133584
50	1.198950353741111
25	-2.046358165775416
12.5	-0.633956401784815
0	0.563887602493494
100	1.832705601078601
50	-0.844912938120160
25	1.980037264978989

表 6

相关系数 R^2	F 值	与 F 对应的概率 P	估计误差方差
0．9985210281929	405．0872464611553	0.0001928592100	5.8200279444505

表 7

的概率 $p < \alpha$，拒绝 H_0，根据 F 检验，回归模型（Ⅳ）成立。

$$y = 1309.832425214150 - 7.825296356378B + 4.949407980110G - 4.722511350698R - 9.850745634837H - 3.572004296212S \quad （Ⅳ）$$

3．工业碱浓度与颜色读数之间的函数关系：

根据工业碱的实验数据（见表 8），使用 matlab 对数据进行多元线性回归（代码见附录中程序 1），画出残差图（图 4）并给出具体的残差值（表 9）和其置信区间（表 10）。

浓度（ppm）	B	G	R	H	S
7.34	153	140	132	108	35
8.14	151	142	133	104	29
8.74	158	126	127	120	52

9.19	161	85	118	132	120
10.18	127	21	119	147	211
11.8	94	6	91	148	237
0	152	142	132	105	32

表 8 工业碱的实验数据

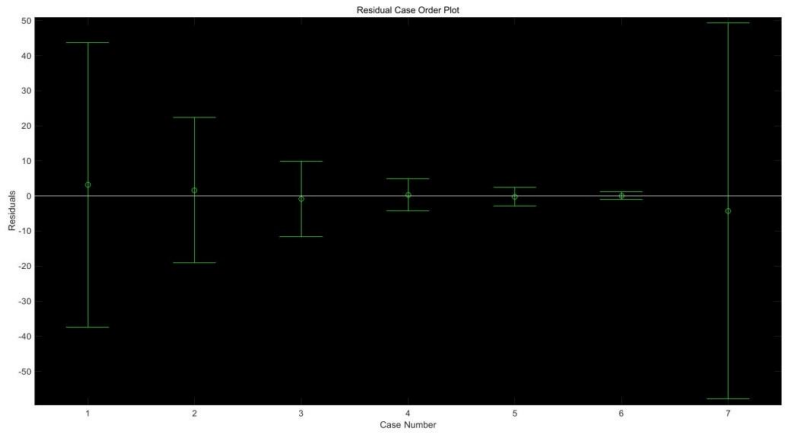

图 4 工业碱浓度与颜色读数线性回归残差图

浓度（ppm）	残差值 r
7.34	3.193187385650077
8.14	1.630046034029125
8.74	-0.845684659598861
9.19	0.362126613702886
10.18	-0.206897864646130
11.8	0.084838267818999
0	-4.217615776955910

表 9

相关系数 R^2	F 值	与 F 对应的概率 P	估计误差方差
0.631383991895078	0.342570033863204	0.851764320877558	31.538101080870671

表 10

由表 10：相关系数 R^2=0.631383991895078，说明回归方程不显著。根据 F 检验，F 对应的概率 $p > \alpha$，接受 H_0，回归模型（Ⅴ）不成立。

$$y = 261.0648697407135 + 0.1642129382467B - 1.3981694093973G - 0.3136416077960R - 0.1305810922948H - 0.8798705475876S \quad （Ⅴ）$$

4．硫酸铝钾浓度与颜色读数之间的函数关系：

根据硫酸铝钾的实验数据（见表 11），使用 matlab 对数据进行多元线性回归（代码见附录中程序 1），画出残差图（图 5）。

浓度（ppm）	B	G	R	H	S
0	116	126	104	76	44
0	114	126	104	74	45

0	118	125	105	78	40
0	113	124	103	73	42
0	114	124	104	75	39
0	113	126	104	72	45
0.5	148	112	47	100	174
0.5	150	111	44	100	178
0.5	138	118	71	98	123
0.5	136	118	70	98	122
0.5	136	117	64	98	134
0.5	136	118	64	97	135
0.5	138	111	50	99	161
1	149	116	48	99	172
1	150	115	49	100	171
1	147	119	55	99	159
1	149	119	64	100	145
1	140	113	54	99	156
1	137	111	51	99	160
1.5	153	113	44	101	180
1.5	153	113	42	100	184
1.5	153	115	50	101	171
1.5	153	115	47	100	176
1.5	152	116	52	100	167
1.5	153	116	49	100	171
2	156	106	34	102	199
2	162	107	37	103	196
2	161	110	40	102	190
2	163	111	38	102	194
2	159	104	35	103	198
2	158	105	35	103	198
5	155	107	34	101	198
5	156	108	34	101	198
5	152	116	48	100	174
5	151	115	51	100	168
5	154	105	33	102	199
5	156	105	35	102	197

表 11 硫酸铝钾的实验数据

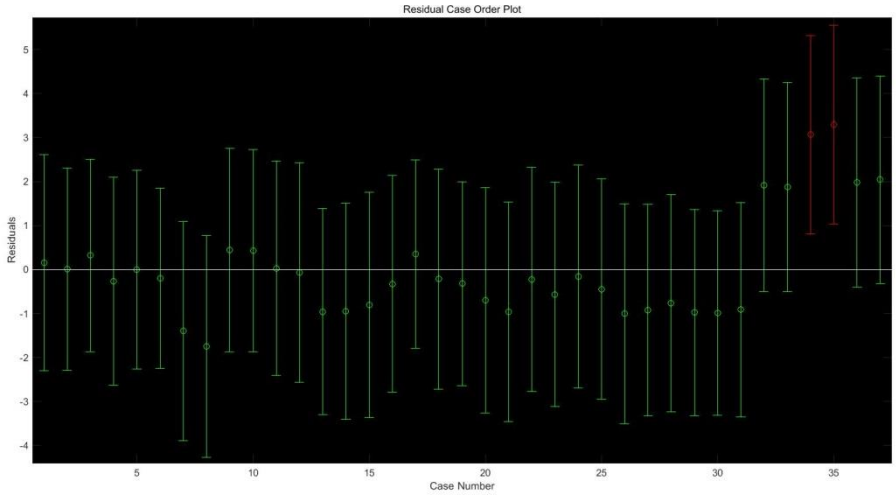

图 5 硫酸铝钾浓度与颜色读数线性回归残差图

从残差图可以看出，除第 34、35 个数据外，其余数据的残差离零点均较近，且残差的置信区间均包含零点，这两个数据可视为异常点(剔除)，剔除后重新进行多元线性回归。其残差图见图 6，其置信区间见表 10，具体的残差值见表 12。

