颜色与物质浓度辨识

摘　要

数码照片比色法一种检测物质浓度的方法，其原理是根据照片中的颜色读数来判断待测物质的浓度。本文根据所给数据的特征，采用合理的颜色读数值建立统计回归模型，对颜色读数与物质浓度之间的关系进行了细致的分析与研究。

针对问题一，首先分别做出各组数据中的 RGB 值与浓度的散点图，大体判断出是否采用 RGB 颜色模型或灰度颜色模型进行回归分析建模。如不可行，再结合数据特征，用 HSV 颜色模型（H 值或 S 值）与浓度建立回归分析模型。具体回归函数可以根据各组数据变化特征进行选择。

如组胺、溴酸钾、工业碱三组数据，可以采用灰度颜色模型建立一元回归分析模型，进而确定出颜色读数与物质浓度之间的关系。对于硫酸铝钾，采用 S 值与浓度建立 Michaelis-Menten 回归分析模型，也可以确定出两者之间的关系。对于奶中尿素，经过对比，最后采用 RGB 中的 B 值与浓度建立回归模型，但该组数据拟合效果相对较差。

对于数据优劣的评价，主要从数据的准确度和精密度进行分析。两者可以分别用实验测量次数和数据的标准偏差大小进行量化。通过两者的比值构造数据优劣度模型，对各组数据进行排序。最终优劣顺序为：溴酸钾、组胺、硫酸铝钾、工业碱与奶中尿素。

针对问题二，结合二氧化硫 H 值和浓度的变化规律，选用 Michaelis-Menten 模型构建回归分析方程，并对计算结果进行误差分析，删去数据异常点，进行模型改进，并做出模型预测值与原始数据的残差图。模型的预测值误差基本可以控制在 10%以内。

针对问题三，首先考虑数据量对模型的影响，一般，数据量越大越好。通过删除问题一中的部分溴酸钾溶液数据，重新建立模型与问题一中结果对比，可以看出模型的拟合效果明显变差。所以数据量应结合实际情况，至少达到一定量，并尽量做到数据分布均匀，当数据间隔较大时，应对同组数据进行多次测量。

其次考虑颜色维度对模型的影响，对问题一中工业碱溶液的模型进行改进，建立灰度值，H 值、S 值与浓度的多元线性回归模型，可以发现拟合的效果反而变差。再建立 RGB 三个值、H 值、S 值与浓度的多元线性回归模型，则模型的效果可以得到提高。由此可以判断颜色维度对模型的影响好坏不能一概而论，要结合具体的实验数据进行讨论。

关键词：数码照片比色法　颜色模型　回归分析　Michaelis-Menten 模型

一、问题的提出

1.1 问题背景

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。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。

1.2 问题重述

现根据附件所提供的有关颜色读数和物质浓度数据，建立数学模型，解决以下问题：

问题一：对附件 Data1.xls 中给出的 5 种物质在不同浓度下的颜色度数进行讨论，从 5 种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 5 组数据的优劣。

问题二：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。

问题三：讨论数据量和颜色维度对模型的影响。

二、问题分析

2.1 预备知识

数字照片比色法是一种对采集图片数字化处理的分析方法，该种方法操作简便，耗时少，成本低。其原理是根据显色溶液的特点来选择不同的颜色模型，并由分析软件得出最终结果。常用的颜色模型有 RGB、灰度、HSV 等[1] [2]。

RGB 颜色模型是由红、绿、蓝三基色通过颜色加权混合而成的一种模型，其每种颜色的取值范围为[0,255]。

灰度颜色模型是用 0 到 255 的不同灰度值来表示图像，0 表示黑色，255 表示白色，灰度模式可以由 RGB 模式直接转换得到。在比色法中，用灰度颜色模型对显色结果进行分析是比较简便的。

HSV 颜色模型是由每一种颜色都是由色调，饱和度和明度三个变量所决定的颜色模型。其在计算机图像处理、车牌识别等领域用途较为广泛。

2.2 问题的分析

针对问题一，首先对 5 组数据画出 RGB 值与浓度的散点图，从而大体判断能否用 RGB 颜色模型或灰色颜色模型进行回归分析建模。如果 RGB 值与浓度关系不明显或拟合效果不佳，再用 HSV 颜色模型（H 值或 S 值）与浓度建立回归分析模型。具体回归函数可以根据各组数据变化特征进行选择，从而建立各组颜色读数与浓度的数学模型。

对于评价数据的优劣，可以从数据的准确度和精密度进行分析。准确度主要从测量次数分析，精密度主要依靠数据的标准偏差大小进行量化。通过两者的比值构造数据优劣度模型，对各组数据进行排序。

针对问题二，首先观察二氧化硫溶液的 RGB 值、H 值与 S 值与浓度变化趋势，可以发现 H 值与浓度变化关系最为明显，结合 H 值数据的变化规律，选用 Michaelis-Menten 模型构建回归分析方程，并对计算结果进行误差分析，筛选掉数据异常点，建立更精确的回归模型。并给出预测值与原始数据的残差图，模型预测值的误差基本控制在 10%以内。

针对问题三，首先考虑数据量对模型的影响，单纯从建模需要来讲，样本容量肯定是越大越好。若删除问题一中的部分溴酸钾溶液数据，重新建立模型，可以得到模型的拟合效果明显变差。因此，根据实验要求不同，数据量至少达到一定量，并尽量做到数据分布均匀，当数据间隔较大时，可对同组数据进行多次测量。

其次考虑颜色维度对模型的影响，对问题一中工业碱溶液的模型进行改进，结合数据特征，将灰度颜色模型，与 H 值、S 值一起建立多元线性回归模型，发现拟合的效果反而变差。再将 RGB 三个值、H 值、S 值一起建立多元线性回归模型，则模型的效果可以得到提高。由此可以判断颜色维度对模型的影响好坏不能一概而论，要结合具体的实验数据进行讨论。

