Model of Colour Value and Substance Concentration Based on Ridge Regression

Zhonghua Ling 1, Yushan Gao 2 and Luling Duan 1, *

1 School of Mathematics and Information Science, Guangxi College of Education, Guangxi, China
2 Nanning Number Two High School, Guangxi, China

*Corresponding author e-mail: duanluling2006@163.com

Abstract. Based on the data of colour value and substance concentration, the quantitative relationship between colour reading and substance concentration in digital photos is studied in this paper. Correlation coefficient, variance inflation factor (VIF) and condition indices are used to analyse the collinearity of variables, and the regression model of colour value and material concentration of digital photos is established by ridge regression method. The ridge trace shows that the least square regression coefficient is very unstable at high collinearity. As the bias parameter increases from 0 to 0.0477, the ridge regression coefficient becomes stabilized and the model's coefficient of determination does not decrease significantly. The maximum VIF (k) is 3.5161 (less than 10), which indicates that the coefficients of the model constructed by the ridge regression method is more robust, and the prediction effect of the model is better than that estimated by the least square method.

Keywords: Ridge regression; Collinearity; Substance concentration

1. Introduction

Colorimetry is a commonly used method to detect the concentration of substances, but because of the sensitive difference of each person's colour and the observation error, the accuracy of this method is greatly affected. According to the study, the colour of the solution of the chemical substance will change from light to deep with the increase of the concentration of the substance [1], so it is considered to measure the concentration of the substance by detecting the colour. With the improvement of photographic technology and colour resolution, the method of using colour reading to detect material characteristics has been widely used, for example, using the colour degree of digital camera photos to establish regression model to detect the nitrogen content crop leaves [2, 3, 4], using RGB digital images to measure the concentration of melanin, oxygen-containing blood and anoxic blood in skin tissue [5]. Therefore, we hope to establish a quantitative relationship between the colour values of digital photos and the concentration of substances, and hope that the concentration of the substances to be tested can be obtained by entering the colour values in the photos.

In regression analysis, the predictor variables are strongly interrelated that the regression results are ambiguous. It is impossible to estimate the unique effects of each variable in the regression equation. The estimated values of the coefficients are very sensitive to slight changes of the data and to the addition
or deletion of the variables in the equation. The estimation of regression coefficient has large sampling error, which will affect the inference and forecasting based on regression model [6]. So, eliminating multiple collinearity in predictor variables is an important part of parameter estimation in regression analysis. Ridge estimation is a common method to eliminate collinearity. In this paper, a mathematical model of colour value and material concentration identification is established by ridge regression method.

2. Materials and Methods

2.1. Data

Data from the 2017 National College students Mathematical Modeling Competition C questions see Table 1. The data consist of variables that Sulfur dioxide concentration(C), colour value of green (G), colour value of red (R), value of hue (H) and value of saturation (S).

| Numble | C   | R   | G   | B   | S   | H   | Numble | C   | R   | G   | B   | S   | H   |
|--------|-----|-----|-----|-----|-----|-----|--------|-----|-----|-----|-----|-----|-----|
| 1      | 0   | 153 | 148 | 157 | 138 | 14  | 14     | 50  | 141 | 99  | 174 | 137 | 109 |
| 2      | 0   | 153 | 147 | 157 | 138 | 16  | 15     | 50  | 142 | 99  | 176 | 136 | 110 |
| 3      | 0   | 153 | 146 | 158 | 137 | 20  | 16     | 80  | 141 | 96  | 181 | 135 | 119 |
| 4      | 0   | 153 | 146 | 158 | 137 | 20  | 17     | 80  | 141 | 96  | 182 | 135 | 119 |
| 5      | 0   | 154 | 145 | 157 | 141 | 19  | 18     | 80  | 140 | 96  | 182 | 135 | 120 |
| 6      | 20  | 144 | 115 | 170 | 135 | 82  | 19     | 100 | 139 | 96  | 175 | 136 | 115 |
| 7      | 20  | 144 | 115 | 169 | 136 | 81  | 20     | 100 | 139 | 96  | 174 | 136 | 114 |
| 8      | 20  | 145 | 115 | 172 | 135 | 83  | 21     | 100 | 139 | 96  | 176 | 136 | 116 |
| 9      | 30  | 145 | 114 | 174 | 135 | 87  | 22     | 150 | 139 | 86  | 178 | 136 | 131 |
| 10     | 30  | 145 | 114 | 176 | 135 | 89  | 23     | 150 | 139 | 87  | 177 | 136 | 129 |
| 11     | 30  | 145 | 114 | 175 | 135 | 89  | 24     | 150 | 138 | 86  | 177 | 136 | 130 |
| 12     | 30  | 146 | 114 | 175 | 135 | 88  | 25     | 150 | 139 | 86  | 178 | 137 | 131 |
| 13     | 50  | 142 | 99  | 175 | 137 | 110 | 14     | 50  | 141 | 99  | 174 | 137 | 109 |

