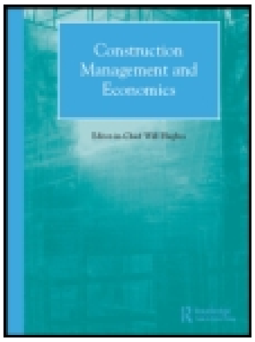

Routledge
Taylor & Francis Group

Construction Management and Economics

ISSN: 0144-6193 (Print) 1466-433X (Online) Journal homepage: https://www.tandfonline.com/loi/rcme20

Project cost estimation using principal component regression

Swee Lean Chan Corresponding author & Moonseo Park

To cite this article: Swee Lean Chan Corresponding author & Moonseo Park (2005) Project cost estimation using principal component regression, Construction Management and Economics, 23:3, 295-304, DOI: 10.1080/01446190500039812

To link to this article: https://doi.org/10.1080/01446190500039812

Construction Management and Economics (March 2005) **23**, 295–304

Project cost estimation using principal component regression

SWEE LEAN CHAN[1]* and MOONSEO PARK[2]

[1]*Department of Building, National University of Singapore, 4 Architecture Drive, Singapore 117566*
[2]*Department of Architecture, Seoul National University, San 56-1 Shinrim-dong, Seoul, Korea*

Received 21 August 2003; accepted 15 November 2004

Factors affecting construction project cost include project-specific factors and those reflecting the characteristics of the project team. Multiple regression is often used to estimate a project's cost, but independent variables with a high degree of correlation are likely be left out of such a model. As a result, only a limited number of factors are included in the estimate of project cost and predictions from such models will not be accurate. To overcome this technical inefficiency, the aims of this study are: to identify factors that contribute to project cost, to construct a predictive project cost model using the principal component technique and to assess the relative importance of determining factors. The data are obtained from a random sample survey comprised of Singapore building projects completed after 1992 costing more than US$5 million in value. Three main groups of variables are identified, pertaining to characteristics of the project, contractors and owner/consultants. Special project requirements such as high technological level; contractor's specialized skills; and public administered contract have significant effects on cost. Other factors include contractor's technical expertise; owner's level of construction sophistication and contractor's financial management ability. The model assesses the impact of individual factors on project cost and provides a decision support tool to estimate cost more accurately.

Keywords: factors, principal component regression model, project cost

Introduction

The cost of a construction project is an important concern in any construction project. In order to control the cost within an acceptable level, it requires appropriate and accurate measurement of various project-related determinants and the understanding of the magnitude of their effects. Past studies (Songer and Molenaar, 1997; Konchar and Sanvido, 1998) have identified a list of metrics that measure and compare the performance of construction projects. Other empirical studies (Akintoye, 2000; Chan *et al.*, 2001) identified the determining factors and assessed their impacts on project cost.

A common finding of such studies is that cost is affected by a large number of factors. This can be explained by the fact that construction is a multidisciplinary industry and its work involves many parties such as the owner, professionals, contractors and suppliers. Therefore, integrated efforts of the various parties and their decisions regarding the design, technology and implementation of the project can have significant effect on the overall project cost.

In establishing the relationship between determining factors and cost, many studies (Trost and Oberlender, 2003; Ling *et al.*, 2004) apply the ordinary least square regression approach and choose the best model based on the coefficient of multiple determination or the R^2 value. However, this approach tends to produce regression coefficient estimators that will perform poorly in the presence of multicollinearity, which is likely to surface due to high correlation among a large group of variables. In addition, the variance of the ordinary least squares estimator becomes inflated, which results in the low possibility of the estimator being close to the true value of the regression coefficient.

Technically, the analysis can be enhanced and the accuracy of the predictive model improved, if the

*Author for correspondence. E-mail: bdgcsl@nus.edu.sg

Construction Management and Economics
ISSN 0144-6193 print/ISSN 1466-433X online
http://www.tandf.co.uk/journals
DOI: 10.1080/01446190500039812

procedure allows for prior selection and determination of uncorrelated variables to be included in the regression model. This study seeks to address this technical inefficiency and proposes an alternative solution. In addition, it aims to (i) identify factors that contribute to project cost; (ii) construct a predictive model using the proposed principal component technique that produces smaller standard error of estimate compared to the classical ordinary least square method; and (iii) examine the relative importance of these factors based on the significance of their contributions. The unit cost, measured by the ratio of final project cost to the area of the project, is used as the predictor of the model which is regressed against chosen components represented by the contributing variables.

The following sections review the past work, outline the principal component technique, describe the data, present and discuss results of the empirical study. The last section concludes.