三、模型假设

1、假设各组照片的拍摄环境是一致的。

2、忽略拍摄环境（距离、角度、温度）对读数的影响。

3、假设各组颜色数据的读取设备是同一台设备的。

4、假设试纸没有过期、无破损。

5、假设溶液与试纸已充分反应。

四、符号说明

L	灰度值
R	红色数值
G	绿色数值
B	蓝色数值
H	表示色调值
S	表示饱和度值
M	溶液浓度
$\hat{M}$	溶液浓度预测值

五、模型的建立与求解

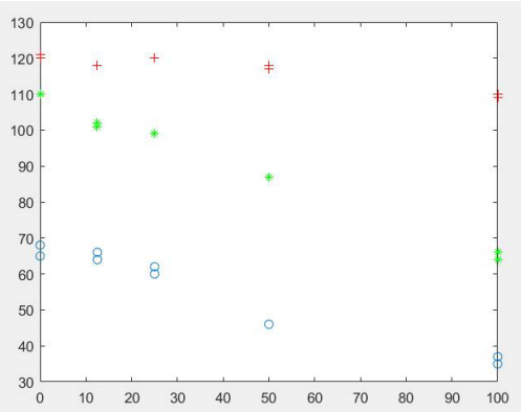
5.1 问题一的模型建立与求解

5.1.1 颜色读数与浓度关系的确定

由于数字照片比色法要根据不同的数据情况选择合适的颜色模型来分析结构，且在附件 Data1 所给数据中包含了 5 种颜色读数，包括红（R）、绿（G）、蓝（B）三基色的取值和色调（H）、饱和度（S）的取值，因此要分别对各组的具体数据选择合适的颜色模型进行分析求解。

项目一：组胺

先做出 RGB 三组值与浓度的散点图，见图 5-1。



‘+’代表红色，‘*’代表绿色，‘o’代表蓝色

图 5-1　RGB 值与组胺浓度的散点图

从图 5-1 可以发现，随着浓度的的增加，R 值变化较小，但 G 值与 B 值有比较明显的线性递减趋势，综合上述分析，将结合灰度颜色模型建立一元回归分析模型。

首先利用 RGB 的数据计算出灰度值 L，计算公式[1]如下：

$$L = 0.299 \times R + 0.587 \times G + 0.114 \times B$$

则组胺浓度与灰度值的对应数据如表 5-1。

表 5-1 组胺浓度与灰度值

浓度	100	100	50	50	25	25	12.5	12.5	0	0
灰度值L	76	74	91	92	101	101	103	102	109	108

建立线性回归模型[3]：

$$M = \beta_0 + \beta_1 \cdot L + \varepsilon \qquad (1)$$

其中 β_0，β_1 为回归系数，ε 为随机误差。利用 Matlab 软件的 regress 函数求解，模型（1）的回归系数估计值及其置信区间（置信水平 α=0.05）、检验统计量 R^2, F, p, s^2 的结果见表 5-2.

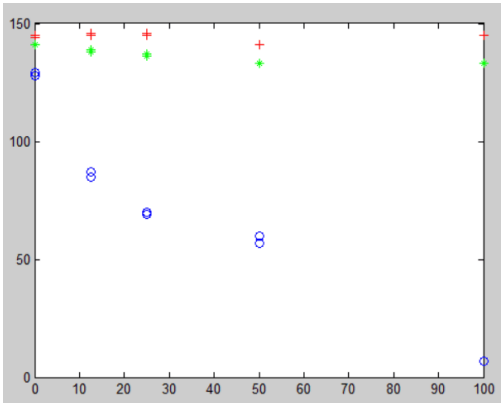
表 5-2 组胺浓度的回归结果

参数	参数估计值	参数置信区间
β_0	325.2068	[304.1089,346.3047]
β_0	-3.0063	[-3.2252,-2.7875]
$R^2 = 1.0$	F=1003.7	p=0.0 $s^2 = 12.4$

结果分析：从表 5-2 中可以看出 R^2 近似为 1，F 值大于 F 检验的临界值，p 远小于 0.05，且每个回归系数的置信区间没有包含零点，说明灰度值 L 对浓度 M 影响是显著的，因此模型（1）从整体上是合理可用的。说明组胺溶液的浓度可以通过颜色读数来确定，其预测方程的关系式为：$\hat{M} = 325.2068 - 3.0063L$ 。

项目二：溴酸钾

针对溴酸钾溶液，同样先做出 RGB 三组值与浓度的散点图。见图 5-2。



‘+’代表红色，‘*’代表绿色，‘o’代表蓝色

图 5-2　RGB 值与浓度(溴酸钾)的散点图

从图 5-2 可以发现，随着浓度的的增加，R 值和 G 值变化较小，但 B 值有明显的线性递减趋势，同理，可以建立灰度颜色模型的一元回归分析模型。

先计算出溴酸钾浓度与灰度值的对应数据如表 5-3。

表 5-3 溴酸钾浓度与灰度值

浓度	100	100	50	50	25	25	12.5	12.5	0	0
灰度值L	122	122.000	127.000	127.000	131.000	132.000	135.000	135.000	141.000	140.000

将上述数据代入模型（1）中，可得出结果（见表 5-4）。

表 5-4 溴酸钾浓度的回归结果

参数	参数估计值	参数置信区间
β_0	729.5510	[548.9463，910.1558]
β_1	-5.2748	[-6.6497，-3.8998]
$R^2 = 0.9073$ F=78.2646 p=0.0000 $s^2 = 144.9031$		

