Forecast of Development Cost of Missile Equipment Based on Partial Least Squares

Qiaoyang Li*, Guiming Chen, Ziqi Li and Boxuan Fan

Xi'an Research Institute of High-Tech, Xi'an 710025, China

*Corresponding author e-mail: liqiaoyang2019@163.com

Abstract. Due to the complex and diverse influencing factors of the missile equipment development cost forecast and the lack of data, taking into account the partial least squares regression method (PLS) can perform main component analysis, canonical correlation analysis, and multiple regression on variables, and handle multiple linear correlations. The small sample multivariate data has the advantage of high accuracy, and a missile equipment development cost prediction model based on the partial least square method is established. Compared with multiple linear regression, least squares support vector machine and other methods, the results show the effectiveness of the model.

1. Introduction

Due to the special nature of the missile equipment, the models available for cost prediction are very limited. At the same time, there are many variables reflecting the characteristics of the missile and the correlation is complex, and the number of samples used for prediction is significantly smaller than the number of variables. How to accurately predict the development cost of missile equipment has become one of the problems in current cost management. For such problems, more researches are on parameter estimation methods. However, there are usually multiple correlations between various factors that have a significant impact on the development cost of missile equipment. Therefore, it is difficult for the parameter estimation method to effectively predict the development cost of missile equipment. The partial least squares regression method has a great advantage in processing small samples and even poor samples, and can effectively deal with the multiple correlation problems between variables. Therefore, this paper proposes a prediction method for the development cost of missile equipment based on the partial least square method.

2. Basic principles of partial least squares

Suppose the dependent variable $y$ and $k$ independent variables $\{x_1, x_2, \ldots, x_k\}$, of which the number of samples is $n$, constitute the dependent variable vector $\{y, y_2, \ldots, y_n\}$ and independent variable matrix $X=\{x_{i} \}_{i=1}^{n \times k}$. The principal component $t_1$ is extracted in $X$, and $t_1$ is a linear combination of $x_1, x_2, \ldots, x_k$. It is required that $t$ carry the mutated signal in $X$ as much as possible and that it has the greatest correlation with $Y$. In this way, $t_1$ can represent $X$ to the maximum, and it has the strongest interpretation ability for $Y$. 
After the first principal component \( t_1 \) is extracted, the regression of \( Y \) and \( X \) pairs on \( t_1 \) is implemented. If the regression equation has reached satisfactory accuracy at this time, the algorithm stops; otherwise, the residual information after \( X \) is interpreted by \( t_1 \) and \( Y \) is interpreted by \( t_1 \). Residual information is extracted from the second principal component \( t_2 \), and \( Y \) and \( X \) are returned to \( t_1 \) and \( t_2 \). Repeat this process until you can reach a satisfactory accuracy. If a total of \( m \) components \( t_1, t_2, \cdots, t_m \) are extracted for \( X \), the partial least squares method will perform a regression of \( Y \) on \( t_1, t_2, \cdots, t_m \). Since \( t_1, t_2, \cdots, t_m \) is a linear combination of \( x_1, x_2, \cdots, x_k \), it can be finally expressed as a regression equation of \( Y \) on independent variable \( x_1, x_2, \cdots, x_k \).

3. Partial least squares regression modelling

Step1: Standardize the sample \( X \) and obtain the standardized independent variable matrix \( E_0 \) and the dependent variable matrix \( F_0 \). After the standardization process, the center of the collection of sample points coincides with the origin of the coordinates, so that data of different dimensions and different orders of magnitude can be compared. The standardized processing method is as follows:

\[
E_0 = (x^*_ij)_{n \times k}, x^*_ij = (x_{ij} - \bar{x}_j)/s_j
\]

\[
F_0 = (y^*_i)_{n \times 1}, y^*_i = (y_i - \bar{y})/s_y
\]

\[i = 1, 2, \ldots, n; j = 1, 2, \ldots, k\] (1)

In the formula, \( E_0 \) and \( F_0 \) are normalized matrices and vectors of \( X \) and \( y \) respectively, \( x_{ij} \) represents the \( i \)-th sample value of the \( j \)-th variable in the independent variable matrix \( X \), \( \bar{x}_j \) represents the mean value of the \( j \)-th variable \( x_j \) in the independent variable matrix \( X \), \( s_j \) represents the standard deviation of \( x_j \), \( y_i \) represents the \( i \)-th sample value of the dependent variable \( y \), \( \bar{y} \) represents the mean of \( y \), \( s_y \) represents the standard deviation of \( y \), \( x^*_ij \) represents the value normalized by \( x_{ij} \), \( y^*_i \) represents the value normalized by \( y_i \).