Previous work

Akintoye (2003) examined factors that influence a contractor's estimate of a project cost. Of the 24 factors, the researcher identified seven main factors using the factor analysis approach: project complexity, technological requirements, information, team requirement, contract requirement, duration and market requirement.

Trost and Oberlender (2003) established a predictive model to determine project cost. Factor Analysis was applied to group 45 variables into 11 orthogonal factors. Multiple regression analysis was performed on the 11 factors using their factor scores to predict the cost and assess its accuracy. Of the 11 factors, 5 were found to be significant at 10% level: process design, team experience and project information, time allowed to prepare the estimate, site requirements and bidding and labour climate.

Hegazy and Ayed (1998) developed a parametric cost-estimating model for projects where little information is known about the scope of the project. This type of parametric model consists of one or more functions, or cost-estimating relationships, between the cost as the dependent variable and the cost-determining factors. The neural networks approach adopted by this study is appropriate due to the presence of difficult tasks involving intuitive judgment and the need to detect data patterns that elude conventional analytical techniques.

There are studies that examined the performance of a construction project from various aspects. By performance, the authors mean how well can the project (i) complete within budget and schedule; (ii) achieve acceptance level of quality; and (iii) fulfill the owner's and consultant's satisfaction.

Other researchers measured the performance of a construction project by different metrics. Konchar and Sanvido (1998) and Songer and Molenaar (1997) identified 11 performance metrics that are classified under the broad categories of cost, time, quality and other aspects of the construction project. They measured project cost in three ways: (i) unit cost expressed as a ratio of final project cost to area; (ii) cost growth determined by the percentage change in final project cost and contract project cost; and (iii) intensity, defined as the unit cost of design and construction work put in place in a facility per unit time.

Bennett *et al.* (1996) studied project selection and performance from the owners' perspective. They constructed three statistically viable models to predict these three performance metrics: unit cost, construction speed and delivery speed. In determining one of the performance metrics, the study included other performance metrics as predictor variables, which clearly showed the interdependence of the three factors.

Bromilow's (1969) log–log time cost model is widely used by researchers to estimate the construction completion time based on an estimate of the size of the project measured by the cost.

Chan *et al.* (2001) identified a set of project success factors for design and build projects and examine the relative importance of these factors. In their study, the success of the projects is measured by the ability to complete on time and within budget. Terms such as 'behind schedule', 'on schedule'; 'ahead of schedule'; 'overrun budget'; on budget' and 'underrun budget' are used to categorize the projects. Some project success factors are highlighted: team commitment, contractor's competences, risk and liability assessment, client's competencies, end users' needs and constraints imposed by end users.

Pinto and Slevin (1988) stressed the important influences of project team experience, contracts, resources and information availability on the success of a project.

Ashley *et al.* (1987) conclude that effective allocation of resources has impact on project performance. Songer and Molenaar (1997) relate specific project characteristics such as project, owner, market and relationship variables to the project success.

Some researchers reckon that success factors are different for different types of projects. Tiong (1996) believes that the success of build–operate–transfer projects is affected by a limited number (six) of critical factors. As for design and build projects, the number of factors to be considered is much more. Chan *et al.* (2001) identified 31 factors, Ashley *et al.* (1987) 46, Songer and Molenaar (1997) 15 and Konchar and Sanvido (1998) had nearly 100 factors that affect project cost, schedule and quality performance of a project.

Recently, Cox *et al.* (2003) reported the management perceptions of key performance indicators (KPI) utilized in the construction industry. Correlation analyses are performed for both quantitative and qualitative indicators to determine which type of indicator is used most extensively. Basic statistical analyses and frequency distributions provide evidence in support of some of the hypotheses of the research. The results of the survey data analysis conclude that KPIs vary according to management's perspective, and there is a substantial difference between construction executive and project management's perceptions. Six indicators were reported as being most useful by every segment of the construction industry involved in their study. The correlation between quantitative indicators and qualitative indicators proved to be inconclusive.

Most of the previous studies, including those reviewed above (Songer and Molenaar, 1997; Konchar and Sanvido, 1998), undertook the multivariate regression approach for estimating the performance metrics, as the technique is straightforward, easy to use and widely available within statistical computer packages. Kaming *et al* (1997) applied the factor analysis, and extracted 'factors' or 'components' out of the original variables that influence construction time: equipment usage, resource estimates, buildability and human resource shortages. Chan *et al.* (2001) grouped a large number of variables into the reduced number of 'factors' whose loadings indicate the relative importance of each factor. However, these factor analysis approaches are not fully adequate for explicitly displaying the relationship between the factors and the performance metric. For example, when a stepwise multiple regression model was constructed using the success factors as dependent variable and the factor scores of the factors extracted by factor analysis as independent variables, the derived model establishes the relationship between project success level and the relative importance of the factors, not the individual variables contributing to the factors. As a result, the model cannot be reused to predict the success level of a new project given its project characteristics and details.