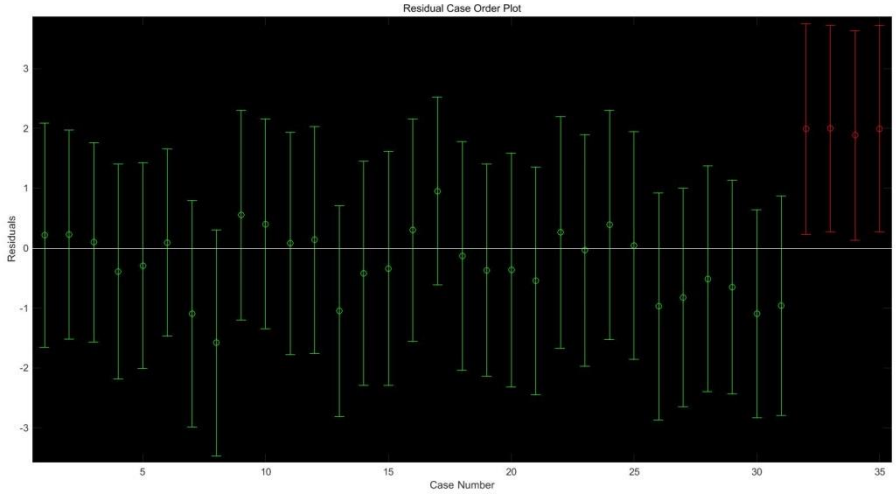

图 6 剔除异常点后硫酸铝钾浓度与颜色读数线性回归残差图

相关系数 R^2	F 值	与 F 对应的概率 P	估计误差方差
0.618028668332955	9.384385630950229	0.000021075916702	0.953423023427052

表 12

由表 12：相关系数 R^2=0.618028668332955，说明回归方程不够显著。根据 F 检验，F 对应的概率 $p < \alpha$，拒绝 H_0，回归模型（Ⅵ）成立。

$$y = 33.301599085809258 + 0.064396228061359B - 0.074826220081641G - 0.197724103131009R - 0.099368368917937H - 0.078317444150693S \quad (Ⅵ)$$

浓度（ppm）	残差值 r

0	0.217812495378475
0	0.226185657816012
0	0.097384883538230
0	-0.391115358786932
0	-0.294003078333490
0	0.091845148041501
0.5	-1.094599168574980
0.5	-1.578120377569594
0.5	0.550772518151472
0.5	0.403523426992495
0.5	0.082161917933112
0.5	0.135937213247507
0.5	-1.049785941526949
1	-0.417969670398088
1	-0.338417090642832
1	0.301243393927612
1	0.954892016792217
1	-0.129616765715801
1	-0.365983054485273
1.5	-0.365653364371072
1.5	-0.547200162948258
1.5	0.265486697222034
1.5	-0.035466760335465
1.5	0.387519206106351
1.5	0.043220445254729
2	-0.972466812655620
2	-0.826429615083248
2	-0.513655451206029
2	-0.649800116906356
2	-1.096532909104727
2	-0.957310460961725
5	1.989069822418745
5	1.999499814439030
5	1.883775320254449
5	1.993796182092362

表 13

5.奶中尿素浓度与颜色读数之间的函数关系：

根据奶中尿素的实验数据（见表 14），使用 matlab 对数据进行多元线性回归（代码见附录中程序 1），画出残差图（图 7）。

浓度（ppm）	B	G	R	H	S
0	118	136	139	25	37
500	117	137	139	27	41

1000	108	136	138	28	54
1500	110	136	139	26	52
2000	108	140	142	28	60
0	120	136	138	26	33
5	119	140	142	26	40
500	111	139	142	27	55
1500	107	136	139	26	58
2000	105	136	137	28	58
0	125	135	140	20	27
500	114	134	138	25	44
1000	112	132	134	27	42
1500	105	134	138	26	60
2000	107	135	138	26	57

表 14 奶中尿素的实验数据

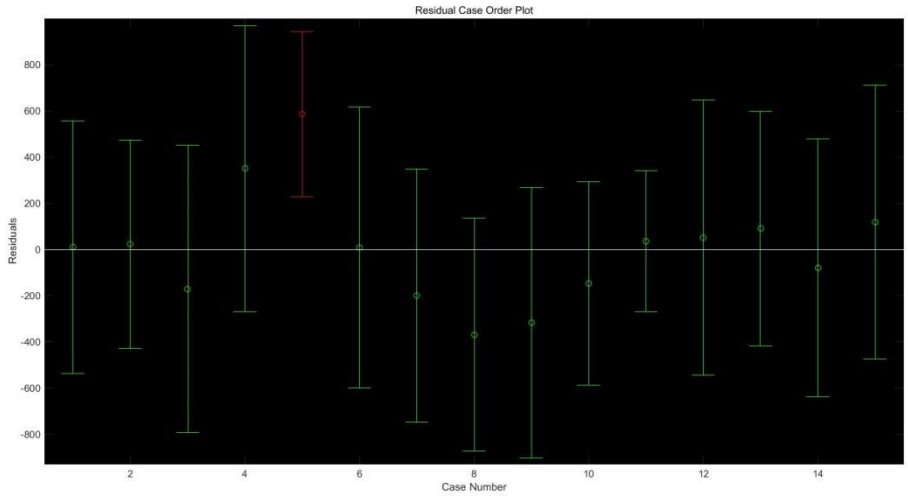

图 7 残差图

从残差图可以看出，除第 5 个数据外，其余数据的残差离零点均较近，且残差的置信区间均包含零点，这说明回归模型能较好的符合原始数据，而这个数据可视为异常点(剔除)，剔除后重新进行多元线性回归。其残差图见图 8，其置信区间见表 15，具体的残差值见表 14。