结果分析：从表 5-4 中可以看出 R^2 为 0.9073，F 值大于 F 检验的临界值，p 远小于 0.05，且每个回归系数的置信区间没有包含零点，说明灰度值 L 对浓度 M 影响是显著的。说明溴酸钾溶液的浓度可以通过颜色读数来确定，其预测方程的关系式为：

$$\hat{M} = 729.551 - 5.2748L \text{ 。}$$

模型改进：如果用二次函数建立回归模型[3]

$$M = \beta_0 + \beta_1 L + \beta_1 \hat{L} + \qquad (2)$$

其中 β_0，β_1，β_2 为回归系数。代入数据可得计算结果如表

表 5-5 溴酸钾浓度的回归结果（二次函数）

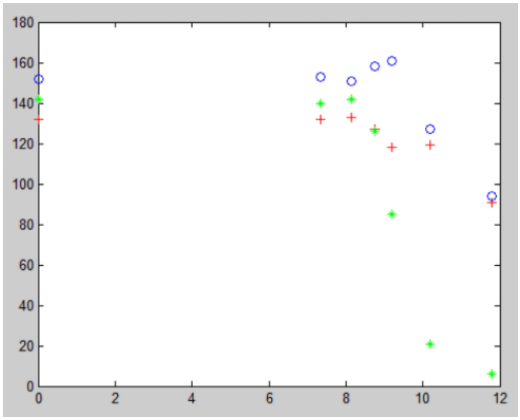
参数	参数估计值	参数置信区间
β_0	5561.2	[4603.5，6519.0]
β_1	0.3	[0.2，0.3]
β_2	-79.1	[-93.7，-64.5]
$R^2 = 0.9957$ F=803.0623 p=0.000 $s^2 = 7.7489$		

结果分析：从表 5-5 中可以看出 R^2、F 值明显增大，p 远小于 0.05，且每个回归系数的置信区间没有包含零点，说明该模型拟合效果更好，其预测方程的关系式为：

$$\hat{M} = 5561.2 + 0.3L - 79.1L^2$$

项目三：工业碱

针对工业碱溶液做出 RGB 三组值与浓度的散点图。见图 5-3。



‘+’代表红色，‘*’代表绿色，‘o’代表蓝色

图 5-3　RGB 值与浓度(工业碱)的散点图

从图 5-3 可以发现，除个别点外，随着浓度的的增加， RGB 三个值都有较明显的线性递减趋势。同样可以灰度颜色模型的一元回归分析模型。

计算出工业碱浓度与灰度值的对应数据如表 5-6。

表 5-6 工业碱浓度与灰度值

浓度	11.8	10.18	9.19	8.74	8.14	7.34	0
灰度值L	41	62	104	130	140	139	140

可以发现，本组的数据量明显偏少，且所给出的浓度变化范围主要集中在 7~12 之间，将上述数据代入模型（1）中，可得出结果（见表 5-7）。

表 5-7 工业碱溶液回归结果

参数	参数估计值	参数置信区间
β_0	14.4799	[5.3960，23.5637]
β_1	-0.0608	[-0.1401，0.0186]
$R^2 = 0.4368$　F=3.8779　p=0.1060　$s^2 = 9.6347$		

从表 5-6 中可以看出 R^2 为 0.4368，p 大于 0.05，说明模型不可用。做出残差示意图，如图 5-4 所示。

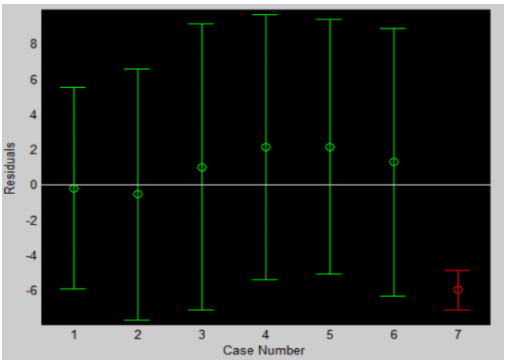


图 5-4 残差示意图

明显第 7 个数据点是一个异常点。因此删除该异常点（浓度为 0），重新代入模型计算，结果见表 5-8。

表 5-8 工业碱浓度的回归结果

参数	参数估计值	参数置信区间
β_0	12.9142	[11.2751，14.5533]
β_1	-0.0359	[-0.0508，-0.0209]
$R^2 = 0.9174$ F=44.3981 p=0.0026 $s^2 = 0.2585$		

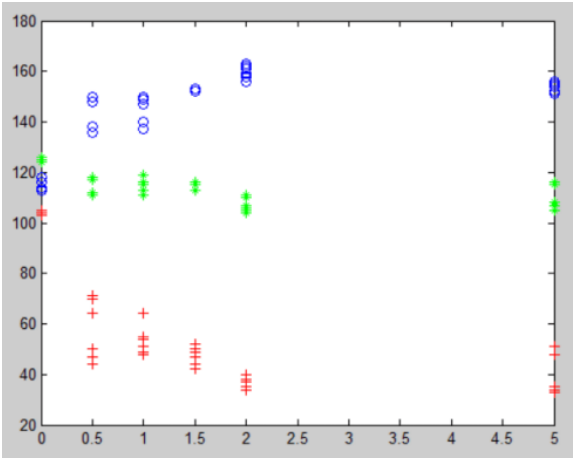
结果分析：从表 5-8 中可以看出 R^2 为 0.9174， F 值大于 F 检验的临界值， p 小于 0.05，且每个回归系数的置信区间没有包含零点，说明灰度值 L 对浓度 M 影响是显著的。说明工业碱溶液的浓度可以通过颜色读数来确定，其预测方程的关系式为：

$$\hat{M} = 12.9142 - 0.0359L$$

注意：由于该组数据量偏少，且浓度变化范围主要集中在 7~12 之间，因此预测方程应用范围也只能大体限定在[7,12]上。

项目四：硫酸铝钾

针对硫酸铝钾溶液做出 RGB 三组值与浓度的散点图。见图 5-5。



‘+’代表红色，‘*’代表绿色，‘o’代表蓝色

图 5-5　RGB 值与浓度(硫酸铝钾)的散点图

从图 5-5 发现，随着浓度的的增加， RGB 三个值没有明显的变化趋势。同时，把该组数据代入灰度值的回归分析模型，发现模型不成立。因此，我们考虑用 HSV 颜色模型来进行回归分析建模。