Belassi and Tukel (1996), however, proposed a new framework for determining the critical success factors for projects by ranking them in four areas related to project; project manager and the team members; organization; and external environment. The study examines the factor on an individual basis instead of analysing the collective effects of success factors.

Among some analyses, such as a study by Ashley *et al.* (1987) that determines the relationship between 46 factors and project performance based on only eight projects, the number of observations is very much less than the number of independent variables or the model is over specified. This makes the formulation of the model ineffective (Rencher, 1995, p. 415).

Attala and Hegazy (2003) developed a model using artificial neural network (ANN) to estimate cost deviation from the planned values in reconstruction projects. They compared the ANN model with the regression model and concluded that both models produced relatively close results. The main difference was that ANN was more sensitive and used 18 variables in the estimation, while regression model used only five variables. However, the ANN technique is more suitable for problems involving high level of uncertainty and in cases where some decision support system is required. ANN is not appropriate for this study.

Research method

Survey data

Based on the literature review, a total of 57 determinants (Table 1) relating to the project, the construction team and the contractor are identified to have an impact on project cost. Most of the determinants are qualitative variables expressed in the Likert scale of between a minimum of 1 and a maximum of 5. All, except the dummy variables, are classified under three broad categories, pertaining to (A) project design, complexity and time; (B) professional level of the project team; and (C) contractor's competency.

A questionnaire survey was designed to obtain the primary data for this research. A pilot study was first carried out to test the relevance and comprehensiveness of the questionnaires before a full-scale survey was conducted. The random sample was selected from the Singapore Building and Construction Authority's (BCA) website (BCA, 2002). It comprised building projects that exceed US$5 million in value, and were completed after 1992. The project data were collected from owners, architects, engineers and contractors.

The respondents were given a choice of being interviewed face-to-face or self-administer the questionnaires, and sending them back to the researchers. In total, 400 randomly selected projects were identified, and questionnaires were sent to 40, 57 and 60 owners, architects, engineers and contractors, respectively.

Altogether, data from 87 projects were received, giving rise to a 22% response rate. The profile of the projects is presented in Table 2. More than half the sampled projects were residential type and public owned.

This study adopts two stages in the analysis of the data, namely, extracting components and using

Table 1 Determinants of project cost

Codes	Explanatory variables	Definition
LUM	Type of contract	Dummy variable: 1=yes; 0=no
SIA	Singapore Institute of Architects Form of contract	Dummy variable: 1=yes; 0=no
PSC	Public sector standard conditions of contract;	Dummy variable: 1=yes; 0=no
RES	Residential building	Dummy variable: 1=yes; 0=no
COM	Commercial building	Dummy variable: 1=yes; 0=no
IND	Industrial building	Dummy variable: 1=yes; 0=no
PUB	Public owned	Dummy variable: 1=yes; 0=no
X6	Level of design complexity	1=not complex; 5=highly0complex
X7	Level of construction complexity	1=not complex; 5=highly complex
X8	Level of technologically advancement	1=not complex; 5=highly complex
X9	Level of specialization required of contractors	1=not complex; 5=highly complex
X10	Percentage of repetitive elements	6=0–10%; 5=11–20%; 4=21–30%; 3=31–40%; 2=41–50%; 1=>50%
X11	Presence of special issues	1=Yes; 2=No
X12	Type of specification	1=performance; 3=combination; 5=prescriptive
X13	Extent to which bid documents allow additions to scope	1=prevented addition; 5=encouraged additions
X14	Flexibility of scope of works when contractor is hired	1=flexible; 5=not flexible
X15	Project scope definition completion when bids are invited	1=low; 5=high
X16	Design completion (by owner) when bids are invited	1=0%; 2=up to 10%; 3=11–25%; 4=26–49%; 5=⩾50%
X17	Design decisions made (by owner) when bids are invited	1=up to 10%; 2=11–20%; 3=21–30%; 4=31–49%; 5=⩾50%
X18	Design completion when budget is fixed	1=0%; 2=1–5%; 3=6–10%; 4=11–25%; 5=26–50%; 6=>50%
X19	Good bidding information	Dummy variable: 1=yes; 0=no
X20	Importance for project to be completed within budget	1=not crucial; 5=very critical
X21	Importance for project to be delivered	1=not crucial; 5=very critical
X22	Preparation for bidding	1=inadequate; 5=adequate
X23	Time given to owners/consultants to evaluate bids	1=inadequate; 5=adequate
X24	Extent to which the contract period is allowed to vary during bid evaluation stage	1=firmly fixed; 5=variable
X25	Importance for the project to be completed on time	1=not crucial; 5=very critical
X30	Bidding environment	1=scarce; 5=high/plentiful
X31	Consultant's level of construction sophistication	1=low; 5=high
X32	Owner's level of construction sophistication	1=low; 5=high
X33	Consultant's experience with similar projects	1=no similar projects; 5=nearly all those types
X34	Owner's experience with similar projects	1=no similar projects; 5=nearly all those types
X35	Consultant's staffing level to attend to contractor	1=low; 5=high
X36	Owner's staffing level to attend to contractor	1=low; 5=high
X37	No. of DBB/DB projects handled by consultant in the past	1=0; 2=1; 3=2–3; 4=4–6; 5=7–10; 6=>10
X38	No. of DBB/DB projects handled by owner in the past	1=0; 2=1; 3=2–3; 4=4–6; 5=7–10; 6=>10
X39	Contractor's experience with similar types of projects	1=no similar projects; 5=nearly all those types
X40	Contractor's experience with similar size of projects	1=no similar projects; 5=nearly all those types
X41	Contractor's experience with projects in Singapore	1=no similar projects; 5=nearly all those types
X42	Subcontractors' experience and capability	1=poor; 5=excellent
X43	Communication among project team members	1=poor; 5=excellent
X44	Contractor's prior working relationship with the owner	1=poor; 5=excellent
X45	Contractor's prior working relationship with consultants	1=poor; 5=excellent
X46	Contractor's track record for completion on time	1=poor; 5=excellent
X47	Contractor's track record for completion on budget	1=poor; 5=excellent
X48	Contractor's track record for completion to acceptable quality	1=poor; 5=excellent
X49	Contractor's staffing level	1=low; 5=high
X50	Adequacy of contractor's plant and equipment	1=low; 5=high
X51	Magnitude of change orders in contractor's past projects	1=low; 5=high