相关系数 R^2	F 值	与 F 对应的概率 P	估计误差方差
0.95791774884	36.42077968757	0.00002688234	37904.20690587087

表 15

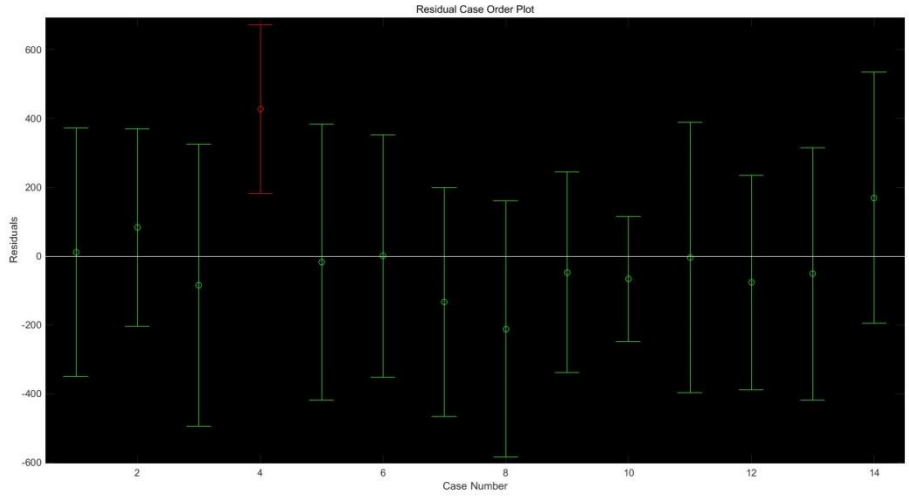

图 8 剔除异常点后奶中尿素浓度与颜色读数线性回归残差图

浓度（ppm）	残差
0	10.25536938636
500	83.881946665621
1000	-84.676233083668
1500	426.617345751831
0	-16.945348469715
5	0.421074517217
500	-133.810181334869
1500	-211.046368939387
2000	-46.846723118331
0	-65.449023536308
500	-4.712865495465
1000	-76.907322702158
1500	-51.308416740812
2000	169.256579546811

表 16

由表 16：相关系数R^2=0.95791774884，说明回归方程显著。根据 F 检验，F 对应的概率$p < \alpha$，拒绝H_0，回归模型（Ⅶ）成立。

$$y = 18784.81577777840 + 285.91798514465B + 454.96769353348G - 823.02411983538R - 369.46611698178H + 249.38672115051S \quad (Ⅶ)$$

6．评价 5 组数据的优劣

基于问题一的分析，我们首先给出评价 5 组数据优劣的五大准则：

准则一：能通过 F 检验；

准则二：相关系数R^2越接近于 1 越好；

准则三：$P < 0.05$且越接近于 0 越好；

准则四：误差方差越小越好；

准则五：是否剔除异常数据。

物质种类	R^2	F	P	S^2	模型检验	数据完整性
组胺	0.999580	1904.461358	0.0000007	1.312155	模型成立	完整
溴酸钾	0.998521	405.087246	0.000192	5.820027	模型成立	剔除 1 个
工业碱	0.631383	0.342570	0.851764	31.538101	模型不成立	完整
硫酸铝钾	0.618028	9.384385	0.000021	0.953423	模型成立	剔除 2 个
奶中尿素	0.957917	36.420779	0.000026	37904.206905	模型成立	剔除 1 个

表 17 五种物质线性回归方程的显著性检验指标

由表 17，五组数据中，只有工业碱的多元线性回归模型不成立，因此工业碱的数据是最差的。组胺的数据完整，多元线性回归模型的相关系数R^2、F 值最大，P 值最小，数据是最好的。硫酸铝钾的相关系数R^2比较小，模型不是很显著，而且剔除了两个数据，因此数据不够好。溴酸钾和奶中尿素都是剔除一个数据之后建立的模型，相关系数R^2很高，但溴酸钾的相关系数R^2更大一些，数据相对更优。

因此数据优劣的排序为：组胺>溴酸钾>奶中尿素>硫酸铝钾>工业碱。

问题二：

根据前面对问题二的分析，首先我们仍然建立与问题一一致的线性回归模型，利用 Data2.xls 提供的 25 组数据（即表 18），采用 matlab 进行线性回归（代码见附录中程序 2），得到了二氧化硫的浓度与颜色读数之间的线性回归方程。

2.1 多元线性回归模型

利用多元线性回归，绘出残差图（见图 9）.从残差图可以看出，除第 15 个数据外，其余数据的残差离零点均较近，且残差的置信区间均包含零点，这说明回归模型能较好的符合原始数据，而这个数据可视为异常点(剔除)，剔除后重新进行多元线性回归，得到残差图（见图 10），回归方程的显著性检验指标（见表 19）和具体残差值（见表 20）。

由表 19：相关系数R^2=0.9250310882931，说明回归方程较为显著。根据 F 检验，F 对应的概率$p < \alpha$，拒绝H_0，回归模型（Ⅷ）成立。但是估计误差方差偏大。

$$y = 2910.630153554265 + 3.587352490846x_1 - 21.155917919245x_2 + 4.796418968805x_3 - 6.750902382498x_4 - 10.532016102969x_5 \quad (Ⅷ)$$

浓度（ppm）	B	G	R	H	S
0	153	148	157	138	14
0	153	147	157	138	16
0	153	146	158	137	20
0	153	146	158	137	20
0	154	145	157	141	19
20	144	115	170	135	82
20	144	115	169	136	81
20	145	115	172	135	83
30	145	114	174	135	87
30	145	114	176	135	89
30	145	114	175	135	89
30	146	114	175	135	88
50	142	99	175	137	110
50	141	99	174	137	109
50	142	99	176	136	110
80	141	96	181	135	119
80	141	96	182	135	119
80	140	96	182	135	120
100	139	96	175	136	115
100	139	96	174	136	114
100	139	96	176	136	116
150	139	86	178	136	131
150	139	87	177	137	129
150	138	86	177	137	130
150	139	86	178	137	131

