Dynamic real estate project cost prediction based on SIMCA-P

Tang Xiaodong¹, Zhang Xiaojie*¹
1School of Civil Engineering, Shandong Jianzhu University, Jinan.250101, Shandong, China
1*School of Civil Engineering, Shandong Jianzhu University, Jinan.250101, Shandong, China
Corresponding author’s e-mail: zxj@sdjzu.edu.cn

Abstract. with real estate projects as study cases, this study explored the influencing factors (contract term, project visa, project changes, etc.) for dynamic cost control and performed numerical analysis on these factors. The SIMCA-P software was used to perform partial least squares (PLS) regression on the actual cost and a regression prediction model was built. The difference between the checked cost and that forecasted by the model was compared, the reliability of the mathematical models for the residence projects was verified. The research results show that the explanation capacity of the explanatory variables for the dependent variables reached 0.937, which means that all these before-mentioned factors had significant impacts on dynamic cost control. The average error of the predicted residence cost obtained by the model based on actual data was 0.006582346, indicating the high prediction accuracy and that the PLSs regression method performed well in solving the multiple correlations between independent variables. Therefore, the PLS regression method could provide a good solution for prediction of dynamic cost of real estate projects.

1. Introduction
Dynamic cost is an important part of cost control and runs throughout the whole project. Excellent dynamic cost control can facilitate capital management of enterprises, increase their economic gains, maximize the profit of projects and realize precision cost control. In this study, the dynamic cost analysis was analyzed from three perspectives: project visa, project changes and contract terms.

Project overview: the Northern 2# plot project was taken as the study case. The plot covers a total area of 626,700 m² and belongs to a tourist resort of the city. The construction site takes up 11,948 m² and the building occupies 119,489.58 m². The project was to build 5~7-storied residence buildings, a 2-story public building and underground garages.

2. Building the model
Many methods and technologies have been used to prediction of dynamic cost control, such as gray correlation prediction, regression analysis, combination prediction, neural network prediction, etc. Regression analysis is a widely used prediction method, but in multi-linear regression analysis, multiple correlations occur among independent variables. In prediction of dynamic costs, the sample data are usually finite, and the models built lack stability, leading to large errors. The PLS regression method, however, provides a solution. The SIMCA-P software is combined with the PLS regression method to perform prediction of dynamic cost analysis in this study.
2.1. Data source and model fitting

2.1.1 Data source.

The data used in this study are from the dynamic cost database for the Northern 2# real estate project. Take the residence building programs in the Northern 2# project as an example. Variables including the agreed charges on the contract, the project visa fee, the contract term and the project area were used, and the SICMCA-P software was used for analysis and fitting of variables of calculated costs. The statistics are shown in Tables 1.

| Construction site | Contract indicator (yuan/m²) | Change of indicators (10^4 yuan/m²) | actual cost (yuan) |
|-------------------|-------------------------------|--------------------------------------|-------------------|
| 1                 | 64.649                        | 20                                   | 10955128.22       |
| 2                 | 35.37                         | 132                                  | 5994634.01        |
| 3                 | 31.68                         | 132                                  | 5730475.8         |
| 4                 | 16.56                         | 0                                    | 2806412.11        |
| 5                 | 0.16                          | 0                                    | 27460.22          |
| 6                 | 4.25                          | -26                                  | 719720.44         |
| 7                 | 87.58                         | -128.49                              | 14842414.9        |
| 8                 | 103.62                        | 85                                   | 17562102.8        |
| 9                 | 153.66                        | 85                                   | 26044592.71       |
| 10                | 131.09                        | 85                                   | 22218750.56       |
| 11                | 63.25                         | 45                                   | 10720344.1        |
| 12                | 227.45                        | 58                                   | 38550255.78       |
| 13                | 18.94                         | 40                                   | 3210393.74        |
| 14                | 90.38                         | 129                                  | 15317625.07       |

Note: The contract indicator in the tables refer to the ratio of the actual contract charges and the building area, and the changing indicator is the ratio of the actual changing cost to the building area.

2.2. Model building and optimization

2.2.1. Selection of independent variables.

According to the features of dynamic cost control, three variables, i.e., project visa, project changes and contract term are taken as the independent variables, and the project’s actual cost as the dependent variable. Multivariate linear regression was performed on the actual cost and the major indicators. There are issues worth noting in selection of independent variables: first, the major indicators selected in the aforementioned study should be taken into account and the actual visa conditions of the northern 2# project should be considered in determination of the independent indicators; second, the basic data should be changed according to the actual engineering conditions of the Northern 2# project, the quantification capacity and measurability of the independent variable indicator should be fully considered. Given these two and the major control indicators of independent variables regarding the change patterns of the dynamic cost changes in the northern 2# plot, three independent variables that were chosen were: contract indicator (X1), the project change indicator (X2) and the contract term (X3). The building area was the assistant calculation indicator for final accounting.