Table 1 (cont'd)

Codes	Explanatory variables	Definition
X52	Magnitude of claims and disputes in contractor's past projects	1=low; 5=high
X53	Contractor's key personnel's management ability	1=poor; 5=excellent
X54	Contractor's ability in financial management	1=poor; 5=excellent
X55	Contractor's quality control and management capability	1=poor; 5=excellent
X56	Contractor's health and safety management capability	1=poor; 5=excellent
X57	Contractor's technical expertise	1=poor; 5=excellent
X58	Contractor's design capability	1=no in-house capability; 5=full in-house designers
X59	Size of contractor by paid-up capital (US$)	1=14–27K; 2=28–82K; 3=83–138K; 4=139–277K; 5=278–833K; 6=834K–4=139–277K; 5=278–833K; 1.39M 7=1.4M–2.8M; 8=>2.8M

principal component regression technique to model the data.

Extracting components

In regression modelling it is known that the sample size, N, must be greater than the number of explanatory variables, K, to avoid over specification problem. In cases where the response variable, y, is determined by a large number of independent variables such that the value of N is less than K, the classical regression analysis will not be appropriate. Furthermore, the classical regression analysis assumes that there is no linear relationship among independent variables.

In this study, N is 87 while K is 57. With a large number of explanatory variables, the occurrence of correlation among independent variables is highly probable. To avoid multicollinearity problem in the analysis, another approach that offers intuitive appeal is that of principal components (Maddala, 1992, p. 285). Consider a set of independent variables represented in an X matrix, the first stage is to extract from it a small number of variables that, in some sense, account for most of or all the variation in X (Greene, 1997, p. 425). This is termed extracting components.

The next stage would be to apply the principal component regression method to regress the response variable on a smaller set of components with minimal loss of information. Because the components are uncorrelated, it eliminates the possible occurrence of multicollinearity that will inflate the variance of the ordinary least-squares estimator. Hence, the resultant model will produce estimated regression coefficients with smaller mean square errors.

Table 2 Profile of projects surveyed

Type of building	Number	%
Residential	50	57
Factory/industrial building	19	22
Office	7	8
School	11	13
Total	87	100
Gross floor area (m^2)		
10 000–50 000	48	55
50 001–100 000	20	23
>100,000	19	22
Total	87	100
Ownership		
Public	48	55
Private	39	45
Total	87	100
Contract sum		
US$5–14.9 million	31	36
US$15–29.9 million	13	15
US$30–49.9 million	17	20
⩾ US$50 million	26	30
Total	87	100

Selecting principal components for regression

After the extraction, the next step is to select significant principal components to be used for the estimation. Only some principal components but not all of them are to be chosen, otherwise, it will be the same as regressing on all the Xs. The guideline is to choose only a small number of principal components that contributes satisfactorily to explaining the variation in y (Greene, 1997, p. 426) and if the components have a large number of variables, only those with high contributions should be selected.