表 18 二氧化硫的实验数据

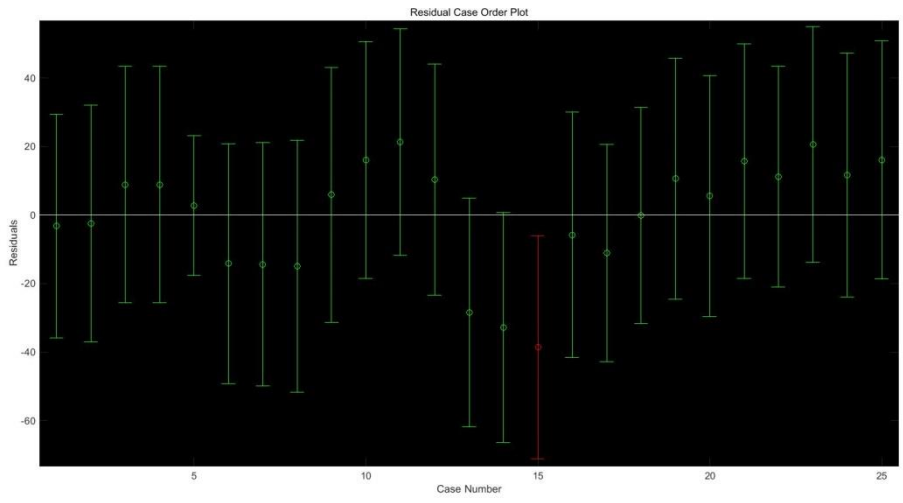

图 9 线性回归残差图

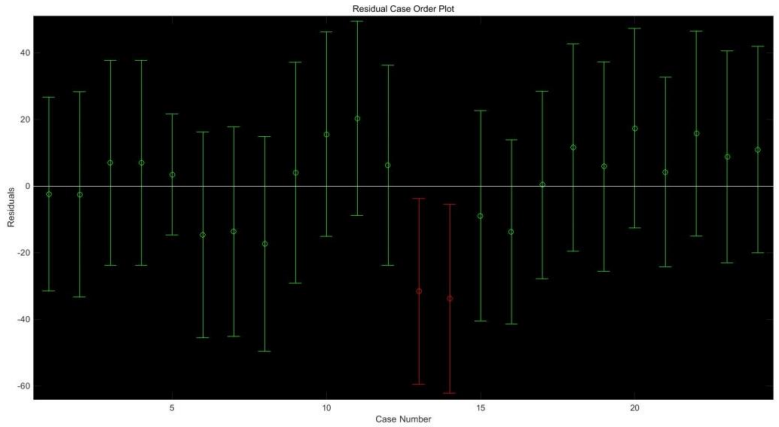

图 10 剔除异常点后二氧化硫浓度与颜色读数线性回归残差图

相关系数 R^2	F 值	与 F 对应的概率 P	估计误差方差
0.9250310882931	44.4199047583412	0.0000000016617	270.6516543935724

表 19 二氧化硫线性回归方程的显著性检验指标

浓度（ppm）	残差值
0	-2.384256481544441
0	-2.476142194851974
0	6.948682946475856
0	6.948682946475856
0	3.473424932212367
20	-14.672434139092047
20	-13.657128890758031
20	-17.320608464579323
30	4.058700090441562
30	15.529894358769411
30	20.326313327574439
30	6.206944733759315
50	-31.576255061221445
50	-33.724499704539312
80	-8.948829979216725
80	-13.745248948021299
80	0.374119645794281
100	11.627226785928087
100	5.891629651764106
100	17.362823920092069
150	4.191048334563675
150	15.830255399174575
150	8.793706073744261
150	10.941950717061900

表 20

通过残差值可以看出，模型还有待优化。可以通过剔除新的残差图中的异常点来继续优化多元线性回归模型，但续优化作用有限，而且数据完整性越来越差。该线性回归模型的结果需要进一步进行优化改进，于是尝试多元非线性二次回归。

2.2 多元二次回归模型

2.2.1 多元二次回归模型的建立与求解

利用 rstool（x，y，'model'，alpha）建立多元二次回归方程。其中'model'这一选项是指从下列 4 个模型中选择 1 个（用字符串输入，缺省时为线性模型）：

linear（线性）：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$

purequadratic（纯二次）：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{j=1}^{n} \beta_{jj} x_j^2$

interaction（交叉）：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \le j \ne k \le m} \beta_{jk} x_j x_k$

quadratic（完全二次）：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \le j,k \le m} \beta_{jk} x_j x_k$

rstool 函数输出包括回归参数，剩余标准差以及残差，可以通过修改 model 的值比较多个模型的标准差来确定哪个最好。

本题最终采用完全二次的方法进行多元非线性二次回归，即采用模型（Ⅸ）

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \\ &+\beta_6 x_1 x_2 + \beta_7 x_1 x_3 + \beta_8 x_1 x_4 + \beta_9 x_1 x_5 + \beta_{10} x_2 x_3 + \beta_{11} x_2 x_4 + \beta_{12} x_2 x_5 \\ &+\beta_{13} x_3 x_4 + \beta_{14} x_3 x_5 + \beta_{15} x_4 x_5 + \beta_{16} x_1^2 + + \beta_{17} x_2^2 + \beta_{18} x_3^2 + \beta_{19} x_4^2 + \beta_{20} x_5^2 \end{aligned} \qquad (Ⅸ)$$