2.2. Ridge regression method

Multiple linear regression model can be expressed as:

\[ Y = X\beta + \epsilon, \]

Where \( Y \) is \( n \times 1 \) vector of response variables, \( X \) is \( n \times p \) matrix of predictor variables, \( \beta \) is \( p \times 1 \) vector of regression coefficients, \( \epsilon \) is \( n \times 1 \) vector of random error and \( E(\epsilon) = 0, \Var(\epsilon) = \sigma^2 I_n \). The least squares estimator \( \hat{\beta} \) of \( \beta \) can be written explicitly as

\[ \hat{\beta} = (X^T X)^{-1} X^T Y. \]  

(1)

It will be convenient to assume that the predictor variables and the response have been centered and scaled to unit length. At this time, the established regression model \( Y = X\beta + \epsilon \) has no constant term, \( X^T X \) is \( p \times p \) matrix of coefficients between the predictor variables, \( X^T Y \) is \( p \times 1 \) vector of correlations between predictor variables and response variable. The total mean square of regression coefficient estimation is expressed as:

\[ E[(\hat{\beta} - \beta)^T (\hat{\beta} - \beta)] = \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}, \]  

(2)

\( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0 \) are the eigenvalues of \( X^T X \) [6]. If there is collinearity in the predictor variable, the determinant \( \det(X^T X) \) will be very small, so that at least one of the eigenvalues will be small,
and (2) implies that the total mean square error of estimator $\hat{\beta}$ may be very large, suggesting imprecision in the least squares estimation method.

To deal with the problem that collinearity leads to the instability of regression coefficient, Hoerl and Kennard [7] put forward ridge regression. The basic idea is to add a diagonal matrix to the matrix $X^TX$, so that the matrix $X^TX$ does not have smaller eigenvalues and the stability and reliability of parameter estimation can be improved. The ridge regression estimators of regression coefficient $\beta$ are [6]

$$\hat{\beta}(k) = (X^TX + kI)^{-1}X^TY = (X^TX + kI)^{-1}X^TX\hat{\beta}$$

(3)

Where $k$ is a Bias parameter, $0 < k < 1$. Because $E[\hat{\beta}(k)] = (X^TX + kI)^{-1}X^TX\beta$, $\hat{\beta}(k)$ is a biased estimator of $\beta$.

Hoerl and Kennard (1970) prove that there exists a value of $k > 0$ such that

$$E[(\hat{\beta}(k) - \beta)^\top(\hat{\beta}(k) - \beta)] < E[(\hat{\beta} - \beta)^\top(\hat{\beta} - \beta)]$$

That is, the mean square error of ridge regression estimation $\hat{\beta}(k)$ is less than that of least square estimation $\hat{\beta}$.

One of the key steps of ridge regression is to select the value of the appropriate bias parameter $k$. Hoerl and Kennard[7] suggested that the value $k$ be determined by observing the ridge trace. The selection principle is to find the minimum $k$ value that makes $\hat{\beta}(k)$ stable, they proposed fixed point method and iterative method in 1975 and 1976, respectively[6].