This study adopts Rencher's (1995, p. 434) criterion of selection: include principal components whose eigenvalues are more than the average of the eigenvalues, $\frac{\sum_{i=1}^{p} \lambda_i}{p}$, where p is the number of principal components extracted from the data, and λ is the eigenvalue of component i.

Table 3 Principal components of categorized variables

Category A variables	Components								Category B variables	Components			Category C variables	Components				
	1	2	3	4	5	6	7	8		1	2	3		1	2	3	4	5
x6	−0.12	0.58	0.16	0.13	−0.11	0.35	0.25	−0.28	x31	0.45	−0.34	0.7	X39	0.27	0.76	0.35	−0.11	−0.14
x7	−0.24	0.54	−0.04	0.38	0.05	0.35	−0.17	−0.14	x32	0.82	−0.34	0.15	X40	0.26	0.71	0.45	0.07	0.08
x8	−0.19	0.76	0.26	−0.28	0.08	−0.21	−0.08	0.29	x33	0.74	−0.23	0.1	X41	0.24	0.70	0.46	−0.10	0.22
x9	−0.25	0.81	0.22	−0.19	0.07	−0.23	0.04	0.15	x34	0.85	−0.17	−0.18	X42	0.37	0.56	−0.20	0.29	−0.29
x10	0.44	0.01	−0.06	−0.14	0.39	0.51	−0.34	0.05	x35	0.53	0.73	0.19	X43	0.62	0.38	−0.15	0.46	0.04
x11	0.24	−0.28	−0.13	0.13	0.22	−0.52	−0.31	−0.44	x36	0.6	0.65	0.12	X44	0.28	0.61	−0.36	0.07	0.29
x12	0.10	0.45	−0.53	0.36	0.38	−0.09	0.13	−0.12	x37	0.5	0.43	−0.5	X45	0.42	0.44	−0.37	−0.14	0.19
x13	−0.24	−0.04	0.55	0.00	0.37	−0.18	0.38	0.02	x38	0.49	−0.48	−0.57	X46	0.72	−0.04	−0.05	−0.19	−0.33
x14	0.38	0.06	−0.18	−0.63	0.17	−0.22	0.34	−0.11					X47	0.82	−0.13	−0.13	−0.16	−0.35
x15	0.56	0.03	−0.25	−0.05	0.43	−0.09	0.24	0.37					X48	0.80	−0.21	0.08	0.21	−0.27
x16	0.52	−0.01	−0.26	0.14	−0.35	0.25	0.26	0.36					X49	0.92	−0.13	0.11	0.02	0.02
x17	0.38	0.29	0.10	−0.23	−0.59	0.01	−0.01	0.03					X50	0.83	−0.12	0.12	−0.04	0.06
x18	−0.05	0.36	−0.75	0.19	0.11	0.10	−0.23	0.13					X51	−0.33	−0.15	0.30	0.74	0.03
x19	0.23	0.21	0.01	−0.65	0.01	0.22	−0.26	−0.25					X52	−0.39	−0.17	0.47	0.60	−0.15
x20	0.46	−0.05	0.27	−0.24	0.41	0.41	0.09	−0.07					X53	0.81	−0.31	−0.09	0.26	−0.08
x21	0.41	0.38	0.44	0.23	0.04	−0.05	−0.12	−0.35					X54	0.80	−0.25	−0.21	0.02	0.22
x22	0.63	0.15	0.29	0.31	0.22	−0.19	−0.08	0.06					X55	0.71	−0.34	0.06	0.00	0.17
x23	0.55	0.00	0.41	0.57	−0.15	−0.01	0.20	0.08					X56	0.76	−0.14	0.32	0.15	0.35
x24	−0.51	−0.16	0.23	0.11	0.43	0.06	−0.23	0.35					X57	0.85	−0.21	0.12	0.02	0.15
x25	−0.41	−0.22	0.04	0.02	0.20	0.45	0.47	−0.21					X58	−0.04	−0.44	0.57	−0.36	0.33
x30	0.08	−0.24	0.50	0.00	−0.05	0.18	−0.44	0.22					X59	0.24	0.08	0.58	−0.38	−0.42
Eigenvalues	2.94	2.70	2.30	1.91	1.68	1.52	1.38	1.13	Eigenvalues	3.28	1.68	1.18	Eigenvalues	7.57	3.21	2.14	1.64	1.15
% Variance	13.99	12.88	10.94	9.12	8.01	7.25	6.55	5.37	% Variance	41.06	21.02	14.75	% Variance	36.97	15.56	9.85	8.21	5.37
Cumulative %	13.99	26.87	37.80	46.92	54.93	62.19	68.74	74.10	Cumulative %	41.06	62.08	76.82	Cumulative %	36.97	52.53	62.38	70.59	75.96