模型（Ⅸ）中 y 代表浓度，x_1, x_2, x_3, x_4, x_5 分别代表 B, G, R, H, S 。

代入数据进行多元二次回归拟合（代码见附录中程序 3），具体结果见图 11、表 19 和模型（Ⅹ）。

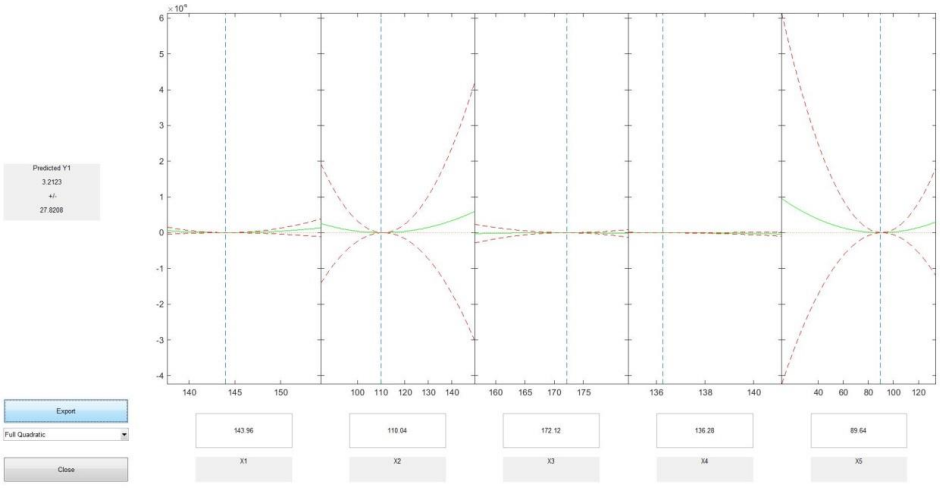

图 11

$$y = -229171.315611749 - 5684.26671497298B - 304.8236533092026$$
$$+4983.90599969629R + 4477.91841203602H - 1706.637003749365$$
$$-2.893107902333498G - 0.437552876691341BR + 15.71793037245668BH$$
$$+2.178457691376458BS - 5.355516884114918GR + 2.959910565388868GH \quad (X)$$
$$+4.387358454936638GS - 26.50954054086838RH - 2.694400813875188RS$$
$$+8.053737134345648HS + 12.97179440223578B^2 + 3.835465824505018G^2$$
$$-1.401816335831348R^2 - 11.91294881639798H^2 + 1.543033896961688S^2$$

浓度（ppm）	残差值 r
0	-0.183023306644486
0	0.295698988685444
0	-0.0573058051979842
0	-0.0573058051979842
0	-0.0112789762431476
20	0.483971591391310
20	0.0544367711663654
20	-0.617107404175840
30	0.438573762845408
30	0.223558673598745
30	-0.471712868260511
30	0.0136040522083931
50	0.551662284327904
50	-0.775350805535709
50	0.0956137145112734
80	0.0435429326025769
80	0.243119500199100
80	-0.357436173322640
100	-1.73816418466959
100	0.0231545045717212
100	1.60688363169902
150	0.0254071982417372
150	0.875543694299267
150	0.637870694485173
150	-1.34395668067009

表 21

2.2.2 多元二次回归模型的检验

本文采用剩余标准差对该二次回归模型的进行检验。回归残差$e_i = Y_i - Y_i$有助于衡量回归模型拟合样本数据的程度。应用线性回归分析，需要计算回归剩余标准差，而回归剩余标准差是表示回归方程用来预测的精度标志，可用来检验模型预测的可靠程度，回归剩余标准差（记作S_Y）：$S_Y = \sqrt{\frac{\sum (Y_i - Y_i)^2}{n-2}} = \sqrt{\frac{Q}{n-2}}$；$S_Y$越接

近于 0，说明模型对样本数据的偏差越小，预测的可靠程度(精度)越高；S_Y 数值越大，模型偏离样本数据越大，用于预测的可靠程度就越差在实际问题中，S_Y 往往较大，为评价预测模型的优劣，通常采用指标 S_Y/Y。当 S_Y/Y 小于15%时，可以认为预测模型较好。

根据结果可以计算出回归剩余标准差 RMSE= 1.65062261369908 ，S_Y/Y =0.02821577117，预测模型非常好。而且从残差值可以看出，二次模型拟合的效果非常好，并且没有剔除原始数据，保证了数据的完整性，不足之处应该是方程较为复杂。

问题三：

首先分析数据量对模型的优劣有无影响。根据问题一和问题二的求解结果发现：不管是多元线性回归，还是非线性二次回归，对同一物质而言,数据量的大小会影响模型的优劣；下面先对两个代表性物质的数据进行分析。

硫酸铝钾：硫酸铝钾的实验数据是最多的一组，按照各种浓度依次去掉最后一个数据点，逐步降低数据量，每次减少 6 个数据点，直接多元线性回归拟合，未剔除数据。具体结果见表 22

数据量	模型的评价				
	R^2	F	P	S^2	模型检验
37	0.468666833926450	5.468761515146521	0.001013036543586	1.650977684294797	成立
31	0.447711409551480	4.053237902197904	0.007854854622201	1.778012946024591	成立
25	0.393434181372659	2.464777676723364	0.069859466359325	2.063600764003753	不成立
19	0.476901392112480	2.370382181860196	0.097518202246353	1.963208135675026	不成立

表 22

由表 22 可知：随着数据量减少，P 值一直在增大，当数据量减小到 25 个数据点时，P>0.05，回归模型不再成立。

组胺：组胺的实验数据是多元线性回归拟合最优的一组，由于数据量较少，10 组数据，依次去掉最后两个数据点，逐步降低数据量，直接多元线性回归拟合，未剔除数据。具体结果见表 23

数据量	模型的评价				
	R^2	F	P	S^2	模型检验
10	0.999580110101	1904.461358300401	0.000000771022	1.312155935514	成立
8	0.999928630573	5604.240767980095	0.000178414017	0.411907532627	成立
6	1	NaN	NaN	NaN	不成立

表 23

由表 23：随着数据量减少，P 值在增大，当数据量减小到 6 个数据点时，回归模型不成立。

结合比色法的原理：待测物质溶液的浓度越低，测出来的效果越好，我们对 Data1.xls 和 Data2.xls 提供的六种物质六组数据在各自组内删除一些浓度较大的数据，即减少了组内的数据量，再使用 matlab 对组胺、奶中尿素和二氧化硫(因为溴酸钾本身数据量为 7 就少，根据问题 1 的五大准则，工业碱这组数据不是太好，硫酸铝钾的浓度都比较低)重新进行五元的线性回归分析，结果见表 24（代码见附件程序 4)。在进行多元线性回归的过程中发现，数据量不能过少，基本上 10-15 组之间，低于 6 组，回归方程一般效果就不好了甚至不成立了。