首先分别做出 H 值、S 值关于浓度变化的散点图，见图 5-6。

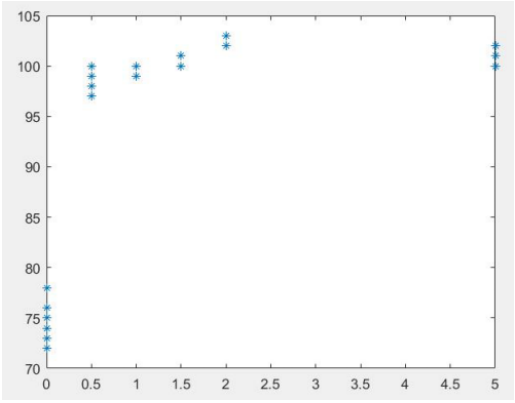


图 5-6(a)H 值与浓度(硫酸铝钾)的散点图

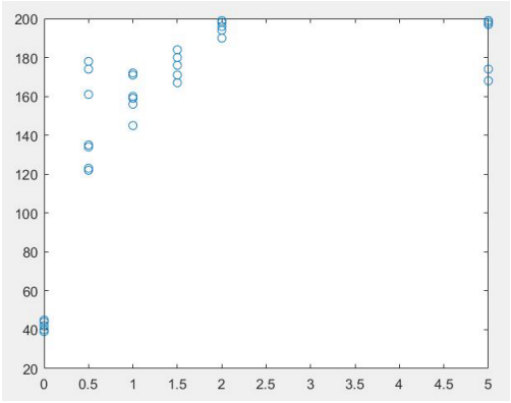


图 5-6(b) S 值与浓度(硫酸铝钾)的散点图

从图 5-6 可以发现，随着浓度的的增加，H 值与 S 值的变化规律基本相似。当浓度较小时，H 值与 S 值都快速增加，而当浓度很大时，H 值与 S 值增加较慢，趋于一个稳定值。一般满足上述性质的模型有两个[3]：

Michaelis-Menten 模型 $$y=\frac{\beta_0 x}{\beta_1+x} \tag{3}$$

指数增长模型 $$y=\beta_0(1-e^{-\beta_1 x}) \tag{4}$$

下面我们用 Michaelis-Menten 模型对该组数据中的 H 值和 S 值进行回归分析。

首先，将 S 值与浓度的数据代入模型（3），其中，浓度作为自变量，S 值作为因变量。作为利用 Matlab 软件的 nlinfit 函数求解，可得出结果（见表 5-9）。

表 5-9 硫酸铝钾浓度的回归结果

参数	参数估计值	参数置信区间
β_0	200.9061	[183.1188，218.6934]
β_1	0.1945	[0.0781，0.3109]

拟合后的结果与原始数据进行对比见图 5-7，且模型的剩余标准差 s=22.396。

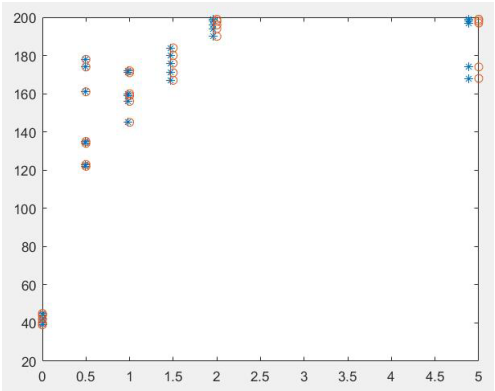


图 5-7　模型（2）的预测结果(S 值与硫酸铝钾浓度)

（‘*’ 代表拟合结果，‘o’代表原始数据）

由上图可知，拟合效果良好，且回归系数的置信区间没有包含零点，说明模型是完全可

用的。将回归得出的系数代入模型（3）可得方程

$$y=\frac{200.9061x}{0.1945+x}$$

求解反函数得：$x=\frac{0.1945\cdot y}{200.9061-y}$，则浓度与 S 值的预测方程为

$$\hat{M}=\frac{0.1945\cdot S}{200.9061-S}.$$

同理，将 H 值与浓度数据代入模型（3），拟合后的效果见图 5-8，其剩余标准差为 30.9423。

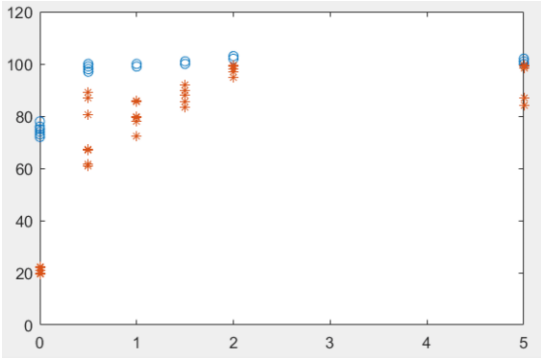


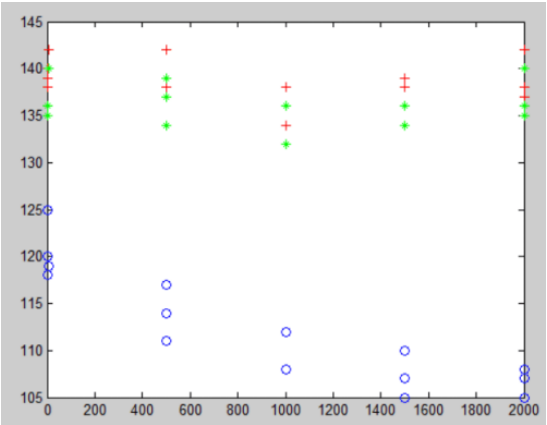
图 5-8　模型（2）的预测结果（H 值）

对比 S 值与 H 值得拟合结果，可以发现 S 值的拟合效果更好，因此可以采用 S 值来建立颜色读数与浓度之间的关系，其预测方程为

$$\hat{M}=\frac{0.1945\cdot S}{200.9061-S}.$$

项目五：奶中尿素

针对奶中尿素做出 RGB 三组值与浓度的散点图。见图 5-9。



‘+’代表红色，‘*’代表绿色，‘o’代表蓝色

图 5-9　RGB 值与浓度(奶中尿素)的散点图

从图 5-9 可以发现，随着浓度的的变化， R 值与 G 值没有明显的变化趋势，而 B

值有较弱的线性递减趋势。同样先用灰度颜色模型的一元线性回归分析模型，结果见表5-10。显然该模型不适用。我们试着采用二次函数模型，也未得到理想的结果。

表 5-10 奶中尿素浓度与灰度值的回归结果

参数	参数估计值	参数置信区间
β_0	19035	[-8846，46915]
β_1	-135	[-342，73]
$R^2 = 0.1315$ F=1.9677 p=0.1841 $s^2 = 562813.1712$		