Estimating the effects

Suppose that of the K columns of X, $L<K$ principal components are used for the modeling, the regression equation is:

$$\hat{y} = \alpha + \beta\, X C_L, + \varepsilon;$$

where $\hat{y}$ is the fitted values of an N-dimensional response vector, α is a constant, X is an $N \times K$ matrix of the original variables, C_L is a $K \times L$ matrix containing the eigenvectors from the selected principal components, β is an L-dimensional vector of unknown regression parameters, while ε is a random vector that satisfies the basic normality assumptions of $E(\varepsilon)=0$ and $\mathrm{var}(\varepsilon)=\sigma^2$.

Hence, the principal components estimator of the full coefficient vector is $\beta\, C_L$. This estimator is a simple function of the original least squares coefficients in the regression of y on X.

Model estimation

Components extraction and selection

Using the factor analysis function in the Statistical Package for Social Sciences (SPSS), eight components are extracted from 21 variables within category A data pertaining to project design, complexity and time. The cumulative percentage variance explained by the eight components is 74%. The percentage variance explained by each of the component is reported in Table 4. Based on the significance of the contribution, and in comparison with the average eigenvalue (1.945), it seems that the first three components are the most important and have to be included in the model. The total percentage variance explained by the three components is about 38%. Of the eight variables from category B data that characterize the professional level of the project team, three components are identified but only the first component is relatively higher than the average eigenvalue (2.05). Within the last category C, comprising 21 variables, five components that amount to 76% of the variance are obtained. The first two components whose eigenvalues are higher than average (3.14) account for 53% of the variance and they contain the key variables in this category.

Independent variables

Altogether 10 variables, as listed in Table 4, which have the highest scores within the respective components identified earlier are chosen as independent variables. Furthermore, seven descriptive variables, namely LUM (lump sum contract), SIA (Singapore Institute of Architect's Form of Contract), PSC (Public sector standard conditions of contract), RES (residential building), COM (commercial building), IND (industrial building) and PUB (public-owned) are also included to test their effects on the cost model. They are represented as dummy variables. The sample statistics of all the independent variables are presented in Table 4.

Estimates of effects

Table 5 reports the estimated effects of the individual variables on the cost of a project using the principal component regression method. The final model

Table 4 Sample statistics of variables

	Mean	Median	SD	Skewness	Kurtosis	Jarque–Bera	Probability
LUM	0.90	1.00	0.30	−2.64	7.95	128.46	0.00
SIA	0.56	1.00	0.50	−0.24	1.06	9.84	0.01
PSC	0.39	0.00	0.49	0.45	1.20	9.94	0.01
COM	0.08	0.00	0.34	4.23	21.23	993.36	0.00
RES	0.61	1.00	0.49	−0.45	1.20	9.94	0.01
IND	0.19	0.00	0.39	1.61	3.59	26.36	0.00
PUB	0.49	0.00	0.50	0.03	1.00	9.83	0.01
X8	2.83	3.00	0.97	0.23	2.77	0.64	0.73
X9	3.07	3.00	0.89	−0.13	3.33	0.45	0.80
X16	4.88	5.00	0.38	−3.43	14.68	397.27	0.00
X32	3.60	4.00	0.91	−0.45	3.02	1.82	0.40
X47	4.02	4.00	0.75	−0.33	2.66	1.14	0.57
X49	3.98	4.00	0.78	−1.04	5.78	24.63	0.00
X50	3.94	4.00	0.69	−0.31	3.14	0.81	0.67
X53	4.06	4.00	0.72	−1.45	8.17	71.72	0.00
X54	3.94	4.00	0.83	−0.78	4.51	9.63	0.01
X57	4.10	4.00	0.74	−1.39	7.57	58.55	0.00

Table 5 Estimates of regression parameters using variables obtained from principal components

Variables	Coefficients	SE	*t*-Statistics	Sign.
(Constant)	786.35	1030.39	0.76	0.450
X8	−669.70	172.90	−3.87	0.000***
X9	883.03	182.77	4.83	0.000***
PSC	−977.65	288.98	−3.38	0.002***
RES	−559.33	273.48	−2.05	0.048**
X54	355.32	195.57	1.82	0.078*
X57	−513.67	203.56	−2.52	0.016**
X32	392.16	163.23	2.40	0.022**
R^2				0.62
Adjusted				0.54
SE of estimate				756.1
F-statistics (Prob)				8.95 (0.00)

Note: *, ** and *** denote rejection at the 10%, 5% and 1% significance level.

comprises seven significant variables: X8 (level of technologically advancement); X9 (level of specialization required of contractors); PSC (whether public sector standard conditions of contract is used or not); RES (whether it is a residential building or not); X54 (contractor's ability in financial management); X57 (contractor's technical expertise); and X32 (owner's level of construction sophistication).