2.3. Collinear detection methods

The main indicators for measuring data collinearity are variance inflation factor and the condition indices.

2.3.1. Variance inflation factor. Supposing predictor variables $X = (X_1, X_2, \ldots, X_p)$, the variance inflation factor for $X_j$ is defined as

$$VIF_j = \frac{1}{1-R_j^2}, j=1, L, p,$$

(4)

Where $R_j^2$ is the coefficient of determination of $X_j$ obtained for regression of the remaining $p-1$ predictor variables. It is suggested that when the value $VIF$ exceeds 10, the collinear phenomenon adversely affect the estimation of coefficients [6].

2.3.2. The Condition Indices. If there is collinearity in the predictor variables, the determinant $|X^TX|$ will be very small, so that at least one of the eigenvalues will be small. Supposing the eigenvalue of $X^TX$ is $\lambda_1 \geq \lambda_2 \geq L \geq \lambda_p$, the conditional indices defining the correlation matrix $X^TX$ are as follows

$$\kappa_j = \sqrt{\frac{\lambda_j}{\lambda_p}}, j=1, L, p$$

(5)

If $\kappa_j$ is more than 15, there is a strong correlation between the data, and when $\kappa_j$ is more than 30, corrective action must be taken to eliminate the influence of collinearity [6].
3. Results

3.1. Collinear analysis of data
We use SPSS software to calculate the correlation coefficient between the colour value and the concentration data, see Table 2, calculate the eigenvalues of the correlation matrix, and obtain the variance expansion factor and the condition index according to (4) and (5). The results are listed in Table 3 and Table 4. It can be seen from Table 3 that, except for value of saturation, the VIF values of other variables all exceed 10, and the maximum value is 1259.666. There are two conditional indexes in Table 4 that exceed 15, and one of them exceeds 30, which indicates the data has a strong collinearity. The influence of collinearity on the model must be eliminated initially.

### Table 2. Correlation coefficients of concentration and colour values data

|     | C   | R   | G   | B   | S   | H   |
|-----|-----|-----|-----|-----|-----|-----|
| C   | 1.000 | -0.844 | -0.867 | 0.696 | -0.150 | 0.830 |
| R   | -0.844 | 1.000 | 0.987 | -0.909 | 0.492 | -0.984 |
| G   | -0.867 | 0.987 | 1.000 | -0.928 | 0.454 | -0.996 |
| B   | 0.696 | -0.909 | -0.928 | 1.000 | -0.667 | 0.956 |
| S   | -0.150 | 0.492 | 0.454 | -0.667 | 1.000 | -0.520 |
| H   | 0.830 | -0.984 | -0.996 | 0.956 | -0.520 | 1.000 |

### Table 3. Variance inflation factors of colour values data

| Variable | R   | G   | B   | S   | H   |
|----------|-----|-----|-----|-----|-----|
| VIF      | 65.745 | 804.959 | 69.320 | 5.043 | 1259.666 |

### Table 4. Condition indices of colour values data

| Numble | 1     | 2     | 3     | 4     | 5     |
|--------|-------|-------|-------|-------|-------|
| Eigenvalues | 0.00049 | 0.00686 | 0.06534 | 0.69547 | 4.23185 |
| Condition indices | 92.961 | 24.831 | 8.048 | 2.467 | 1.000 |

3.2. Ridge regression model
We centralize and unitize variables $C, R, G, B, S, H$ to $\tilde{C}, \tilde{R}, \tilde{G}, \tilde{B}, \tilde{S}, \tilde{H}$, we establish a standardized model

$$\tilde{C}' = \theta_0 \tilde{R} + \theta_1 \tilde{G} + \theta_2 \tilde{B} + \theta_3 \tilde{S} + \theta_4 \tilde{H} + \epsilon'$$