其次，分别做出 H 值、S 值关于奶中尿素浓度变化的散点图，见图 5-10。

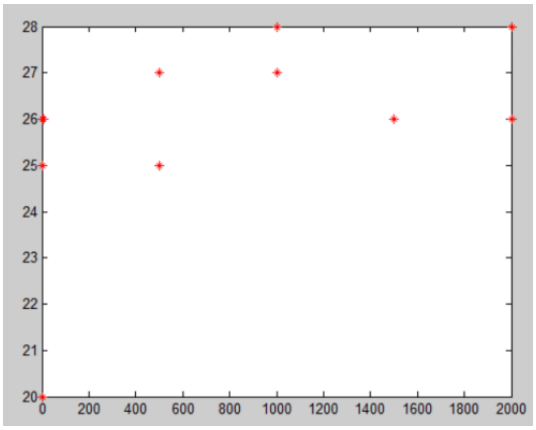


图 5-10(a) H 值与浓度(奶中尿素)的散点图

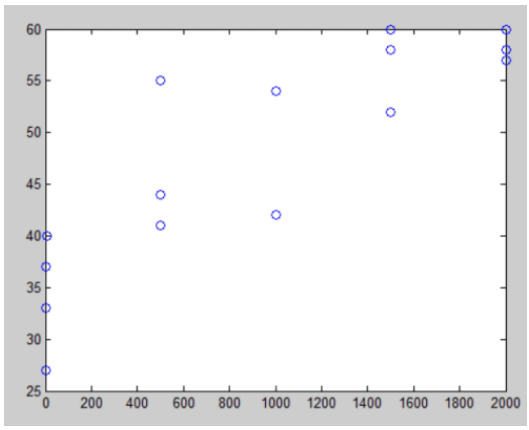


图 5-10(b) S 值与浓度(奶中尿素)的散点图

从图 5-10 可以看出，随着浓度的的增加，H 值的变化很小，而 S 值有较弱的线性增加关系。 因此用 S 值与浓度做一元回归分析模型，其结果见 5-9。

表 5-11 奶中尿素浓度与 S 值的回归结果

参数	参数估计值	参数置信区间
β_0	-2042.3	[-3112.3，-972.4]
β_1	62.2	[40.3，84.0]
$R^2 = 0.7441$ F=37.8100 p=0.000 $s^2 = 165794.5104$		

其中， R^2 为 0.7441，说明拟合效果一般。

从图 5-9 可知，B 值随着浓度有较弱的线性递减趋势，因此用 B 值与浓度做一元回归分析模型，其结果见表 5-12。

表 5-12 奶中尿素浓度与 B 值的回归结果

参数	参数估计值	参数置信区间
β_0	13576	[9728，17424]
β_1	-112	[-147，-78]
$R^2 = 0.7953$ F=50.5149 p=0.000 $s^2 = 132630.6165$		

其中，R^2 为 0.7953，p 小于 0.05，且每个回归系数的置信区间没有包含零点，说明模型整体可用，但效果比前 4 组数据明显较差。综上所述，可以建立颜色读数与奶中尿素浓度之间的关系，其预测方程为

$$\hat{M} = 13576 - 112B.$$

5.1.2 数据的评价

一般，评价实验数据的准则[4]主要包括两个方面：准确度和精密度。

由于原始数据并没有给出浓度的真实值，那准确度主要从实验数据的数量来考虑，一般实验测定次数越多，数据的准确度越高。5 组项目的实验测定次数见表：

表 5-13 各组项目的实验测定次数

	组胺溶液	溴酸钾溶液	工业碱溶液	硫酸铝钾溶液	奶中尿素溶液
测量次数	10	10	7	37	15

再考虑精密度，可以将已知数据分为两组，分组依据是依照其是否进行重复测量。可将之分为:

A 组：组胺溶液、溴酸钾溶液、硫酸铝钾溶液、奶中尿素溶液

B 组：工业碱溶液

如针对组胺溶液，我们对其 5 个颜色值进行权重系数的计算，方法如下：

$$\frac{\max_j(x_{ij}) - \min_j(x_{ij})}{\sum_i [\max_j(x_{ij}) - \min_j(x_{ij})]}$$

其中 x_{ij} 为第 i 行第 j 列的数据值。具体结果见表 5-14.

表 5-14 组胺溶液权重系数表

权重系数	0.14	0.22	0.25	0.05	0.35

随后计算组胺溶液各组浓度下的不同颜色值标准偏差，并将各组颜色值标准差求和，见表 5-15.

表 5-15 组胺溶液标准偏差计算结果

浓度（ppm）	B	G	R	H	S
100 标准偏差	1.41	1.41	0.71	0.71	2.12
50 标准偏差	0.00	0.00	0.71	0.00	1.41
25 标准偏差	1.41	0.00	0.00	0.00	2.83
12.5 标准偏差	1.41	0.71	0.00	0.00	2.12
0 标准偏差	2.12	0.00	0.71	0.71	2.83
求和结果	6.35	2.12	2.13	1.42	11.31

对求和结果乘上相应的权重系数，得到加权标准偏差和为 5.85。

同理，其他三个项目数据的加权标准偏差和分别为

表 5-16 其余溶液加权标准偏差和

	溴酸钾溶液	硫酸铝钾溶液	奶中尿素溶液
加权总标准偏差和	4.95	30.18	12.31

（其中，奶中尿素溶液中的 5ppm 组数据由于只有一组，视为异常点，将其除去。）

最后，为了排出各组数据的优劣顺序，要综合考虑数据的数据量与加权标准偏差和，则将加权标准偏差和除以相应的数据量，该比值可定义为数据优劣度，数据越小代表该组数据越优。最后排序结果如表 5-17.

