Open Access Library Journal
2018, Volume 5, e4377
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Working with Color Readings: Application of Regression Models for Determining the Concentration of Substance

Xinyuan Pan, Yan Cui

Department of liberal Arts Education, Guangdong Lingnan Institute of Technology, Guangzhou, China
Email: 2006pxy@163.com, 2175910809@qq.com

How to cite this paper: Pan, X.Y. and Cui, Y. (2018) Working with Color Readings: Application of Regression Models for Determining the Concentration of Substance. *Open Access Library Journal*, **5**: e4377.
https://doi.org/10.4236/oalib.1104377

Received: January 26, 2018
Accepted: February 24, 2018
Published: February 27, 2018

Abstract

This study first considered using 5 dimensions/variables of color readings (including blue, green, saturation, red, hue) to predict the concentration of matter. The procedure was demonstrated using internet acquired data in the public domain. A stepwise regression method based on software MATLAB was used to build a model to predict the concentration of the sulfur dioxide. In the course of the study, we also discussed other linear or non-linear models to predict the concentration of the sulfur dioxide, but these models didn't perform as the former, which may be due to the strong collinearity between the original data variables. Statistical test results of the model including the number of observations, root mean squared error, adjusted R-square, F-statistic vs constant model were shown to assess the reliability. We also carried out error analysis and discussion.

Subject Areas

Mathematical Analysis, Mathematical Statistics

Keywords

Stepwise Regression, Statistical Test, Colorimetry, Concentration of Substances

1. Introduction

Colorimetry [1] is a commonly used method to detect the concentration of matter, that is, after the material to be measured into a solution, dripping in a specific white paper surface, such as its full response to obtain a color of the strip, and then the color test paper and a standard colorimetric card to compare, you

can determine the concentration of the material to be measured. Because of each person's sensitivity to color differences and observation errors, this method has a great impact on the accuracy. With the improvement of camera technology and color resolution, it is hoped to establish a quantitative relationship between the color readings and the material concentration, *i.e.* the concentration of the substance to be measured can be obtained as long as the color readings in the photograph are entered. In this article we have considered to predict the concentration of sulfur dioxide using 5 dimensions of color readings such as blue, red, green, hue, saturation. The procedure was demonstrated using internet acquired data in the public domain [2]. A stepwise regression method based on software MATLAB was used to build a model to predict the concentration of the sulfur dioxide. Statistical test results of the models including the number of observations, root mean squared error, adjusted R-square, F-statistic vs constant model, p-value are shown to assess the reliability. All the results reveal that the models are significant for statistical significant level at $\alpha = 0.05$. We also carried out error analysis and discovered that the error of two data samples at 5% significant level is outlier; the other samples' error is within the allowable range. Because the internet acquired data in [2] are rare, we use all the data for regression fitting. In fact, in the study of related problems, if more data can be obtained, we should carry out cross-validation and reduce the problem of over-fitting. This article merely provides a mathematical thinking method for determining the material concentration according to the color readings, and its application has the advantages of generalization and reference.

2. Data Set

Data for a particular material sulfur dioxide (SO_2) are given in **Table 1** from "CUMCM2017Problems\C\Data2" [2]. The data consist of 25 observations showing color readings and the relative material concentration. The "concentration (ppm)" in the table indicates the number of milligrams of the substance to be measured per liter of pure water (or solvent), the "water" in the table indicates a situation where the concentration of the material to be measured is zero. In the following part we take the concentration as the response variable, and the color variables or dimensions (blue, red, green, hue, saturation) as the predictor variables.

3. Data Analysis

We find that the green, blue, and hue readings in **Table 1** vary greatly with the concentration, while the readings of red and saturation are not apparent with the concentration. Then, we plot the graph of concentration and color variables using Python to watch the relationship between them, see **Figure 1**. From **Figure 1** we notice there is strong linear relationship between the concentration of sulfur dioxide and color dimensions. In order to portray their relationship more profoundly, we also use Python to draw the correlation coefficient heat-map (see

Table 1. Concentration (sulfur dioxide) vs color readings.