Collectively, the variables account for 62% and 54% of the total variance of project cost indicated by the R^2 and adjusted R^2, respectively. The F-ratio test indicates that the joint effects of these variables are highly significant at the 1% significance level.

The regression model can be represented mathematically as:

$$\begin{aligned}\text{Cost} &= 786.3 - 669.7 * \text{X8} + 883.0 * \text{X9}\\ &- 977.6 * \text{PSC} - 559.3 * \text{RES} + 355.3 * \text{X54}\\ &- 513.7 * \text{X57} + 392.1 * \text{X32}\end{aligned} \tag{1}$$

It implies that, on the project side, higher technological level imposed on construction can reduce cost; but a stringent requirement on the contractor's specialized skill can result in an overpriced contract sum. Results of the study show that residential projects cost relatively less than other types of construction (commercial, industrial and 'others'). Similarly, the use of public sector standard of conditions of contract, denoted by the dummy variable PSC, implies government administration and probable cost-effective methods of construction, has a negative effect on cost. To ensure that the project can be executed properly and complete on time, a contractor's ability to manage his own finances is important although it may not necessarily contribute to lower project cost as his technical expertise does. Finally, the regression model shows that owner's level of construction sophistication, ranked third in Songer and Molenaar's (1997) survey of project success factors, can add to the cost of the project.

Validity of the model

To test the validity of the model, the predicted values of project cost computed using Equation 1 are compared to the actual values and the average percentage error of the estimation is found to be about 13%.

Some diagnostic tests are also performed to check if the errors are normal, homoskedastic and independent of the regressors, and that the linear specification of the model is correct (refer to Kenkel, 1989, for details of diagnostic tests). The results are presented in Table 6, the Jarque–Bera statistic tests for normality. If the residuals are normally distributed, the Jarque–Bera statistics should not be significant (QMS, 2000). The Breusch–Godfrey LM test checks for serial correlation. The null hypothesis is no serial correlation. The ARCH LM tests for autoregressive conditional heteroskedasticity (ARCH) in the residuals. The null hypothesis is no ARCH. As the result provides evidence of serial correlation in the errors, White's tests, with and without cross terms, are also performed. The above residual tests demonstrate that the basic assumptions underlying the multiple regression analysis have not been violated.

Table 6 Diagnostics of the estimated principal component regression

Test statistics	Size	*P*-value
Jarque–Bera	0.21	0.90
Breusch–Godfrey LM	2.11	0.15
ARCH LM	6.02	0.01
White's test (no cross terms)	32.20	0.00
White's test (cross terms)	39.43	0.17

Table 7 Estimates of the parameters using OLS without components

Parameters	Unstandardized coefficients	SE	t-Statistics
(Constant)	1749.943**	109.813	15.936
PUB	−782.170**	172.289	−4.540
Adjusted R^2			0.387
SE of estimate		478.66	
F-statistics (Prob)			

Note: *(**) denotes rejection at the 5% (1%) significance level.

Discussion

As a comparison, the classical least-squares regression method is used to regress the unit project cost on all the original 57 explanatory variables. Results presented in Table 7 reveal that only one variable qualifies to stay in the equation. This is because high correlation exists among the large group of independent variables and an estimate of individual effects becomes impossible. The model has an R^2 and adjusted R^2 value of 41% and 39%, respectively, which are at least 30% lower than what can be attained by the principal component regression method. In terms of accuracy, the application of principal components reduces the standard error of estimate by about 36%.

In addition, the principal component regression model comprising seven independent variables representing characteristics of the project and members of the construction team is more plausible than the one-variable classical model. For example, the two variables, RES and PSC, present in the final model reflect the prevailing condition of the construction market in Singapore. In the residential sub-market, the public sector is the key supplier of housing. More than 80% of the Singapore population lives in public housing estates. This is the result of the government's long-term strategy to achieve a high home ownership rate for Singaporeans. The construction of these public sector flats is administered by a government body, the Housing and Development Board (HDB), using the public sector standard conditions of contract. The flats are a form of government subsidized housing designed with basic and less expensive materials and methods of construction. Hence, project cost for public residential units are lower.