Step2: Extract the ingredients. Since \( F_0 \) is just a variable, when extracting a component \( u_1 \) from \( E_0 \), \( u_1 = F_0 c_1 \). At this time \( c_1 \) is a scalar, so \( \| u_1 \| = c_1 = 1 \), that is \( u_1 = F_0 \). Extract a component \( t_1 = E_0 w_1 \) from \( E_0 \), where \( w_1 \) is the first major axis of \( E_0 \), that is \( \| w_1 \| = 1 \), where \( w_1 = E_0^T F_0 / \| E_0^T F_0 \| \). Since \( E_0 \) is a normalized matrix and \( F_0 \) is a normalized vector.

\[
E_0^T F_0 = \left( E_{01}, E_{02}, \ldots, E_{0p} \right)^T F_0 = \left( r(x_1, y), r(x_2, y), \ldots, r(x_p, y) \right)^T
\] (2)

\[
w_1 = E_0^T F_0 / \| E_0^T F_0 \| = 1 / \sqrt{\sum_{j=1}^{k} r^2(x_j, y)} = \left[ r(x_1, y) \right]^{1/2} \quad \cdots \quad \left[ r(x_k, y) \right]^{1/2}
\] (3)

\[t_1 = E_0 w_1 = 1 / \sqrt{\sum_{j=1}^{k} r^2(x_j, y)} \left[ r(x_1, y) E_{01} + r(x_2, y) E_{02} + \ldots + r(x_k, y) E_{0k} \right]
\] (4)
In the formula, \( r(x_j, y) \) represents the correlation coefficient between the variables \( x_j \) and \( y \), \( E_a(i = 1, 2, ..., k) \) represents the \( i \)-th column of \( E_a \). The interpretability of the linear combination coefficient is that if the correlation between \( x_j \) and \( y \) is stronger, the combination coefficient of \( 0^{1h}_{ij} \) in the \( t_1 \) component is larger, and the optimization value of the objective function is \( \|E_0^T F_1\| = \sqrt{\sum_{j=1}^{k} r^2(x_j, y)} \) at this time.

Step3: Implement the regression of \( E_0 \) and \( F_0 \) on \( t_1 \), \( E_0 = t_1 p_1^T + E_{r1} \), \( F_0 = t_1 r_1 + F_{r1} \). The regression coefficient are \( p_1 = E_0^T t_1 / \|t_1\|^2 \) and \( r_1 = F_0^T t_1 / \|t_1\|^2 \), \( r_1 \) is a scalar. Therefore, the residual matrix of the regression equation can be written as \( E_1 = E_0 - t_1 p_1^T \), \( F_1 = F_0 - t_1 r_1 \). Check the convergence at this time, and if the satisfactory accuracy has been reached, proceed to the next step. Otherwise, let \( E_1 = E_0 \) and \( F_1 = F_{r1} \), find the second principal axis \( w_2 \) and the second component \( t_2 \) in the same way, and so on. Among them, it should be noted that when extracting the second component, \( E_0 \) is no longer a normalized matrix, so

\[
\begin{align*}
  w_2 &= E_1^T F_0 / \|E_1^T F_0\| = 1 / \sqrt{\sum_{j=1}^{k} \text{cov}^2 (E_{1j}, y)} \begin{bmatrix} \text{cov}^2 (E_{11}, y) \\ \text{cov}^2 (E_{12}, y) \\ \vdots \\ \text{cov}^2 (E_{1p}, y) \end{bmatrix} \\
  t_2 &= E_1^T w_2 \\
  p_2 &= E_1^T t_2 / \|t_2\|^2 \\
  r_2 &= F_1^T t_2 / \|t_2\|^2 \\
  E_2 &= E_1 - t_2 p_2^T \\
  F_2 &= F_1 - t_2 r_2
\end{align*}
\]  

