A Regression Equation Model for Height and Weight Prediction

Xu ZHANG¹,², Chao ZHAO², Hai-tao WANG ², Fang ZHANG², Jing ZHAO² and Gang WU ²,*

¹School of Mathematical Sciences, Capital Normal University, Beijing, China
²China National Institute of Standardization, Beijing, China
wugang@cnis.gov.cn

Abstract. Height and weight prediction has been a popular problem in ergonomics study. In this paper, we reduce the dimension by principal component analysis and choose the best regression equation using various statistical criterion such as Residual Mean Square(RMSq), Mallow Cp and Akaike information criterion(AIC). Finally, compared with the real value, we analyze the fitting accuracy of the regression equation we proposed.

1. Introduction
Regression analysis method is the most widely used in multivariate statistical analysis. Regression analysis is widely used in social, economic, scientific and technological fields of data analysis, the establishment of empirical formula for regular forecasting, such as weather forecasting, earthquake prediction, stock market analysis and so on.

Height and weight inference is an important part of individual identification work in forensic anthropology. In the actual case detection process, height and weight prediction play a key role in finding the dead source as well as detecting the unknown body, albino bone entity, broken corpse and other cases. In previous studies, the height and weight were predicted with a single variable such as foot length, head circumference and shoulder width. Some articles use the neck data, head data for multiple regression.

The primary purpose of this research is to establish a method between the body data with the height and weight. This paper not only introduces the prediction of height and weight by one variable, but also gives the optimal combination of multiple variables when they are fitted. These equations have a great reference value for forensic science and anthropology.

2. Related work
The physical characteristics of the body are an important part of anthropological research. In forensic personal identification, it is important to speculate on height and weight. The more accurately we speculate height and weight, the more quickly we identify the identity of the deceased. Thereby it can shorten the detection time of the case and improve the efficiency of detection. At present, domestic research on these aspects is rare, especially about weight prediction.

We focus on forensic anthropology in height and weight prediction research. At the scene of the incident, we can find many data of the body. It becomes more critical that how to do more accurate height and weight prediction. When the conditions are not allowed, we get limited data. This paper first carries out regression prediction based on a single factor. When we get a lot of data, how to
choose a more appropriate variable is what we need to consider. The second part of this article focuses on how to choose variables.

3. Methodology

Total 2000 participants were recruited for this experiment. Participants were between 18 and 60 years old, without any deformity of body.

| items                              | n   | minimum | maximum | average   | SD    |
|------------------------------------|-----|---------|---------|-----------|-------|
| height(mm)                         | 2000| 1496    | 1893    | 1677.605  | 58.9680|
| weight(hg)                         | 2000| 370     | 880     | 596.667   | 80.5707|
| armreach from back(mm)             | 2000| 712     | 956     | 833.698   | 35.4544|
| hand length(mm)                    | 2000| 153     | 212     | 182.845   | 8.0007 |
| foot length(mm)                    | 2000| 212     | 283     | 246.726   | 10.3434|
| lower extremity length(mm)         | 2000| 834     | 1128    | 992.421   | 43.4268|
| neck circumference(mm)             | 2000| 296     | 428     | 352.664   | 18.0861|
| bust(mm)                           | 2000| 718     | 1085    | 871.142   | 54.0766|
| waist circumference(mm)            | 2000| 571     | 1036    | 748.869   | 76.2112|
| head width(mm)                     | 2000| 512     | 615     | 561.132   | 15.3977|

In the prediction of height and weight, it is often necessary to establish a model of the relationship between variables and height and weight. For the above data to simulate, we find that the variance expansion factor of the explanatory variables is quite large. It indicates that there is a more serious multiple collinearity between the explanatory variables. The least squares regression shows that the sign of the partial regression coefficients is negative and very unstable. At the same time, although the F-test can pass, but most of the regression coefficient estimation t-test was rejected, and the mean square error is as high as 29.76.

When the multiple collinearity of the observed multiple matrices occur is too high, the mean square error of the regression coefficient will be too large. That will affect the goodness of the model fitting. If we can compress the relevant variables into a few independent variables which can reflect the original number of variable information, we will avoid the above problems. Principal component analysis is an effective way to convert multiple variables into a few independent variables by means of linear combinations, which provides convenience for subsequent analysis.

