Hybrid approach for reducing estimating overfitting and collinearity

Journal:	*Engineering, Construction and Architectural Management*
Manuscript ID	ECAM-08-2018-0353.R2
Manuscript Type:	Original Article
Keywords:	Estimating, Novel Model, Approach
Abstract:	

Hybrid approach to reducing estimating overfitting and collinearity

Abstract:

Purpose: The purpose of this paper is to present an approach to address the overfitting and collinearity problems that frequently occur in predictive cost estimating models for construction practice. A case study, modelling the cost of preliminaries is proposed to test the robustness of this approach.

Design/methodology/approach: A hybrid approach is developed based on the Akaike information criterion (AIC) and principal component regression (PCR). Cost information for a sample of 204 UK school building projects is collected involving elemental items, contingencies (risk), and the contractors' preliminaries. An application to estimate the cost of preliminaries for construction projects demonstrates the method and tests its effectiveness in comparison with such competing models as: alternative regression models, three artificial neural network data mining techniques, case-based reasoning, and support vector machines.

Findings: The experimental results show that the AIC-PCR approach provides a good predictive accuracy compared with the alternatives used, and is a promising alternative to avoid overfitting and collinearity.

Originality/value: This is the first time an approach integrating the Akaike information criterion (AIC) and principal component regression (PCR) has been developed to offer an improvement on existing methods for estimating construction project Preliminaries. The hybrid approach not only reduces the risk of overfitting and collinearity, but also results in better predictability compared with the commonly used stepwise regression models and traditional PCR approach under the sum of squares error (SSE) criterion.

Keywords: Akaike information criterion, principal component regression, overfitting, collinearity, construction cost estimation.

Introduction

Estimates for such variables as scope, cost, and schedule are needed for most construction projects and many papers have demonstrated the use of estimation methods such as multiple linear regression (MLR) for this purpose (Cheung and Skitmore, 2006b; Li *et al*., 2005; Skitmore and Patchell, 1990). Two main problems involved in the use of such methods are *overfitting* and *collinearity*, as the existence of either can produce significantly biased results. Overfitting occurs when too many independent variables are incorporated into the developed (training) model. An extreme example is where there are as many variables as cases so that, although a perfect fit is obtained with the sample data, the model has little prospect of representing the population and making accurate predictions. Collinearity occurs when the independence assumption is violated, i.e., when the independent variables are strongly intercorrelated and can be largely represented by fewer variables. In response, some studies have sought such other methods as artificial neural networks (ANN), case-based reasoning (CBR), and support vector machines (SVM) for solutions.

Unlike these black box or indirect approaches, however, MLR can produce the desired parameter estimates directly and accurately if collinearity and overfitting are dealt with properly. The collinearity problem in ordinary least squares (OLS) regression was recognised several decades ago (Skitmore and Marston, 1999), for instance, and many treatments have been developed. Although it is possible to improve a model by simply deleting one or more predictors with a high R_i^2 (see Eq. (8)) (O'brien, 2007), keeping or removing a variable should depend on the theoretical underpinning involved (Andersen and Bro, 2010). Ridge regression (RR), partial least squares regression (PLS), and principal component regression (PCR) (see Liu *et al*. (2003)) are three popular methods developed to deal with collinearity and avoid loss of information when deleting variables (Næs and Martens, 1988; Vigneau *et al*., 1997). Despite these methods being comparable in predictive ability, RR and PLS still produce biased estimates of the regression coefficients of the predictor variables (O'brien, 2007). PCR, on the other hand, is more consistent with stepwise MLR and collinearity diagnostics. In addition, although PCR does not necessarily lead to improved predictions relative to OLS, such improvements do nevertheless occur quite often in practice (Næs and Martens, 1988), and the currently recommended method of overcoming collinearity problems is therefore to use it to correct parameter estimates (Liu *et al*., 2003). This involves the application of the default stepwise regression approach in selecting predictor variables by simultaneously minimizing the sum of squares error (SSE) and maximizing adjusted R^2 in principal component selection.

To deal with overfitting problems, it is necessary to invoke the principle of parsimony in variable selection (Andersen and Bro, 2010). Including too many variables in the model leads to a high variance in parameter estimates and an overfitted model with poor predictability, while too few variables lead to the lack of necessary information and a decreased model fit (Johnson and Omland, 2004). For overfitting problems, the Akaike information criterion (AIC) is an asymptotically unbiased estimator of the expected relative Kullback-Leibler information quantity (Kullback and Leibler, 1951), and has been recommended for choosing suitable predictor variables (Akaike, 1974). This statistic represents the amount of information lost in the model fit when adding predictor variables, to help avoid overfitting with a comparatively small sample size (Posada and Buckley, 2004) and is given by

$$AIC = -2\iota + 2K \qquad (1)$$

with maximized log-likelihood (ι) and K estimable parameters. Despite the "superficial" form of the AIC formula, it is well founded in information theory and with a non-arbitrary "penalty term" $2K$ (Burnham and Anderson, 2002).