Step4: When the \( h \)-th \((h = 1, 2, ..., m)\) component is extracted, \( t_h = E_{h-1}^T w_h \). Among them, \( w_h = E_{h-1}^T F_{h-1} / \|E_{h-1}^T F_{h-1}\| \), the regression equation is \( E_{h-1} = t_h p_h^T + E_{r2} \), \( F_{h-1} = t_h r_h^T + F_r \), the regression coefficient are \( p_h = E_h^T t_h / \|t_h\|^2 \) and \( r_h = F_h^T t_h / \|t_h\|^2 \), and the regression equation of \( F_0 \) with respect to \( t_1 \sim t_h \) is \( \hat{F}_0 = r_1 t_1 + r_2 t_2 + ... + r_m t_m \). According to the nature of the partial least squares method, it can be known that \( t_1 \) to \( t_h \) are all linear combinations of \( E_0 \), that is \( t_h = E_{h-1}^T w_h = E_0 \prod_{j=1}^{h-1} (I - w_j p_j^T) w_h = E_0 w_h^* \), where \( I \) is the identity matrix. So \( \hat{F}_0 \) can be written as a linear combination of \( E_0 \), that is \( \hat{F}_0 = r_1 E_0 w_1^* + r_2 E_0 w_2^* + ... + r_m E_0 w_m^* = E_0 (r_1 w_1^* + r_2 w_2^* + ... + r_m w_m^*) \). At this time, \( x_j^* = E_0^T, y^* = F_0 \).
If the recorded regression coefficient of \( x_j \) is \( a_j = \sum_{h=1}^{m} r_{hj} w_{hi}^* \) \( (w_{hi}^* \) is the \( j \)-th component of \( w_{hi}^* \), \( j=1,2,...,k \) ), the regression equation of the standardized variable \( y^* \) with respect to \( x_j \) is:

\[
y^* = a_1x_1^* + a_2x_2^* + ... + a_kx_k^*
\]  
(11)

In addition, let \( a_i = a_i S_i / S_y \). Through the standardized inverse process, the regression equation of the original variable \( y \) with respect to \( x_j \) is:

\[
y = \left( y - \sum_{i=1}^{m} a_i x_i \right) + a_1x_1 + a_2x_2 + ... + a_kx_k
\]  
(12)

Step5: Cross validity analysis can determine the number of components that should be extracted. Let \( y_i \) be the original data and \( t_1, t_2, \ldots, t_n \) are the component extracted in the partial least squares regression process. \( \hat{y}_i \) is the fitted value of the \( i \)-th sample point after regression modeling using all the sample points and \( t_1, t_2, \ldots, t_n \) components. \( y_{i_{-o}} \) is the sample point \( i \) deleted during modeling. After \( t_1, t_2, \ldots, t_n \) component regression modeling, this model is used to calculate the fitted value of \( y_i \). Note:

\[
S_{ss,h} = \sum_{i=1}^{n} \left( y_i - \hat{y}_{i_{-o}} \right)^2
\]  
(13)

\[
S_{SPRESS,h} = \sum_{i=1}^{n} \left( y_i - \hat{y}_{i_{-o}} \right)^2
\]  
(14)

\[
Q^2_h = 1 - \frac{S_{SPRESS,h}}{S_{ss,h-1}}
\]  
(15)

When \( Q^2_h \geq (1 - 0.95^2) = 0.0975 \), the marginal contribution of the \( t_h \) component was considered significant. The introduction of a new principal component \( t_h \) will significantly improve the prediction ability of the model.