2.2.2. Analysis of model parameters.

Two numbers of observation N=9 and N=14 were taken for the PLSs models in the SIMCA-P software. The X independent variable in the two models was 3 and the sum of the dependent variables Y was 1. Via model fitting, the PLS models for the residence dynamic costs were obtained, as shown in Figures 1.
Figure 1. Residence cost PLS model

Table 2 lists the parameter explanations of the model.

| Component | R²X | R²X (cum) | Eigenvalue | R²Y | R²Y (cum) | Q² | Limit | Q² (cum) |
|------------|-----|----------|------------|-----|----------|----|-------|----------|
| 1          | 0.476 | 0.476 | 1.43      | 0.881 | 0.881 | 0.821 | 0.05 | 0.821 |
| 2          | 0.349 | 0.826 | 1.05      | 0.118 | 0.999 | 0.979 | 0.05 | 0.996 |

The parameters of the model fitting are as follows:
- R²X – the changing percentage of the independent variable X in the model;
- R²X (cum) – the percentage accumulation of R²X of a given element;
- Eigenvalue – the eigenvalue is the multiplication of the number of the independent variable X and R²X;
- R²Y – the percentage of the changes of the dependent Y in the model;
- R²Y (cum) – the overall cross verification of the elements, mainly used to verify the cross validity of R²;
- Limit – the threshold value of Q²; if the value equals 1, the element is insignificant;
- Q² (cum) – accumulation of the assigned element Q². Different from R² (cum), Q² (cum) is not simple accumulation. Q² (cum) is not simple accumulation, but means the cross validity of the accumulation modeled with m elements (t1, t2, ..., tm).

Based on the parameter explanation of the SIMCA-P model, the cross validity of the Y variable was analyzed. As Table 3 shows, when the second element was extracted, the Q² of the models were 0.979, respectively, while the Q² (cum) were 0.996, which means that the models reached the expected prediction accuracy, thus the partial least squares elements of the models could be extracted. Figures 2 show the fitting histograms that give the results of the regression model fitting. The parameters show that the fitting results were satisfactory.

2.2.3. Outliers of the recognition model.
The outliers of the model data could be recognized based on the distribution map of the t[1]/t[2] ratio. Figure 2 shows the elliptical distribution of the cost analysis model of the residence, and no outlier was
there. Therefore, the model fitting effect was good and no change was needed. If outliers occurred outside the ellipse, the outliers should be checked and removed according to the actual conditions.

2.2.4. Correlation verification
To verify the linear correlation the residence cost models, PLS correlation analysis was performed. Figures 3 shows the correlation between all Y variables (u1) and X variables (t1) of the models. Whether X and Y are highly correlated is reflected on the scattering points around the diagonal line. The size of the scattering range was the measurement indicator of the changeability. As the figure shows, the residence model shows strong linear correlation.

2.3. Verification of model overfitting
To avoid model overfitting, 200 combination sets of the two models were analyzed. Figure 4 shows that of the residence model. The results show that the prediction results of the original model was effective.
The standards to verify the effectiveness of the model was: all blue R2 values on the left were lower than the initial values on the right, or the blue regression line of Q2 crossed with the vertical axis (left) or was lower than the abscissa. The R2 value indicates the feasibility level of the model. When all green R2 values on the left were lower than the initial values on the right, the original model was proved effective.

![Figure 4. Sequence distribution of the residence cost analysis PLS model](image)

3. Model analysis

3.1. Correlation analysis between the independent variables and the dependent variables

By calling the command Plot/Lists→Lists in the software, we obtained the standard PLS model of the actual cost and the three independent variables.

- Actual residence cost = 0.826639×contract indicator× building area -0.397194×changing indicator× building area -0.51684× contract term (1)

The regression coefficient graph shows the explanation of the independent variables for the actual cost, as Figures 5 show. figures are for the regression equations of the standard data.
3.2. Importance of the independent variables in explaining the dependent variables

The VIP analysis is table. The VIP value could reflect the importance of the independent variables in explaining the dependent variables. The VIP value is the weighted sum of squares of the proportion of each independent variable X. the sum of squares of all VIPs equal the number of terms of in the model. Therefore, the average VIP is 1. If the VIP value exceeds 1, it means the X variable is “important”, and if the value is below 0.5, the variable is considered “unimportant”. The area between 0.5 and 1 is a gray area, and the importance level depends on the size of the data set. The VIP diagram is sequenced from high to low, and shows the confidence interval of the VIP values, which is usually 95%.