表 5-17 A 组溶液优劣度排序

	溴酸钾溶液	组胺溶液	硫酸铝钾溶液	奶中尿素溶液
数据优劣度	0.4649	0.5848	0.8158	0.8207

由此可知溴酸钾溶液数据最优，组胺溶液数据其次，奶中尿素数据最差。

针对 B 组的工业碱溶液，由于各浓度均只有一组数据，故无法计算其标准偏差，所以选择对其各组数据的浓度间隔进行分析，如下表所示：

表 5-18 工业碱溶液各组数据的浓度间隔

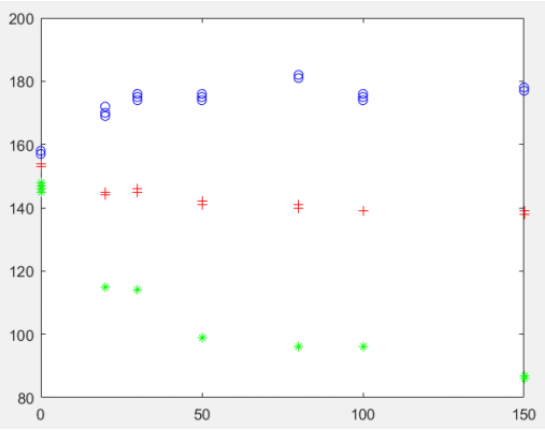
1-2组间隔	2-3组间隔	3-4组间隔	4-5组间隔	5-6组间隔	6-7组间隔
1.620	0.990	0.450	0.600	0.780	7.360

由此可知，该组溶液中，前 5 个浓度数据间隔较为均衡，但 6-7 组数据间隔过大。且该组溶液测量数据较少，所以认为其数据优劣度较差。

5.2 问题二的模型建立与求解

5.2.1 模型的建立与求解

与问题一类似，附件 Data2 中给出的数据也包括 RGB 的取值和 H、S 的取值，分别作出上述取值与浓度的散点图，见图 5-11、图 5-12。



‘+’代表红色，‘*’代表绿色，‘o’代表蓝色

图 5-11 RGB 值与浓度(二氧化硫)的散点图

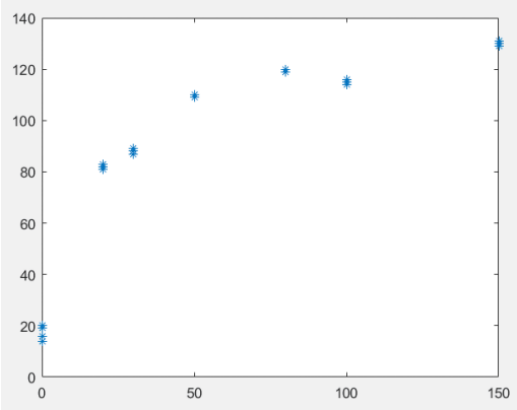


图 5-12(a) H 值与浓度(二氧化硫)的散点图

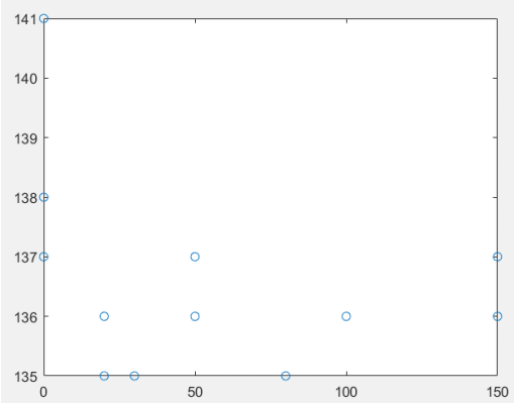


图 5-12(b) S 值与浓度(二氧化硫)的散点图

从上述图中可以发现，随着浓度的的增加，RGB 值与 S 值的没有明显的变化规律。而 H 值有明显的变化，当浓度较小时，H 值快速增加，当浓度很大时，H 值增加较慢，趋于一个稳定值。因此，可以根据 Michaelis-Menten 模型构建 H 值与浓度的回归方程。

将 H 值与二氧化硫浓度值的数据代入模型（3），其中，浓度作为自变量，H 值作为因变量。作为利用 nlinfit 函数求解，可得出结果（见表 5-19）。

表 5-19 二氧化硫浓度的回归结果

参数	参数估计值	参数置信区间
β_0	140.8184	[129.4161，152.2207]
β_1	15.8583	[10.2000，21.5167]

拟合后的结果与原始数据进行对比见图 5-13，且模型的剩余标准差 s=9.1231。

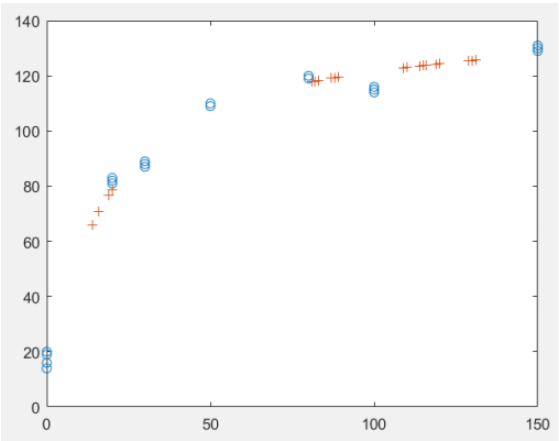


图 5-13 模型（3）的预测结果(H 值与二氧化硫浓度)