Principal component analysis is implemented in the R software through the princomp function. The principal component analysis of the nine variables is taken as the principal component analysis. We get the result that the total variance contribution rate is 87.1%, the maximum value is 3.7985 and the minimum is 0.1982. The four principal components are used as the new variable values for further analysis. For the convenience of the description below, the four principal components are referred to as limb data.
3.1. The relation between human height with limb data

In the study of the predicted model, the simplest and commonly method is that the two characteristic parameters of the system are closely related to the linear relationship. For such model, it is generally use a linear regression method—the least squares method. For the data that basically conforms to the linear relationship, the least squares method is to obtain the coefficients \( a \) and \( b \) of the regression line with minimize the distance between the regression line and the point of the hash in the Y direction. Given \( n \)-th column \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\), the regression equation is

\[
y = bx + a.
\]

We use \(\sum_{i=1}^{n} [y_i - (a + bx_i)]^2\) to quantitative description the point in the y direction to the straight distance of the straight line. So it can be seen as a binary function

\[
Q(a, b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2.
\]

So the question that finding a straight line that is the most close to the \( n \) points of the problem, into the problem that finding two numbers \( \hat{a}, \hat{b} \) that can make the binary function \( Q(a, b) \) to a minimum. Derived by the formula, we have

\[
b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})}.
\]

where \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i; \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i; a = \bar{y} - b\bar{x} \).

Scatter plot shows the dependent variable with the independent variable and the general trend, which can choose the appropriate function of the data points to fit. We draw the scatter plot of height with other factors based on the data.
By calculating the correlation coefficient, we found that height with limb data were highly correlated and they have a proportional relationship. According to the independent variable, we respectively estimated linear regression equation. The equation has extremely significant statistical significance by t-test and F-test.

Table 2 Regression equation of height and the related detection value

| items                  | height Y                                      | T value | Adjusted R-squared | F-statistic |
|------------------------|-----------------------------------------------|---------|--------------------|-------------|
| armreach from back      | $Y = 1.1322X + 733.69265$                     | 41.54   | 0.4631             | 1725        |
| hand length            | $Y = 4.8645X + 788.1596$                     | 39.27   | 0.4353             | 1524        |
| foot length            | $Y = 3.92466X + 709.28839$                   | 42.42   | 0.4737             | 1800        |
| lower extremity length | $Y = 1.09538X + 590.52738$                   | 61.02   | 0.6506             | 3723        |

Compare the adjusted R-squared in the table, $0.6507 > 0.4739 > 0.4631 > 0.4356$, we get the conclusion that it is better to establish the regression equation according to the lower extremity length. The range of the estimated error is smaller lower extremity. Therefore, the priority order used to estimate the height of the indicators should be the lower extremity length.

When we choose all the variables, the regression equation is

$$ Y = 339.54799 + 0.27558X_1 + 0.7695X_2 + 1.06205X_3 + 0.71095X_4 $$  \hspace{1cm} (4)

and the adjusted R-squared rise to 0.7323. Compared to the prediction of lower extremity length, the accuracy has improved greatly.

3.2. Prediction weight by optimal subsidence regression

Now, we use the same method to do weight analysis on the four variables. We draw the scatter plot of weight with other factors based on the data.
Fig. 3 The scatter plot of weight with other factors

Table 3 Regression equation of weight and the related detection value

| items X               | weight Y                     | T value | Adjusted R-squared | F-statistic |
|-----------------------|------------------------------|---------|--------------------|-------------|
| armreach from back    | Y=1.1308-346.0583            | 25.641  | 0.2472             | 657.5       |
| hand length           | Y=3.9348X-122.7905           | 18.974  | 0.1522             | 350         |
| foot length           | Y=3.3096X-219.8888           | 20.979  | 0.1801             | 440.1       |
| lower extremity length| Y=0.74869X-146.34772         | 19.714  | 0.1624             | 388.6       |

We find that the adjusted R-squared are less than 0.3. It indicates that in the case of a single variable, the weight of the fitting effect is not very good. So we introduce all the variables to fit the regression. It can be seen from the regression equation that the coefficient of X2 is not significant. The next step is to select the variable so that all variable coefficients are significant.

First, we can compute the independent variables of all possible subsets and the corresponding $RMS_q$ and $C_p$ statistic values. It's defined by

$$RMS_q = \frac{1}{n-p} \cdot RSS_q,$$  \hspace{1cm} (5)

where $RSS_q = y'(I - X_q(X_q'X_q)^{-1}X_q')y$ is the robust residual sum from a robust fit of the selected model.

We define. Mallow’s Cp is a technique for model selection in regression (Mallow’s1974). The Cp Statistic is defined as a criterion to assess fits when models with different number of parameters are being compared. It is given by

$$C_p = \frac{RSS_q}{\sigma^2} - (n - 2q).$$  \hspace{1cm} (6)
Table 4 The \( RMS_q \) and \( C_p \) values for the regression equation

| Variable subset | \( C_p \) | \( RMS_q \) |
|-----------------|-----------|-------------|
| \( X_2 \)       | 981.9697  | 7015.616    |
| \( X_3 \)       | 3.53E+33  | 8.31E+33    |
| \( X_1 \)       | 3.62E+33  | 8.52E+33    |
| \( X_4 \)       | 3.97E+33  | 9.36E+33    |
| \( X_1,X_2 \)   | 3.561356  | 4708.288    |
| \( X_2,X_3 \)   | 445.0235  | 5748.823    |
| \( X_3,X_4 \)   | 516.1024  | 5916.357    |
| \( X_1,X_3 \)   | 3.36E+33  | 7.92E+33    |
| \( X_1,X_4 \)   | 3.49E+33  | 8.22E+33    |
| \( X_2,X_4 \)   | 3.57E+33  | 8.42E+33    |
| \( X_1,X_2,X_4 \)| 3.543942  | 4705.89     |
| \( X_1,X_3,X_4 \)| 5.517213  | 4710.543    |
| \( X_2,X_3,X_4 \)| 340.461   | 5500.407    |
| \( X_1,X_2,X_3 \)| 3.36E+33  | 7.92E+33    |
| \( X_1,X_2,X_3,X_4 \)| 5      | 4706.965    |