物质	数据量	R^2	F	P	S^2
组胺	10	0.9996	1904.4613	0.00000077	1.3122
	8	0.9982	220.2804	0.0045	2.4781
奶中尿素	15	0.9579	36.4208	0.00002688	37904.2069
	12	0.9265	15.1348	0.00239	45914.34
	9	0.9411	9.5897	0.04597	29441.39
二氧化硫	25	0.925031	44.41990	0.00000000166	270.651654
	21	0.8919	24.7434	0.00000095	183.52
	18	0.9721	83.6880	0.0000000066	31.24
	15	0.9704	59.0661	0.000001312	16.56
	12	0.9882	100.1326	0.00001075	4.1447

表 24

其次分析颜色维度对模型的优劣有无影响。我们通过改变颜色维度来找其对模型的影响。从显示技术上看，目前通行的 RGB 标准（红绿蓝）和 HSL 标准（色调、饱和度、亮度）是等价的，结合题目给出的数据，我们考虑是否将维度降为 3 维、2 维和 1 维，再重新进行回归分析，具体代码见程序 5，具体数据见表 25-表 28。

颜色维度 / 物质	五维 *RGBHS*	三维 *RGB*	二维 *HS*	一维 *R*	一维 *G*	一维 *B*	一维 *H*	一维 *S*
组胺	0.9958	0.99521	0.9714	0.02074	0.9940	0.9456	0.9561	0.9268
溴酸钾	0.9948	0.94078	0.9478	0.15210	0.7532	0.9142	0.4844	0.9074
工业碱	0.6314	0.48220	0.5187	0.07503	0.4408	0.2407	0.5018	0.4334
硫酸铝钾	0.4687	0.409105	0.4576	0.06742	0.3829	0.3677	0.2531	0.3757
奶中尿素	0.93299	0.83778	0.83354	0.14353	0.0044	0.8180	0.3048	0.8304
SO_2	0.85844	0.84860	0.2379	0.03607	0.7522	0.4844	0.2295	0.0224

表 25 六种不同物质在不同颜色维度下的 R^2 表

颜色维度 / 物质	五维 *RGBHS*	三维 *RGB*	二维 *HS*	一维 *R*	一维 *G*	一维 *B*	一维 *H*	一维 *S*
组胺	1904.4613	415.7519	1.1884	52.2822	1335.1724	138.971	174.222	101.3408
溴酸钾	152.446	31.7715	63.5718	0.2184	24.4117	85.2298	7.5168	78.4393
工业碱	0.3426	0.9313	2.1558	3.1897	3.94211	1.5854	5.0361	3.8246
硫酸铝钾	5.4688	7.6158	14.3433	22.1524	21.71318	20.3511	11.8590	21.0669
奶中尿素	25.0615	18.9359	30.04391	0.0635	0.05783	58.4191	5.6993	63.6545
SO_2	23.0444	39.2369	3.4339	57.1701	69.8123	21.6076	6.8496	0.5260

表 26 六种不同物质在不同颜色维度下的 F 表

颜色维度 / 物质	五维 *RGBHS*	三维 *RGB*	二维 *HS*	一维 *R*	一维 *G*	一维 *B*	一维 *H*	一维 *S*
组胺	0.00000077	0.0000002	0.000004	0.00009	0.0000000003	0.000002	0.000001	0.00000808
溴酸钾	0.00011861	0.012338	0.000032	0.6527	0.001134	0.00002	0.02538	0.0000209
工业碱	0.8518	0.106056	0.2316	0.1342	0.103849	0.26357	0.0748	0.1079

硫酸铝钾	0.0010	0.069	0.00003	0.000039	0.000045	0.00007	0.000151	0.000055
奶中尿素	0.0000496	0.02765	0.000021	0.8050	0.81370	0.000004	0.03286	0.0000023
SO_2	0.00000019	0.02085	0.05036	0.0000001	0.0000000202	0.000112	0.01540	0.475593

表 27 六种不同物质在不同颜色维度下的 P 表

颜色维度 / 物质	五维 *RGBHS*	三维 *RGB*	二维 *HS*	一维 *R*	一维 *G*	一维 *B*	一维 *H*	一维 *S*
组胺	0.00013	0.000997	0.0051	0.02074	0.00093	0.0085	0.0069	0.0114
溴酸钾	0.00163	0.012338	0.0093	0.15210	0.0386	0.0134	0.0806	0.0145
工业碱	0.2265	0.106056	0.074	0.07503	0.0687	0.0933	0.0612	0.0696
硫酸铝钾	0.0660	0.069	0.0615	0.06742	0.0679	0.0696	0.0822	0.0687
奶中尿素	0.013961	0.02765	0.02601	0.14353	0.14359	0.0263	0.1003	0.0245
SO_2	0.021544	0.02085	0.10017	0.03607	0.03116	0.0648	0.0969	0.1229

表 28 六种不同物质在不同颜色维度下的 S^2 表

从表 25-表 28 的数据发现：降低颜色的维度，模型的通用性降低了，即该模型就不适用于表示有些物质浓度与颜色读数的关系了。比如将颜色维度降低到一维，模型的四个指标都显示模型不是太好甚至是不成立的。将维度降低到 2 维或 3 维，虽然模型的四个指标都有所变化，但是模型仍然是成立的。进一步对比发现颜色维度的大小比数据量的多少对模型的影响更大。

最后我们通过建立一个层次分析模型来分析数据量和颜色维度对模型的影响因子：

第一步，构建层次结构模型，具体见图 13。

第二步，利用层次分析法对数据量和颜色维度对模型的影响因子进行求解：首先根据相对权重标度值构造图 13 中两层的 5 个成对比较矩阵，并对其一致性进行检验，然后借助于 MATLAB 的命令[v,d]=eig(A)进行求解成对比较矩阵的最大