（‘+’代表拟合结果，‘o’代表原始数据）

由上图可知，拟合效果良好，且回归系数的置信区间没有包含零点，说明模型是完全可用的。将回归得出的系数代入模型（3）可得方程

$$y=\frac{140.8184x}{15.8583+x}$$

求解反函数得：$x=\dfrac{15.8583\cdot y}{140.8184-y}$，则浓度与 H 值的预测方程为

$$\hat{M}=\frac{15.8583\cdot H}{140.8184-H}.$$

5.2.2 误差分析及模型改进

利用 nlintool 函数做出上述 Michaelis-Menten 模型的预测和结果输出图。

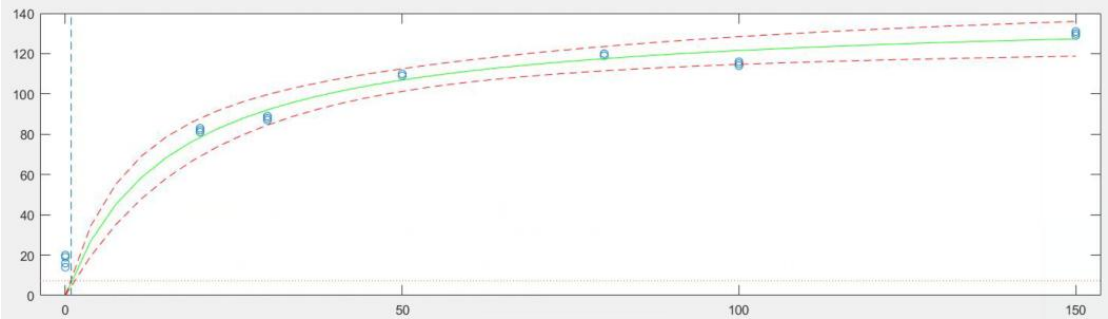


图 5-14（a） H 值与二氧化硫浓度的预测及结果输出

可以看到，除了浓度为 0 和 100 时的点，其他的原始数据都在模型的预测区间内，说明模型整体是可行的。

如果删去上述异常点，重新求解，其预测和结果输出图，

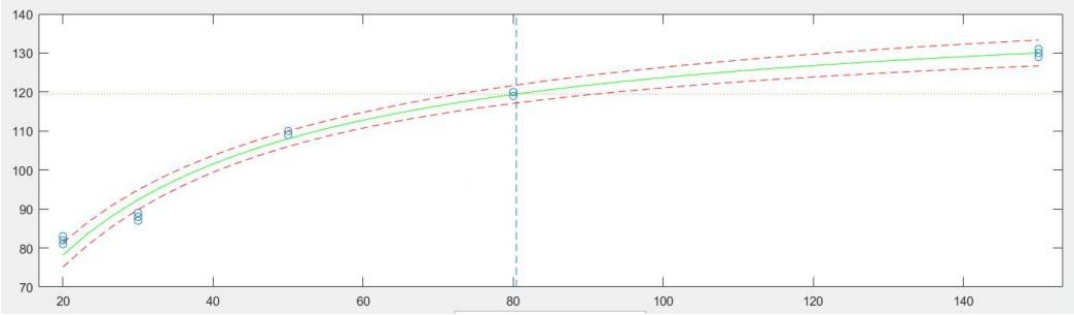


图 5-14（b）H 值与二氧化硫浓度的预测及结果输出

剩余标准差 s= 2.9233，可以发现剩余标准差较原模型更小，且预测区间的长度也大幅缩短。

表 5-20 预测值与预测区间的比较（预测区间为预测值 ±Δ ）

浓度	实际数据平均值	模型1预测值	△1	模型2预测值	△2
20	82.0000	78.5415	9.4315	78.2285	3.1057
50	109.6667	106.9101	5.6072	108.0142	1.9913
80	119.3333	117.5221	6.0305	119.3775	2.2909

计算出删去上述异常点后模型的回归结果（见表 5-20）。

表 5-21 二氧化硫浓度的回归结果(删去上述异常点)

参数	参数估计值	参数置信区间
β_0	144.7591	[140.4610，149.0571]
β_1	17.0093	[15.0094，19.0092]

同理求解反函数可得到浓度与 H 值的预测方程为

$$\hat{M} = \frac{17.0093 \cdot H}{144.7591 - H}.$$

给出上述 $\hat{M}$ 与实际浓度的残差图，可以发现该模型的预测值的误差基本控制在 10%以内。

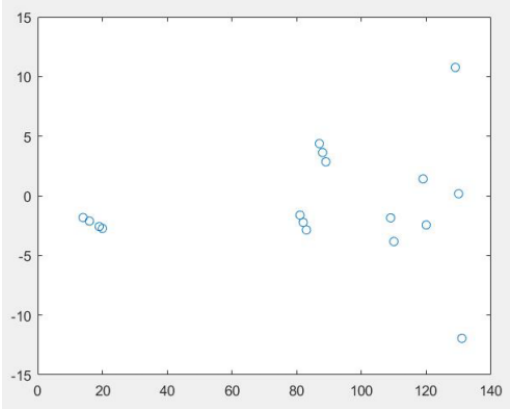


图 5-15 实际浓度残差图

5.3 问题三模型的建立与求解

5.3.1 数据量对模型的影响

由于本文是采用回归分析模型，它对样本数据量具有很强的依赖性。样本的容量太小会导致难以保证参数估计值的精确度和可靠性[4]。因此，单纯从建模需要来讲，样本容量肯定是越大越好。如在问题一中的溴酸钾溶液数据，共 10 组数据，如果去掉有重复值的 5 组，再建立模型，其计算结果如表 5-22：

表 5-22 溴酸钾溶液的回归结果（去重复值）

参数	参数估计值	参数置信区间
β_0	707.9887	[272.6000，1143.4000]
β_1	-5.1104	[-8.4000，-1.8000]
$R^2 = 0.8892$ F=24.0795 p=0.0162 $s^2 = 230.8016$		