This paper aims to provide a solution to the situation where estimation collinearity and overfitting exist simultaneously. As pointed out by Xu (1994), using traditional PCR to counter multicollinearity problems increases the risk of overfitting, while using the SSE and adjusted R^2 criteria for variable selection in OLS regression models can result in some irrelevant variables being input. The AIC is the most commonly used information theoretic approach to measuring how much information is lost between a selected model and the true model. It has been used widely as an effective model selection method in many scientific fields, including ecology (Johnson and Omland, 2004) and phylogenetics (Sullivan and Joyce, 2005). Compared with the use of adjusted R^2 to evaluate the model solely on fit, AIC also takes model complexity into account (Johnson and Omland, 2004). In addition, AIC has several important advantages over the likelihood ratio test (Posada and Buckley, 2004).

Research framework

Acknowledging the effectiveness of AIC in model development, this paper presents a hybrid Akaike Information Criterion-Principal Component Regression (AIC-PCR) approach to deal with the problems of overfitting and collinearity that frequently occur in OLS regression. The following sections illustrate the literature review, descriptions of the approach, its experimental results, and a general estimation process. The relevant literature, especially that focusing on estimation techniques, is firstly reviewed. Then, the procedure involved in the hybrid approach is described in detail. Thirdly, an experimental evaluation of the hybrid approach is conducted with an application in modelling the preliminaries of construction projects. The results of the hybrid approach are further compared with those of other estimation methods, including an artificial neural network (ANN), case-based reasoning (CBR), and support vector machines (SVM). Finally, a standard estimation process is proposed for future applications.

Literature review

Construction cost estimation

The increasing scrutiny of construction costs by both clients and contractors has led to a strong desire for better utilisation of data and analytics (Ahmed *et al*., 2018). The term "building cost modelling", formally introduced in a Building Cost Research Conference held in 1982 (Newton, 1991), highlights the importance of the accuracy of early stage construction cost estimates in project decision making, and their large impact on downstream life-cycle stages (Ahmed *et al*., 2018; Lowe *et al*., 2006; Skitmore *et al*., 1990). Understanding the properties of a cost model is therefore vital for the effective control and development of future techniques (Skitmore and Marston, 1999). Although the accuracy of cost estimating is expected to improve as more information is released as the design evolves (Skitmore, 1987), clients still require accurate cost advice before design work to assist in assessing the feasibility of different development proposals. For such organizations as government authorities and real estate developers, inaccurate early estimates can result in the inefficient use of money, missed development opportunities, and unsuccessful project management (Oberlender and Trost, 2001). Furthermore, it has become increasingly common for the final cost of projects to exceed the estimated costs and by an increasing margin (Williams *et al*., 2005). For example, Flyvbjerg *et al*.'s (2003) analysis of 258 transportation infrastructure projects worth US$90 billion found 90 percent of cost overrun projects to be a direct result of inaccurate estimation in the early project stages. Similarly, Merrow *et al*. (1979) found that 74% of the cost growth of projects undertaken by the chemical, oil, and minerals industries in North America is also caused by underestimation in the early project stages.

Traditional early stage estimation for building construction projects is by floor area models, the main difficulty of which is that a building's floor area is not always representative of the overall cost. Similar to the floor area approach, some researchers and practitioners propose using the storey enclosure method and cost-duration model as an alternative method for simplicity (Cheung and Skitmore, 2006a; Dang and Long, 2018; Xiao *et al*., 2018). However, factors such as drawings and specification, pricing experience, project complexity, clear scope definition, site constraints (access, storage, and services), material availability, financial capabilities of the client, and availability of relevant cost information are also critical for achieving an accurate cost estimate (Muhammad *et al*., 2018). Estimates made in this way are therefore very much dependent on the knowledge and experience of the estimator in making the necessary subjective adjustments for these other factors; the floor area method, for example, being said to have a coefficient of variation around the true cost of 20-30% (Skitmore and Patchell, 1990). Moreover, including risk-related factors in cost estimation modelling is also important for reducing cost overruns (Ökmen and Öztaş, 2015).