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

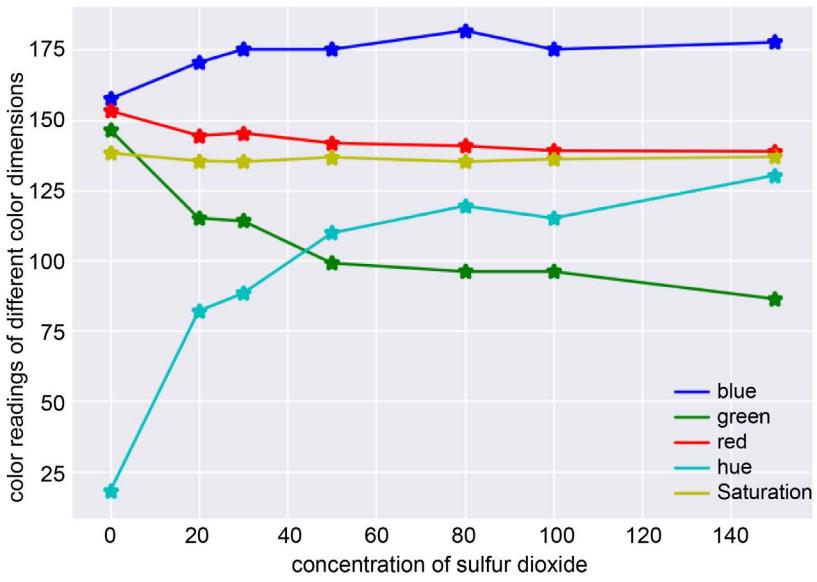

Figure 1. Concentration of SO_2 and the relative color readings.

Figure 2) and the correlation coefficient table of each variable (see Table 2). The Figure 2 and Table 2 all shows that not only the concentration of sulfur dioxide is strongly linearly related to each color variable except saturation, but there are also significant linear relationships within the color variables. Therefore, we can use stepwise regression method to screen and eliminate the variables that cause multiple collinearity, thus establishing the regression model.

In this table, the correlation coefficients are calculated by the following formula:

$$r(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{1}$$

Here, $Cov(X,Y)$ is the covariance, $Var(X),Var(Y)$ is the variance of X and Y respectively.

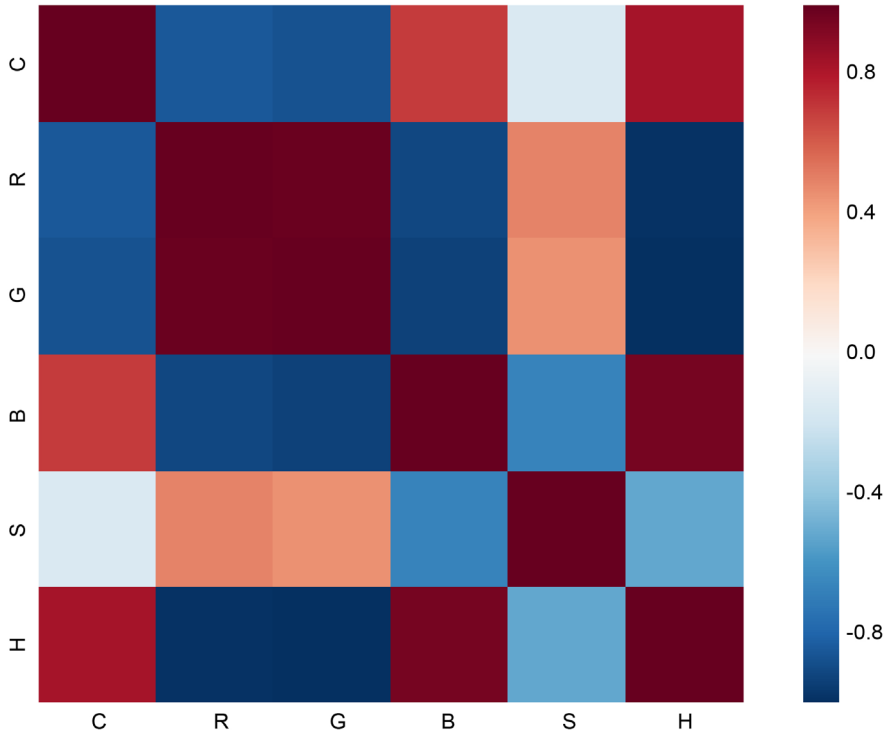

Figure 2. The correlation coefficient heat map.

Table 2. Ordinary correlations.

	C	*H*	*B*	*S*	*R*	*G*
C	1.00					
H	0.83	1.00				
B	0.70	0.96	1.00			
S	−0.15	−0.52	−0.67	1.00		
R	−0.84	−0.98	−0.91	0.49	1.00	
G	−0.87	−1.00	−0.93	0.45	0.99	1.00

4. Introduction of Stepwise Regression

In statistics, stepwise regression [3] is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some pre-specified criterion. The main approaches include forward selection, backward elimination, and bidirectional elimination.