For $k \in [0, 1]$, the ridge estimated coefficients can be calculated with formula (2) $\hat{\theta}(k)$ = $(\hat{\theta}_1(k), \hat{\theta}_2(k), \hat{\theta}_3(k), \hat{\theta}_4(k), \hat{\theta}_5(k))$. We calculate $\hat{\theta}(k)$ taking the value from 0 to 0.1 every 0.002, and the data are shown in figure 1.
When \( k = 0 \), the parameter estimation method of the model is the least square method and the coefficient of determination is \( R^2 = 0.8996 \). It can be seen from the ridge trace that when the value \( k \) is small, the value \( \hat{\theta}(k) \) is unstable, \( \hat{\theta}_1(k) \) changes from positive to negative, \( \hat{\theta}_2(k) \) changes from negative to positive.

When \( k \) increases to 0.04, the estimated values of each regression coefficient are stable. We use the iterative method to determine the value \( k \). The values of the 14th, 15th and 16th iterations are 0.04763, 0.04770, 0.04772, respectively, which shows little difference of \( k \) between the iteration values. When \( k = 0.0477 \), the coefficient of determination of regression model is \( R^2 = 0.8248 \), and does not decrease significantly. Each predictor variable VIF\((k)\) is reduced to 2.6901, 0.0163, 3.5161, 1.3751 and 0.5224 (less than 10). Therefore, the range of \( k \) values (0.04, 0.0477) is reasonable.

When \( k = 0.0477 \), the standardized model is

\[
C = -0.3267R - 0.5191G - 0.1220B + 0.2774S + 0.2410H.
\]

The model transformed into the original variable is

\[
C = -634.9980 - 3.2223R - 1.2882G - 0.7832B + 10.3179S + 0.3137H.
\]

4. Conclusion
Collinear phenomena can seriously affect inference and prediction in regression analysis. In this paper, a digital photo colour reading and substance concentration model is established by ridge regression method. The ridge trace map shows that the estimated value of least square regression coefficients are very unstable under high collinearity, which affects the accuracy of model prediction. We use the iterative method to determine a value of 0.0477 combining ridge map and expansion factor of predictive variables. It can be seen from the ridge trace map that the ridge regression coefficients stabilize when \( k \) increases from 0 to 0.0477. When \( k = 0.0477 \), the maximum VIF\((k)\) is 3.5161 (less than 10). This shows that the coefficient of the model constructed by the ridge regression method is more robust, and the prediction effect of the model is better than that of the model estimated by the least square method.

Acknowledgments
This work was financially supported by the Basic Ability Improvement Projects for Young and Middle-aged Teachers from Colleges of Guangxi in 2019 (NO.2019KY1679, 2019KY1681), the Subproject of Bagui Scholars Innovation Team of Guangxi College of Education and the Guangxi College of Education (No. A2017001).
References

[1] W. J. H, Research on measurement of Cd(II) concentration based on color detection of RGB image, Transducer & Microsystem Technologies. 35 (2016) 17-19.

[2] Y. Wang, D. Wang, P. Shi, Estimating rice chlorophyll content and leaf nitrogen concentration with a digital still color camera under natural light, Plant methods. 10 (2014) 36.

[3] E. R. Hunt Jr, P.C. Doraiswamy, J.E. McMurtrey, A visible band index for remote sensing leaf chlorophyll content at the canopy scale, International Journal of Applied Earth Observation and Geoinformation. 21 (2013) 103-112.

[4] Q. Niu, H. Feng, C. Li, “Estimation of leaf nitrogen concentration of winter wheat using uav-based RGB imagery”, in IFIP International Federation for Information Processing 2019, edited by D. Li and C. Zhao. International Federation for Information Processing ,Springer, Cham, 2017, pp. 139-153.

[5] I. Nishidate, K. Sasaoka, T. Yuasa, Visualizing of skin chromophore concentrations by use of RGB images, Optics letters. 33 (2008) 2263-2265.

[6] S. Chatterjee, A.S. Hadi, Regression analysis by example, fifth ed., John Wiley & Sons Inc., New Jersey, 2013.

[7] A.E. Hoerl, R.W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics. 12 (1970) 55-67.