特征值对应的特征向量，确定每个因素对上一层次该因素的权重，将两层的权重矩阵进行相乘即可得到数据量和颜色维度对模型的影响权重。

第三步，给出结果：数据量和颜色维度对模型的影响因子分别为 0.414 和 0.586。

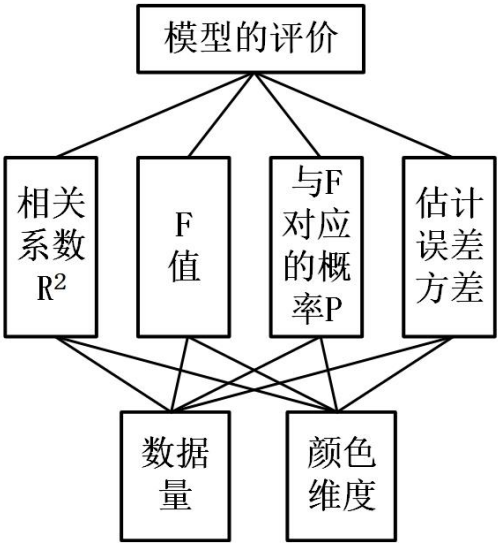

图 13 模型评价的层次结构模型

通过以上的分析和求解，我们即能分析出模型的优劣和数据量的多少、颜色维度的大小是有关的，并且通过层次分析法得出了具体的影响因子。

六、模型的评价与推广

本文给出的多元线性回归模型和改进的非线性二次回归模型不仅能很好的表示物质浓度与颜色读数之间的函数关系，而且具有如下优点：

（1）本模型是基于 Data1.xls 和 Data2.xls 中的所有数据信息建立的，并且不断的分析、检验和完善改进使得模型具有了较高的准确性，同时也确保了模型结构的严谨性。

（2）本模型具有一定的通用性，就是能反映不同物质浓度与颜色读数之间的关系；

（3）数据处理及模型求解时充分运用了 matlab 等数学软件，较好的解决了

问题，得到了较为合理的结果。

本模型可以应用于机器视觉技术在物质内部特性分析方面的研究，即机器学习领域。

七、参考文献

[1]杨海燕，贾贵儒.基于数字色度学的有色透明溶液浓度快速检测方法[J].中国农业大学学报.2006，11（3）：47-50.
[2]王岩，隋思莲，王爱青.数理统计与 MATLAB 工程数据分析[M].北京：清华大学出版社.2006:126-177.
[3]姜启源，谢金星，等.数学模型（第 3 版）[M].北京：高等教育出版社.2003.
[4]王雷震.物流运筹学[M].上海：上海交通大学出版社.2008:186-210.
[5]乔珠峰，田凤占，黄厚宽.趋势数据处理方法的比较研究[J].计算机研究与发展，2006,43（1）：171-175.
[6]黄永安，李文成，高小科.MATLAB 7.0/Simulink 6.0 应用是实例仿真与高效算法开发.北京：清华大学出版社，2008

八、附录

程序 1：Data1.xls 中的数据进行多元线性回归：

%所有数据均存在 data*.txt 文件中，*.m 文件请参见附件

f1_1.m

```
load data1_1.txt
y=data1_1(:,1);
x1=data1_1(:,2);
x2=data1_1(:,3);
x3=data1_1(:,4);
x4=data1_1(:,5);
x5=data1_1(:,6);
X=[ones(length(y),1),x1,x2,x3,x4,x5];
Y=y;
[b,bint,r,rint,stats]=regress(Y,X);
rcoplot(r,rint)
title('组胺线性回归的残差图')
format long
```

f1_2.m

```
%剔除异常点之后，请load data1_2_2.txt
load data1_2.txt
y=data1_2(:,1);
x1=data1_2(:,2);
x2=data1_2(:,3);
x3=data1_2(:,4);
x4=data1_2(:,5);
x5=data1_2(:,6);
X=[ones(length(y),1),x1,x2,x3,x4,x5];
Y=y;
[b,bint,r,rint,stats]=regress(Y,X);
rcoplot(r,rint)
title('溴酸钾线性回归的残差图')
format long
```

f1_3.m

```
load data1_3.txt
y=data1_3(:,1);
x1=data1_3(:,2);
```

```
x2=data1_3(:,3);
x3=data1_3(:,4);
x4=data1_3(:,5);
x5=data1_3(:,6);
X=[ones(length(y),1),x1,x2,x3,x4,x5];
Y=y;
[b,bint,r,rint,stats]=regress(Y,X);
rcoplot(r,rint)
title('工业碱线性回归的残差图')
format long
```

f1_4.m

```
%剔除异常点之后，请load data1_4_2.txt
load data1_4.txt
y=data1_4(:,1);
x1=data1_4(:,2);
x2=data1_4(:,3);
x3=data1_4(:,4);
x4=data1_4(:,5);
x5=data1_4(:,6);
X=[ones(length(y),1),x1,x2,x3,x4,x5];
Y=y;
[b,bint,r,rint,stats]=regress(Y,X);
rcoplot(r,rint)
title('硫酸铝钾线性回归的残差图')
format long
```

f1_5.m

```
%剔除异常点之后，请load data1_5_2.txt
load data1_5.txt
y=data1_5(:,1);
x1=data1_5(:,2);
x2=data1_5(:,3);
x3=data1_5(:,4);
x4=data1_5(:,5);
x5=data1_5(:,6);
X=[ones(length(y),1),x1,x2,x3,x4,x5];
Y=y;
[b,bint,r,rint,stats]=regress(Y,X);
rcoplot(r,rint)
title('奶中尿素线性回归的残差图')
format long
```

程序 2：Data2.xls 中的数据进行多元线性回归：

f2.m

```
%剔除异常点之后，请load data2_2.txt
load data2.txt
y=data2(:,1);
x1=data2(:,2);
x2=data2(:,3);
x3=data2(:,4);
x4=data2(:,5);
x5=data2(:,6);
X=[ones(length(y),1),x1,x2,x3,x4,x5];
Y=y;
[b,bint,r,rint,stats]=regress(Y,X);
rcoplot(r,rint)
title('二氧化硫线性回归的残差图')
format long
```

程序 3： Data2.xls 中的数据进行多元二次回归：

f2_2.m

```
load data2.txt
y=data2(:,1);
x1=data2(:,2);
x2=data2(:,3);
x3=data2(:,4);
x4=data2(:,5);
x5=data2(:,6);
X=[x1,x2,x3,x4,x5];
Y=y;
rstool(X,Y,'quadratic')
format long
```