Forward selection, which involves starting with no variables in the model, testing the addition of each variable using a chosen model fit criterion, adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit, and repeating this process until none improves the model to a statistically significant extent.

Backward elimination, which involves starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if any) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss of fit.

Bidirectional elimination, a combination of the above, testing at each step for variables to be included or excluded.

We can use the algorithm flowchart [3] in **Figure 3** to understand the bidirectional elimination.

5. Modeling

In this paper, we use the bidirectional elimination stepwise regression to determine a final model in MATLAB [4]. At each step, the method searches for terms to add to or remove from the model based on the value of the "Criterion" argument. That is to say, if a term is not currently in the model, the null hypothesis is

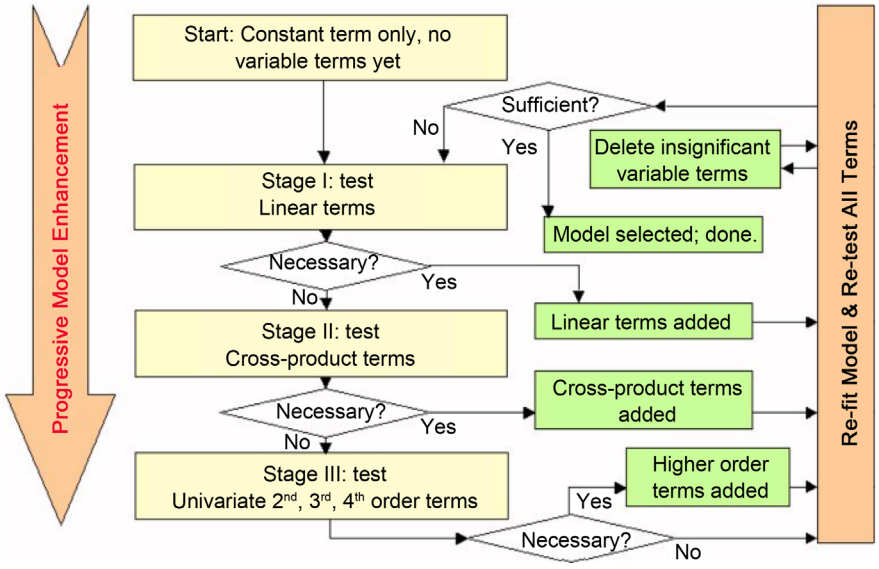

Figure 3. The algorithm flowchart of bidirectional elimination.

that the term would have a zero coefficient if added to the model. If there is sufficient evidence to reject the null hypothesis, for example, the *p*-value for an *F*-test of the change in the sum of squared error of the model is smaller than the default value 0.05, add the term to the model the term is added to the model. Conversely, if a term is currently in the model, the null hypothesis is that the term has a zero coefficient. If there is insufficient evidence to reject the null hypothesis, the term is removed from the model.

Using MATLAB function "*stepwiselm*" [4] to achieve the automatic selection process above, the detailed procedure displayed by the program is as follows: (here *G*, *H*, *B*, are color variables in **Table 1**).

1) Adding *G*, *F*-Stat = 69.8123, *p*-Value = 2.01854e−08

2) Adding *H*, *F*-Stat = 24.5393, *p*-Value = 5.89362e−05

3) Adding *G***H*, *F*-Stat = 25.4857, *p*-Value = 5.34767e−05

4) Adding *B*, *F*-Stat = 11.9566, *p*-Value = 0.00248622

We see stepwise algorithm adds color variable *G* (green), *H* (hue), the interaction item *G***H*, and *B* (blue) to the model with respectively the corresponding *p*-values less than default 0.05. That is to say, there is sufficient evidence to reject the null hypothesis, the term is added to the model. The last model is as follows:

$$C = -1565.4 + 20.661(G) - 10.399(B) + 18.763(H) - 0.059739(G \times H) \quad (1)$$

From this model, we can see the concentration of sulfur dioxide is mainly influenced by the linear influence of green, blue, hue, and the interaction between green and hue, which is consistent with the previous data analysis.

In MATLAB function "*stepwiselm*", the *F*-Stat (*F* statistic) of *F*-test and other parameters are calculated by the following formula [5]:

$$\beta = \left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}Y, \quad SSE = \sigma^2 = \frac{1}{n}\left|Y - X\beta\right|^2 \quad (2)$$

$$F\text{-}Stat = \frac{MSR}{MSE} = \frac{SSR/(p-1)}{SSE/(n-p)} \quad (3)$$

$$SSR = \frac{\beta^{\mathrm{T}}X^{\mathrm{T}}X\beta}{n}, \quad SST = SSE + SSR \quad (3)$$

$$R^2 = \frac{SSR}{SST} \quad (4)$$

Here Y is the $n\times 1$ vectors of the response variable, X is $n\times p$ matrix in which the first column are all 1, n is the number of samples, p is the number of predictor variables (including constant term, the interaction item) in each stepwise procedure.