According to the properties of \( RMS_q \) and \( C_p \), we choose the subset of independent variables that is based on the principle of the smaller of the value. From the table we can see that in the subset of independent variables without \( X_3 \), both \( RMS_q \) and \( C_p \) criteria choose \( \{ X_1,X_2,X_4 \} \) as the optimal subset. Thus, the optimal subset that can be used for prediction is

\[
Y = -472.24592 + 0.83625 X_1 + 0.8322 X_2 + 0.22124 X_4. \tag{7}
\]

3.3. Prediction weight by stepwise regression

We can use stepwise regression method for variable selection. The stepwise regression of the multivariate regression equation is based on the least squares principle. It is used to remove the factors that have little or no effect on the dependent variable. The significant factor is selected and the optimal regression model is obtained. Since Akaike(1978) put forward the AIC criterion (Akaike Information Criterion) from the principle of maximum likelihood of information theory and promotion, the method of successive test hypothesis in model selection has gradually been replaced. The AIC criterion is a generalization of the principle of great likelihood. Let the parameter vector below the k-th model \( M_k \) fall within the parameter space \( \Theta_k \), denote the likelihood function with \( L(\theta) \), and let

\[
L_k = \sup_{\theta \in \Theta_k} L(\theta), k = 0, \cdots, p - 1. \tag{8}
\]

According to the AIC criterion, we use the model \( M_q \) selected by the following formula to take care of \( M_q \):

\[
\hat{q} = \text{Arg min } \{ \text{AIC}(k) : k = 0, \cdots, p - 1 \}, \tag{9}
\]

where \( \text{AIC}(k) = -\log L_k + v(k) \). Here, \( v(k) \) represents the number of free parameters to be evaluated within \( \Theta_k \).

Table 5 The result of variable selection using AIC

| The variables in the model | Adjusted R-squared | AIC      |
|---------------------------|--------------------|----------|
| \( X_1,X_2,X_3,X_4 \)    | 0.2735             | 16922.6  |
| \( X_1,X_2,X_4 \)        | 0.2624             | 16951.81 |
| \( X_1,X_3,X_4 \)        | 0.2738             | 16920.68 |

The set of parameters that make AIC reach the minimum is the optimal parameter selection.
Comparing with AIC criteria, the model is better when the variable is \(X_1, X_3\) and \(X_4\). Accordingly, the regression equation is

\[
Y = -524.66568 + 0.7871X_1 + 1.2947X_3 + 0.14681X_4.
\]  
(10)

4. Discussion

In this paper, we establish the equation between the male body data with the height and weight through the 2000 male physical measurement. When we select only one item in the four variables, the best fit variable is lower extremity length. Using the data obtained by a linear regression equation, we make a conclusion that more than 84.2% of people in the upper and lower error within the range of 5cm. When we use all the limb data to measure height, we find that 5 cm error within the accuracy rate increased to 89.5%. Therefore, the more accurate the variable, the more accurate the results of the height forecast. However, weight prediction doesn’t have such conclusion. In the range of 50kg of error, we calculated that using three variables than using all variables get a higher accuracy of the results prediction. Despite the difference in selection variables is very small, we still choose a more appropriate model.

The above equations have their own different applicable conditions. In the actual calculation of height and weight, we select the most appropriate indicators for prediction that is based on the specific circumstances. Thereby, we can narrow the error and get a more reliable estimate. The research object of this study is adult male, so the regression equation established has some limitations compared with the whole people. For the calculation of height and height of female as well as by age to calculate, we should be further exploration and research.

Acknowledgement

This paper is supported by grants from National Key R&D Program of China (2016YFF0204205) and China National Institute of Standardization (712016Y-4941-2016, 522016Y-4681-2016).

References

[1] Wang B, Yan G. Robust RMS\(^\circ\) criterion of variable selection[J]. Journal of First Military Medical University, 1997.

[2] Pinky Baruah. Application of R2, \(R^2\), Cp and Residual Mean Square in Regression, May-2015

[3] Maggino F, Fattore M. Multivariate Statistical Analysis[J]. Psychophysiology, 1973, 10(5):517-532.

[4] Cai T T, Hall P. Prediction in Functional Linear Regression[J]. Annals of Statistics, 2006, 34(5):2159-2179.