4. Case study
Taking the development cost forecast of a certain series of missiles as an example, five indicators that have a large impact on the development cost of the missile are selected for analysis, which are maximum range, maximum overload, kill probability, flight speed, and bomb length. Table 1 lists the data of 8 samples. In order to perform error analysis and test of prediction results on the established model, the first 7 samples in the table are selected as training samples, and the 8th sample is used as test sample.
Table 1. Missile performance parameters and development cost data.

| Model | Maximum range (km) | Maximum overload (g) | Kill probability (%) | Flight speed (Ma) | Bomb length (m) | Cost (10 thousand yuan) |
|-------|--------------------|----------------------|----------------------|-------------------|----------------|------------------------|
| A     | 300                | 29                   | 74                   | 2.1               | 2.90           | 3215                   |
| B     | 350                | 34                   | 78                   | 2.4               | 3.02           | 3455                   |
| C     | 600                | 36                   | 80                   | 3.0               | 3.25           | 5500                   |
| D     | 650                | 37                   | 82                   | 2.5               | 3.36           | 7030                   |
| E     | 850                | 40                   | 90                   | 4                 | 3.72           | 9400                   |
| F     | 430                | 42                   | 88                   | 2.5               | 2.89           | 4216                   |
| G     | 680                | 38                   | 85                   | 3.2               | 3.8            | 6800                   |
| H     | 700                | 40                   | 85                   | 3.5               | 2.16           | 7600                   |

4.1. Judging relevance
Judge the correlation between independent variables of performance index and dependent variable of missile cost. It can be seen from Figure 1 that there is a clear linear relationship between the independent variable $X$ and the missile cost $Y$ dependent variable, indicating that the model is reasonable.

Figure 1. Correlation between dependent and independent variables.

4.2. Specific points
Identify and eliminate singular points. Take $x_1, x_2, x_3, x_4, x_5$ as independent variables and $y$ as dependent variables, extract two components $t_1$ and $t_2$, calculate the variance of $t_1$ and $t_2$, take confidence of 95% and make an ellipse in the plane, as shown in the Figure 2 shown. It can be seen from the figure that all the sample points are in the ellipse, so there are no specific points.
4.3. Extract the principal components

When a principal component is extracted, the cross-validation is 0.819. Since $Q^2 = 0.819 > 0.0975$, continue to extract components. When two principal components are extracted, the cross-validation is 0.733 and the cumulative cross-validation is 0.952. Since $Q^2 = 0.733 > 0.0975$, continue to extract components. When three principal components are extracted, the cross-validation is 0.029 and the cumulative cross-validation is 0.953. Since $Q^2 = 0.029 < 0.0975$, the extraction of the components is stopped. Finally, two principal components were extracted, as shown in Figure 3.

4.4. Analysis of the importance

Analysis of the importance of variable projection. The projected importance values of each independent variable are: 1.24589, 1.21687, 1.01399, 1.23545, and 0.455633. From the perspective of the projected importance value of each independent variable, the importance values of $x_1$, $x_2$, $x_3$, and $x_4$ are all greater than 1, indicating that these four indicators are obviously important for missile costs. The importance value of $x$ is less than 0.5, which indicates that this index has a weak ability to explain the dependent variable, and this index is excluded in the regression analysis.

A new regression is performed by using partial least squares to $x_1$, $x_2$, $x_3$, $x_4$ and $y$, and Step 1 to 3 are performed and the projection importance values of each independent variable are calculated, which are...
1.20637, 0.83, 0.884105, and 1.0364, respectively. At this time, the importance values of all indicators are above 0.8, indicating that the independent variable plays a significant role in explaining the dependent variable, that is, the cost is closely related to each indicator. Achieve the fitted regression equation:

\[ y = -2151.146559 + 6.895179x_1 - 43.88139454x_2 + 21.1770157x_3 + 1377.278687x_4 \]  \hspace{1cm} (16)

The value of the index parameters of the test sample H-type missile was substituted into the above formula, and \( y = 7540.75 \), and the relative error was only 0.78%.