6. Model Analysis and Error Analysis

At last the program displays the model parameters are as follows:

Number of observations: 25, Error degrees of freedom: 20

Root Mean Squared Error (the σ in Formula (1)): 10.4

R^2: 0.967

F-statistic vs. constant model: 146, *p*-value = 1.69e−14

The above parameters indicate that we used all 25 data in Table 1 as sample points to establish the model, the root mean square error is 10.4, which is reasonable because of the concentration differences of sulfur dioxide in Table 1 are big. The adjusted *R* square is 0.96, which indicates that the fitting is very high, in the case of the default *p* value 0.05, *p*-value is far less than 0.05, so the model at significance level $\alpha = 0.05$ is significant and can be used to fit the concentration efficiently. Figure 4 shows the regression curve and the real sample data curve, which shows the regression model can well depict the sample data.

In Figure 5, the circle shows the residual of each case between the sample data and the predict value, and the line is the confidence interval with a confidence level 0.95 for the residual. If the line does not contain zero then the residual is larger than would be expected at the 95% confidence level. This is evidence that the 13th, 14th observations are outliers. From Table 3 we also see the 13th and 14th cases have larger residuals (marked in red).

A way to test for errors in models created by step-wise regression, is to not rely on the model's F-statistic, significance, or multiple R, but instead assess the model against a set of data that was not used to create the model [3] [6]. This is often done by building a model based on a sample of the dataset available (e.g., 70%)—the "training set"—and use the remainder of the dataset (e.g., 30%) as a validation set to assess the accuracy of the model. Accuracy is then often measured as the actual standard error (SE), MAPE, or mean error between the predicted value and the actual value in the hold-out sample [3] [7]. This method is

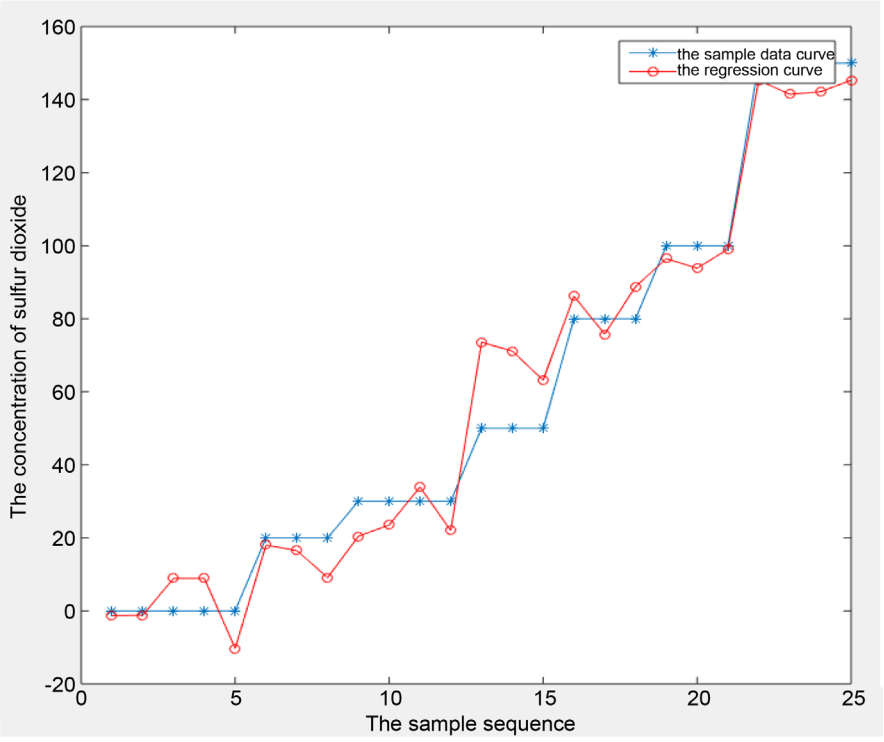

Figure 4. Sample data curve vs regression curve.