该结果较之前表 5-4 的结果有明显的变化，其中 R^2，F 值变小， p, s^2 明显增大，且回归系数的置信区间长度也明显变大，说明模型的拟合效果明显减弱。

另一方面，收集样本数据量是一件困难的工作，因此，选择合适的样本数据量，是一个非常重要的问题。一般，满足基本要求的样本容量需满足 $n \geq 30$ 或者 $n \geq 3(k+1)$，其中 k 为解释变量的数目。

综上所述，建议数字比色法的数据量至少达到 12 组，尽量做到数据分布均匀，当数据间隔较大时，可对同组数据进行多次测量。

5.3.2 颜色维度对模型的影响

一般，数字比色法是根据显色溶液的特点来选择不同的颜色模型进行分析，如在前两问中有分别用灰度值颜色模型、HSV 颜色模型、RGB 颜色模型进行一元回归建模。如果对于某种溶液，RGB 值与 HSV 值都随着浓度的变化有明显的变化规律，那我们可以综合多种颜色模型来建立模型。

如问题一中工业碱溶液的模型是根据灰度值颜色模型来进行回归分析，可以发现其 H 值、S 值与浓度也有着明显的线性关系，见图 5-16：

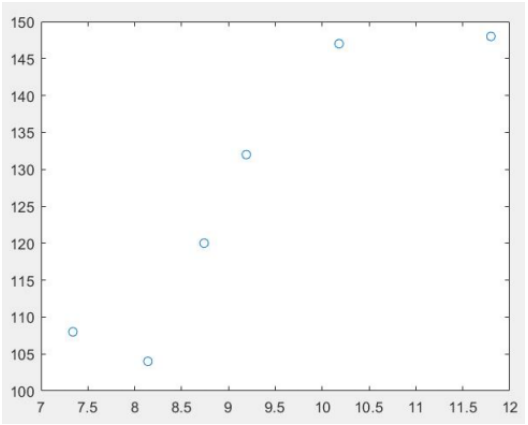


图 5-16（a） H 值与浓度(工业碱)的散点图

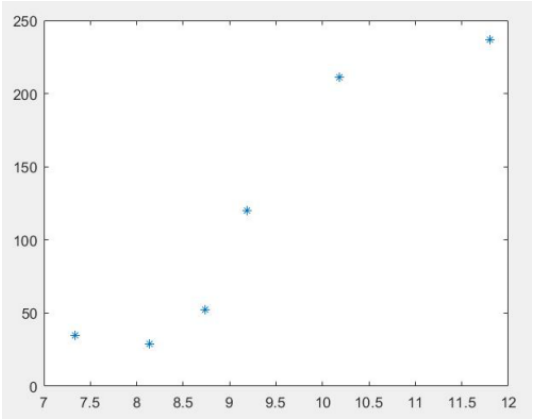


图 5-16（b） S 值与浓度(工业碱)的散点图

因此，我们可以用灰度值、H 值和 S 值对浓度进行多元回归分析，建立多元线性回归模型：

$$M = \beta_0 + \beta_1 H + \beta_2 S + \beta_3 L + \tag{5}$$

其中 $\beta_0, \beta_1, \beta_2, \beta_3$ 为回归系数。利用 regress 函数求解，其结果见表 5-23：

表 5-23 多元线性回归结果

参数	参数估计值	参数置信区间
β_0	17.7632	[-32.1469,67.6732]
β_1	-0.1161	[-0.6480,0.4158]
β_2	0.0285	[-0.0475,0.1045]
β_3	0.0283	[-0.0807,0.1373]
$R^2 = 0.9283$ F=8.6265 p= 0.1057 $s^2 = 0.4487$		

其中 $p>0.05$，且每个回归系数的置信区间都包含零点，说明该模型是不可行的。

而如果用 RGB 值、H 值和 S 值对浓度进行多元回归分析，可以计算出结果，见表 5-24.

表 5-24 多元回归分析（RGB、H、S）

参数	参数估计值	参数置信区间
β_0	93.7528	NaN
β_1	0.0497	NaN
β_2	-0.4187	NaN
β_3	-0.1260	NaN
β_4	-0.0934	NaN
β_5	-0.2482	NaN
$R^2=1$ F=NaN p=NaN $s^2=$NaN		

说明拟合的效果非常好。其预测值与原始数据对比见表 5-25：

表 5-25 预测值与原始数据对比

源数据浓度	7.34	8.14	8.74	9.19	10.18	11.8
计算数据浓度	7.34000000000003	8.14000000000005	8.74000000000005	9.19000000000004	10.18000000000010	11.80000000000010
误差绝对值	3.01981E-14	4.9738E-14	4.9738E-14	4.08562E-14	9.9476E-14	9.9476E-14

综上所述，颜色维度对模型的影响好坏不能一概而论，要结合具体的实验数据进行讨论。如果能选择合理的颜色模型搭配进行建模，则可以提高模型精度。反之，盲目的提高颜色维度进行建模，可能会降低模型的效果。

六、模型的评价与推广

6.1 模型的评价

本文根据数字比色法的基本思想，灵活运用 EXCEL， MATLAB 对所给数据进行处理，建立了统计回归模型，且模型的拟合效果整体良好，充分说明了颜色读数与浓度之间的关系。当然，由于对颜色模型的转换关系与所给数据的局限性，无法选用更好的颜色模型来建立模型。

6.2 模型的推广

本文在主要运用了统计回归模型，如一元线性回归模型，非线性回归模型及多元回归模型，其具有有效性高，适用范围广的特点，可推广到农药残留检测、水质检测、生物分子检查等领域。

七、参考文献

[1] 杨冬冬，张校亮，崔彩娥 基于智能手机数字比色法的有机磷农残快速检测技术研究 《分析测试学报》 34 (10) :1179-1184 2015