程序4：

f4_1.m

```
x1=[121;117;120;118;120;118;120;118];
x2=[110;87;99;102;110;87;99;101];
x3=[68;46;62;66;65;46;60;64];
x4=[23;16;19;20;24;16;19;20];
x5=[111;155;122;112;115;153;126;115];
y=[0;50;25;12.5;0;50;25;12.5];
X=[ones(8,1),x1,x2,x3,x4,x5];
```

```
Y=y;
[b,bint,r,rint,stats]=regress(Y,X)
rcoplot(r,rint)
format long
title('组胺')
```

f4_2.m

```
x1=[139;139;138;139;138;142;142;139;140;138;134;138];
x2=[136;137;136;136;136;140;139;136;135;134;132;134];
x3=[118;117;108;110;120;119;111;107;125;114;112;105];
x4=[25;27;28;26;26;26;27;26;20;25;27;26];
x5=[37;41;54;52;33;40;55;58;27;44;42;60];
y=[0;500;1000;1500;0;500;1000;1500;0;500;1000;1500];
X=[ones(12,1),x1,x2,x3,x4,x5];
Y=y;
[b,bint,r,rint,stats]=regress(Y,X)
rcoplot(r,rint)
title('奶中尿素1')
```

f4_2_1.m

```
format long
x1=[139;139;138;138;142;142;140;138;134];
x2=[136;137;136;136;140;139;135;134;132];
x3=[118;117;108;120;119;111;125;114;112];
x4=[25;27;28;26;26;27;20;25;27];
x5=[37;41;54;33;40;55;27;44;42];
y=[0;500;1000;0;500;1000;0;500;1000];
X=[ones(9,1),x1,x2,x3,x4,x5];
Y=y;
[b,bint,r,rint,stats]=regress(Y,X)
rcoplot(r,rint)
title('奶中尿素2')
format long
```

f4_3.m

```
x1=[153;153;153;153;154;144;144;145;145;145;145;146;142;141;142;141;1
41;140;139;139;139];
x2=[148;147;146;146;145;115;115;115;114;114;114;114;99;99;99;96;96;96
;96;96;96];
x3=[157;157;158;158;157;170;169;172;174;176;175;175;175;174;176;181;1
82;182;175;174;176];
x4=[14;16;20;20;19;82;81;83;87;89;89;88;110;109;110;119;119;120;115;1
14;116];
x5=[138;138;137;137;141;135;136;135;135;135;135;135;137;137;136;135;1
35;135;136;136;136];
y=[0;0;0;0;0;20;20;20;30;30;30;30;50;50;50;80;80;80;100;100;100];
X=[ones(21,1),x1,x2,x3,x4,x5];
```

```
Y=y;
[b,bint,r,rint,stats]=regress(Y,X)
rcoplot(r,rint)
format long
title('二氧化硫1')
```

f4_3_1.m

```
x1=[153;153;153;153;154;144;144;145;145;145;145;146;142;141;142;141;1
41;140];
x2=[148;147;146;146;145;115;115;115;114;114;114;114;99;99;99;96;96;96
];
x3=[157;157;158;158;157;170;169;172;174;176;175;175;175;174;176;181;1
82;182];
x4=[14;16;20;20;19;82;81;83;87;89;89;88;110;109;110;119;119;120];
x5=[138;138;137;137;141;135;136;135;135;135;135;135;137;137;136;135;1
35;135];
y=[0;0;0;0;0;20;20;20;30;30;30;30;50;50;50;80;80;80];
X=[ones(18,1),x1,x2,x3,x4,x5];
Y=y;
[b,bint,r,rint,stats]=regress(Y,X)
rcoplot(r,rint)
format long
title('二氧化硫2')
```

f4_3_2.m

```
x1=[153;153;153;153;154;144;144;145;145;145;145;146;142;141;142];
x2=[148;147;146;146;145;115;115;115;114;114;114;114;99;99;99];
x3=[157;157;158;158;157;170;169;172;174;176;175;175;175;174;176];
x4=[14;16;20;20;19;82;81;83;87;89;89;88;110;109;110];
x5=[138;138;137;137;141;135;136;135;135;135;135;135;137;137;136];
y=[0;0;0;0;0;20;20;20;30;30;30;30;50;50;50];
X=[ones(15,1),x1,x2,x3,x4,x5];
Y=y;
[b,bint,r,rint,stats]=regress(Y,X)
rcoplot(r,rint)
format long
title('二氧化硫3')
```

f4_3_3.m

```
x1=[153;153;153;153;154;144;144;145;145;145;145;146];
x2=[148;147;146;146;145;115;115;115;114;114;114;114];
x3=[157;157;158;158;157;170;169;172;174;176;175;175];
x4=[14;16;20;20;19;82;81;83;87;89;89;88];
x5=[138;138;137;137;141;135;136;135;135;135;135;135];
y=[0;0;0;0;0;20;20;20;30;30;30;30];
X=[ones(12,1),x1,x2,x3,x4,x5];
Y=y;
```

```
[b,bint,r,rint,stats]=regress(Y,X)
rcoplot(r,rint)
format long
title('二氧化硫4')
```

程序 5：问题 3 中将颜色维度降为 3，2，1 分别进行多元线性回归：

三维的列举了一个代码：

f3_1.m

```
x1=[121;110;117;120;118;120;109;118;120;118];
x2=[110;66;87;99;102;110;64;87;99;101];
x3=[68;37;46;62;66;65;35;46;60;64];
y=[0;100;50;25;12.5;0;100;50;25;12.5];
X=[ones(10,1),x1,x2,x3];
Y=y;
[b,bint,r,rint,stats]=regress(Y,X)
rcoplot(r,rint)
format long
title('组胺线性回归的残差图')
```

二维的列举了一个代码：

f3_2.m

```
x4=[22;27;27;26;26;23;27;27;26;26];
x5=[27;241;145;133;106;28;242;151;132;102];
y=[0;100;50;25;12.5;0;100;50;25;12.5];
X=[ones(10,1),x41,x51];
Y=y;
[b,bint,r,rint,stats]=regress(Y,X)
rcoplot(r,rint)
format long
title('溴酸钾线性回归的残差图')
```

一维的列举了一个代码：

f3_3.m

```
x1=[139;139;138;139;142;138;142;142;139;137;140;138;134;138;138];
y=[0;500;1000;1500;2000;0;500;1000;1500;2000;0;500;1000;1500;2000];
X=[ones(15,1),x11];
Y=y;
[b,bint,r,rint,stats]=regress(Y,X)
```

```
rcoplot(r,rint)
title('奶中尿素线性回归的残差图')
format long
```