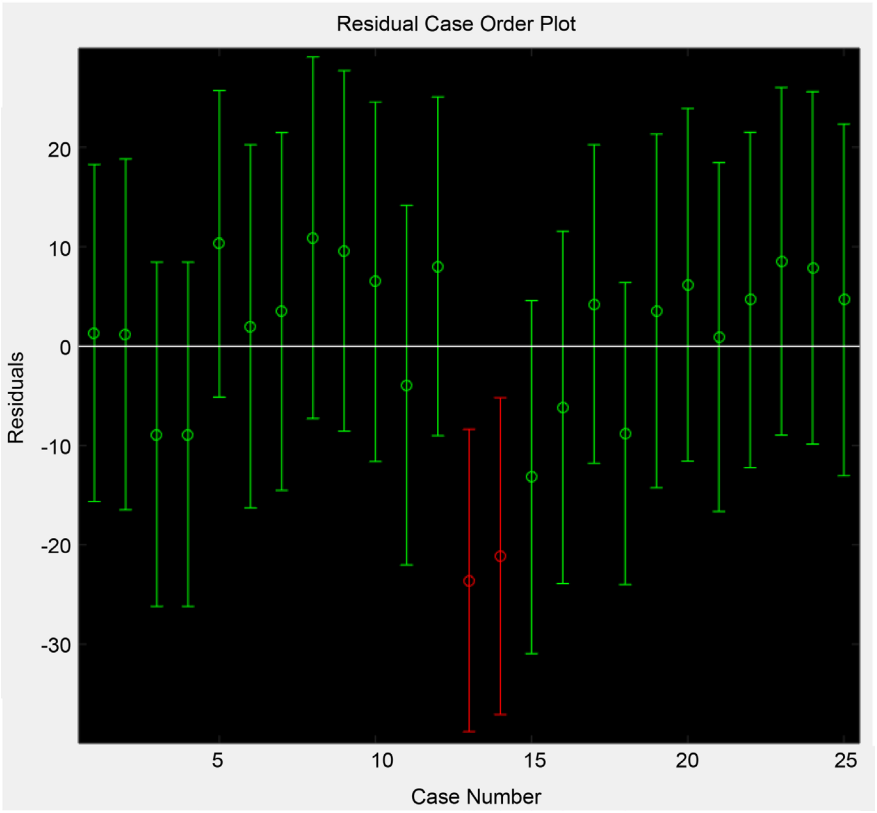

Figure 5. Residual case order plot.

Table 3. Residuals (by columns).

1.31	−21.14
1.17	−13.19
−8.89	−6.18
−8.89	4.22
10.28	−8.80
1.99	3.54
3.48	6.17
10.89	0.91
9.58	4.64
6.47	8.54
−3.92	7.87
8.03	4.64
−23.59	

particularly valuable when data are collected in different settings (e.g., different times, social vs. solitary situations) or when models are assumed to be generalizable. However, in this article, there are only 7 different concentrations of data in Table 1, and there are few other concentrations, so a regression model with higher fitting degree is what we need. So we don't take account into the assess-

ment of the above.

7. Conclusions

In addition to using the stepwise regression method to establish a regression model of color variables and the concentration in Table 1, we also considered several other nonlinear regression models, but the root mean square error is bigger; the fitting degree is not as good as stepwise regression method. In fact, in the previous data analysis, we have pointed out the reasons, mainly because of the significant linearity of the data itself and the collinearity.

Using mathematical models to establish the relationship between color variables and material concentration is of practical value in the study. Combining with modern photography techniques, the material concentration can be predicted through mathematical model relative quickly and accurately. Of course, in practice, we first need to analyze the characteristics of the data itself, and then choose a more reliable model from multiple models to fit and predict, and attention is paid to the relationship between equilibrium fitting and prediction.

References

[1] Concepts, L. (2018) The Definition of Colorimetry. https://en.wikipedia.org/wiki/Colorimetry

[2] Case data, L. (2017) the Concentration of Sulfur Dioxide and the Corresponding Color Readings. http://www.mcm.edu.cn/html_cn/node/460baf68ab0ed0e1e557a0c79b1c4648.html

[3] Concepts, L. (2018) The Definition of Stepwise Regression. https://en.wikipedia.org/wiki/Stepwise_regression

[4] Concepts, L. (2018) The Stepwise Regression Used in MATLAB. https://cn.mathworks.com/help/stats/stepwiselm.html?requestedDomain=true

[5] Concepts, L. (2018) Applied Statistics. http://www.docin.com/p-1712300058.html

[6] Mark, J. and Goldberg, M.A. (2001) Multiple Regression Analysis and Mass Assessment: A Review of the Issues. *The Appraisal Journal*, **1**, 89-109.

[7] Rencher, A.C. and Pun, F.C. (1980) Inflation of R^2 in Best Subset Regression. *Technometrics*, **22**, 49-54. https://doi.org/10.2307/1268382