4.5. Precision comparison

Set the kernel function of the least squares support vector machine model \( \text{gam} = 10000 \), \( \text{sig} = 500 \) and the estimated cost of the H-type missile is 68.687 million yuan. The multiple linear regression model predicts that the cost of the H-type missile is 71.134 million yuan. The results are shown in Table 2.

| Method of prediction          | Cost / 10 thousand yuan | Relative error/% |
|-------------------------------|-------------------------|------------------|
| Partial least squares         | 7540.75                 | 0.78             |
| Least squares support vector Machine | 6868.7                 | 9.60             |
| Multiple linear regression    | 7113.724                | 6.40             |

It can be known from Table 2 that the relative error of the prediction result of the least squares support vector machine model is 9.6%, the relative error of the prediction result of the multiple linear regression model is 6.4%, and the relative error of the prediction result of the partial least squares model is 0.78%. It can be seen from the comparison of results that the prediction accuracy of the partial least squares model is higher than that of the least squares support vector machine and the multiple linear regression model. Therefore, it is feasible to use partial least squares method to predict and analyze small sample data such as missile cost.

5. Conclusions

The development of missile equipment is related to national defense security, and the sample information is incomplete and uncertain. This paper overcomes the shortcomings of traditional methods in processing small samples, and builds a model based on partial least squares method to predict the development cost of missile equipment. Partial least squares model can not only predict, but also extract principal components from sample data, which is not available in traditional methods and least squares support vector machine method. This paper gives the specific implementation steps and considerations of the partial least squares method. By comparing the prediction results with the multiple linear regression model and the least squares support vector machine model, it is shown that the prediction based on the partial least squares method has higher accuracy. This provides an effective method for forecasting the development cost of missile equipment.

References

[1] Yakui Liu, Guogang Zhang, Hao Qin, Yingsan Geng, Jianhua Wang, The Prediction of the Dynamic Contact Resistance of Circuit Breaker Based on the Kernel Partial Least Squares, IET Generation Transmission & Distribution. 12 (2017).

[2] Elsohaby I, Windeyer MC, Haines DM, Homerosky ER, Pearson JM, McClure JT, Keefe GP, Application of transmission infrared spectroscopy and partial least squares regression to predict immunoglobulin G concentration in dairy and beef cow colostrum, J Anim Sci. 96 (2018) 771 - 782.

[3] J. Bouvier-Muller, C. Allain, F. Enjalbert, Y. Farizon, D. Portes, G. Foucras, Somatic cell count–
based selection reduces susceptibility to energy shortage during early lactation in a sheep model, Journal of Dairy Science. 101 (2018) 2248 - 2259.

[4] Wei Luo, Lu Liu, Prediction of military UAV development cost based on partial least-square regression method, Journal of Beijing University of Aeronautics and Astronautics. 36 (2010) 667 - 670.

[5] Xiaodan Qian, Fang Li, Jinlu Bian, Application of partial least-squares regression in forecast of warship maintenance cost, Ship Science and Technology. 29 (2007) 98 - 100.

[6] Liyuan Wang, Jilian Guo, Hengxi Zhang, Application of Hierarchical Partial Least Square Regression to Development Cost Prediction of Aircraft, Acta aeronautica et Astronautica Sinica. 30 (2009) 1380 - 1384.

[7] Gaobo Chen, Aircraft Maintenance Support Cost Prediction Model Based on PSO-LSSVM, Computer Application and Software. 27 (2010) 181 - 183.

[8] Changke Wu, Wei Chen, Application of BP network based on partial least square method in ship maintenance cost prediction, Science and Technology Innovation Herald. 33(2011) 107 - 108.

[9] Xinlong Chen, Cost structure and usability analysis of missile, National University of Defense Technology. 2010.

[10] Weiguo Fang, Weining Cai, Cost prediction of Developing Military UAVs Based on Gram-Schmidt Regression, Industrial Engineering Journal. 16 (2013) 29 - 33.

[11] Luís Filipe Sanches Fernandes, António Fernandes, A. R. L. Ferreira, Rui Cortes, A partial least squares - Path modeling analysis for the understanding of biodiversity loss in rural and urban watersheds in Portugal, Science of The Total Environment. 626 (2018) 1069 - 1085.