Table 3. VIP general list [M1]

| Var ID (Primary)        | M1.VIP[2] | 2.44693 * M1.VIP[2]cvSE |
|------------------------|-----------|--------------------------|
| Contract of indicators | 1.54027   | 0.477501                 |
| contract term          | 0.625204  | 0.444985                 |
| Change of indicators   | 0.486507  | 0.974736                 |

Table 3 shows the VIP values of each independent variable. The importance value of X1 variables in the M1 model exceeds 1, and there is no large difference. In Model 1, the difference between the X2 variable and X3 variable is between 0.5 and 1, which means that the three factors, i.e. the project change, the project visa and the contract term, have strong impact on the dynamic cost and are the major reasons for the changes in the operating cost.
Figure 6. VIP diagram of the residence cost model

Figures 6 is the regression coefficient histogram of the corresponding independent variables of the regression equations based on the standard data. The figures reflect the marginal contribution of each independent variable to the actual cost. As the figures show, the project changes, the project visa and the contract term as well as the project volume play a positive role in the actual cost of the 2# project. In other words, increases of these factors will cause increase in the project target cost. In the residence VIP graph, the VIP value of the changing indicator is 0.486507, which is correlated to the large expenditure on the underground steel concrete in the preliminary building stage.

Analysis of the models shows that building a prediction model using the PLS method to predict the dynamic cost performs well in solving the multiple linearity problem among independent variables, and the built model has strong prediction capacity.

3.3. Verification of the model results

The prediction model equations 1 is used to predict the dynamic cost of the projects. By substituting the actual data into the equations, we obtain the difference between the predicted value and the actual value. The relative errors of the predicted values are shown in Tables 4.

Table 4 Residence cost data error comparison

| Construction site | actual cost (yuan) | Actual accounting (yuan) | Data fitting prediction (yuan) | actual error | data error |
|-------------------|--------------------|--------------------------|------------------------------|--------------|-----------|
| 15                | 1551283.2          | 725289.28                | 894550.8034                  | 1.138847551  | 0.233371054 |
| 16                | 54940769.21        | 13901425.14              | 35582475.09                  | 2.952168116  | 1.559627861 |
| 17                | 3286119.88         | 3605201.51               | 1915030.857                  | -0.088505907 | -0.468814475 |
| 18                | 12341509.03        | 7508132.5                | 7192356.197                  | 0.643752162  | -0.042057902 |
| 19                | 16832333.22        | 17050724.57              | 9809321.276                  | -0.012808333 | -0.424697687 |

Table 4 shows that the relative error of predicted values of all engineering sites is small. Some items of the cost prediction may exceed the expected range, but the overall cost prediction error remains small, thereby proving the feasibility of building a model.

4. Conclusions

Based on the data including the dynamic cost, project changes, project visa and contract term of residence projects from the third season of 2018 to the fourth season of 2019, a PLS regression prediction
model was built, and verification and assessment of the model were performed. The following conclusions are reached.

(1) The dynamic cost of the project is subject to the impacts of the contract indicator, the project change, the contract term and other factors. Correlation verification shows that these factors are highly correlated. The PLS method can solve the problem of multiple linearity.

(2) In the dynamic cost prediction model built based on the PLS method, the explanatory capacity of the explanation variables reaches 0.979. And the relative error of predicted values of all engineering sites is small. This is a high accuracy, which means that the PLS regression method solves the multiple correlation problem among independent variables. Therefore, using the SIMCA software and the PLS regression method is a good solution to predict the project’s dynamic cost.

References
[1] Wold S, Albano C, Dunn WJ. Pattern regression finding and using regularities in multivariate data[J]. Analysis Applied Science, 1983: 147-188.
[2] Wang huwen. Comparative analysis of coastal and inland city economy in China by PLS regression [J]. Mathematical statistics and management, 1998 (5): 10-15.
[3] Wang huwen. Partial least squares regression method and its application [M]. Beijing: national defense industry press, 1999: 200-234.
[4] Wang h w. Linear and nonlinear methods of partial least squares regression [M]. Beijing: national defense industry press, 2006: 97-141.
[5] Hou longwen. Optimization management of real estate architectural design cost [M]. China building materials industry press, 2016.5