Conclusion

Asymptotic principal component regression method as a data reduction and selective procedure prior to modeling to eliminate multicollinearity and other associated problems is used. The technique is particularly applicable in this study because of the relatively large number of independent variables required for the estimation compared to the number of observations. Classical multivariate regression technique in such a situation will render the ordinary least-squares method an inappropriate and ineffective statistical approach. The principal component regression, on the other hand, reduces the large number of independent variables to a few uncorrelated principal components from which variables can be selected, and generates a model that is capable of producing better and more stable estimates of the regression coefficients.

Project cost depends not only on a single factor but a cluster of variables related to the characteristics of the project and the construction team. Technological and project design requirements preset by the client and the consultants; contractor's expertise and management ability; and the client's desired level of construction sophistication play an important role in determining the cost of the project. The application of the predictive model helps practitioners in the construction industry to determine if the budget for the project is attainable, given the general requirements of the project, the profile of the contractor and the owner's expectation.

References

Akintoye, A. (2000) Analysis of factors influencing project cost estimating practice. *Construction Management and Economics*, **18**(1), 77–89.

Ashley, D.B., Laurie, C.S. and Jaselskis, E.J. (1987) Determinants of construction project success. *Project Management Journal*, **18**(2), 69–79.

Attala, M. and Hegazy, T. (2003) Predicting cost deviation in reconstruction projects: artificial neural network versus regression. *Journal of Construction Engineering and Management*, **129**(4), 405–11.

Belassi, W. and Tukel, O.I. (1996) A new framework for determining critical success/failure factors in projects. *International Journal of Project Management*, **14**(3), 141–51.

Bennett, J., Pothecary, E. and Robinson, G. (1996) *Designing and building a world-class industry*, Centre for Strategic Studies in Construction, Reading.

Bromilow, F.J. (1969) Contract time performance expectations and the reality. *Building Forum*, **1**(3), 70–80.

Building and Construction Authority (BCA) (2002) Singapore Building and Construction Authority homepage, http://www.bca.gov.sg (Accessed 19 April 2002.).

Chan, P.C.A., Ho, C.K.D. and Tam, C.M. (2001) Design and build project success factors: multivariate analysis. *Journal of Construction Engineering and Management*, **127**(2), 93–100.

Cox, R.F., Issa, R.R.A. and Ahrens, D. (2003) Management's perception of key performance indicators for construction. *Journal of Construction Engineering and Management*, **129**(2), 141–51.

Greene, W.H. (1997) *Econometric Analysis*, Prentice Hall International, New Jersey.

Hegazy, T. and Ayed, A. (1998) Neural network model for parametric cost estimation of highway projects. *Journal of Construction Engineering and Management*, **124**(3), 210–18.

Kaming, P.F.K., Olomolaiye, P.O., Holt, G. and Harris, F.C. (1997) Factors influencing construction time and cost overruns on high-rise projects in Indonesia. *Construction Management and Economics*, **15**(1), 83–94.

Kenkel, J.L. (1989) *Introductory Statistics for Management and Economics*, PWS-Kent Pub. Co., Boston.

Konchar, M. and Sanvido, V. (1998) Comparison of U.S. project delivery systems. *Journal of Construction Engineering and Management*, **124**(6), 435–44.

Ling, F.Y.Y., Chan, S.L., Chong, E. and Lee, P.E. (2004) Predicting performance of design–build and design–bid–build projects. *Journal of Construction Engineering and Management*, **130**(1), 75–83.

Maddala, G.S. (1992) *Introduction to Econometrics*, Macmillan, New York.

Pinto, J.K. and Slevin, D.P. (1988) Critical success factors across the project life cycle. *Project Management Journal*, **19**(3), 67–75.

Quantitative Micro Software (2000) LLC, *Eviews 4.0 Users' Guide*, California.

Rencher, A.C. (1995) *Methods of Multivariate Analysis*, Wiley, New York.

Songer, A.D. and Molenaar, K.R. (1997) Project characteristics for successful public-sector design–build. *Journal of Construction Engineering and Management*, ASCE, **123**(1), 34–40.

Tiong, R.L.K. (1996) CSFs in competitive tendering and negotiation model for BOT projects. *Journal of Contraction Engineering and Management*, ASCE, **122**(3), 205–11.

Trost, S.M. and Oberlender, G.D. (2003) Predicting accuracy of early cost estimates using factor analysis and multivariate regression. *Journal of Construction Engineering and Management*, ASCE, **129**(2), 198–204.