Applications of MLR, CBR, ANN and SVM

Gaining cost certainty and reduction is a key driver to applying data mining in the AEC sector (Ahmed *et al*., 2018). The regression method has been used as an effective tool in the estimation of project performance for decades. For example, Williams (2003) employs regression models developed using data from five transportation agencies in the U.S. to predict the final cost of highway projects. Li *et al*. (2005) construct regression estimate models for office buildings in Hong Kong. To optimize predictive ability within the sample, the stepwise regression approach can be applied to meet the principle of parsimony. For example, Masrom *et al*. (2013) apply forward and backward stepwise regression to identify key items from 95 possible factors of contractor satisfaction and Guerrero *et al*. (2014) use stepwise regression modelling to predict the construction time of 168 Spanish building projects. Despite its applicability in many situations, the unthinking use of stepwise regression can make the method relatively weak. Son *et al*. (2012), for instance, use the full variable rather than stepwise regression in comparison with the SVM method in a dataset with severe collinearity.

Applications of Artificial Neural Networks (ANN) and Case-Based Reasoning (CBR) methods accounted for less than 5% of 56 publications related to construction cost estimation during 1960-1988 (Newton, 1991), but have developed rapidly since then with the aid of improved computer techniques (Chou and Tseng, 2011). ANN simulates the learning process of the human brain by representing variables as inputoutput nodes in a weighted network trained on data, and has been used to make predictions in a variety of fields (Kim *et al*., 2004; Kim *et al*., 2005). ANN models produce reasonable predictions with nonlinearity in the data (Hassim *et al*., 2018). In applications to construction projects, Kim *et al* (2004), for example, develop an ANN for cost estimation using data from 530 projects in Korea, showing its accuracy to be slightly higher than that provided by regression; Cheung *et al*. (2006) use ANN to predict project performance based on information available at the bidding stage from the Hong Kong Housing Authority; while Ajibade *et al*. (2015) propose the use of ANN as a viable alternative to regression for predicting the costs of electrical services components during the building design stage. However, ANN is a “black box” method and suffers the potential drawback of having to retrain the model completely with all data whenever a new case is added. Additionally, ANN studies have difficulties of generalization because of their overfitting nature (Min and Lee, 2005).

CBR is a method that uses previous experience to solve new cases (Aamodt and Plaza, 1994; Xu, 1994). It is particularly suited to: (1) obtaining a solution with partial understanding; (2) providing a reasonably close match to actual human reasoning; and (3) providing more explanation of its working (Xu, 1994). The inherent logic of CBR is consistent with Skitmore’s (1985) finding that construction experts predict by recalling the estimating details of previous projects and then adjusting these to suit new requirements. The number of publications applying CBR to construction related problems during the last decade is also increasing (Kim and Kim, 2010). For example, Kim *et al* (2005) use both ANN and CBR to model the construction cost of 540 Korean apartment buildings, finding that CBR performs particularly well; while Kim *et al* (2004) examine the estimating capabilities of MLR, ANN, and CBR using data from 530 projects to find that CBR outperforms both MLR and an average of 75 alternative ANN models.

Support vector machines (SVM) are developed mainly by Vapnik (2000) based on structural risk minimization, and have been shown to ensure good generalization (Movahedian Attar *et al*., 2013); An *et al*. (2007b) apply SVM to classify the accuracy of cost estimations for 62 Korean building projects and for regression purposes; and Movahedian Attar *et al*. (2013) use support vector regression (SVR) to forecast how far contractors deviate from client expectations during contractor prequalification, and find that SVR performs better than ANN. Son *et al*. (2012) use PCA-SVR, a SVM approach aided by principal component analysis (PCA) to reduce dimensions, to predict the construction costs of 84 building projects. However, that study ignores the severe collinearity of the 64 predicting variables used in their dataset, and any overfitting resulting from using so many variables/principal components given such a relatively small sample size. In fact, there has been little study generally in construction research of how effectively such methods handle data where the risk of overfitting and collinearity is significant.

Description and procedure of the hybrid approach

The treatment of modelling problems is usually considered in terms of detection and correction (Farrar and Glauber, 1967). Various diagnoses need to be considered before an appropriate approach can be proposed to solve collinearity and overfitting problems.

Collinearity diagnosis

When several variables/predictors in a multivariate regression model are highly correlated, one variable can be linearly and largely explained by the other variables. The coefficient estimates of a multiple regression with this problem may change erratically in response to small changes in the model or the data, rendering the coefficient estimates unreliable. The variance inflation factor (VIF) is commonly used as a standard way to detect collinearity, with larger VIF values indicating more severe correlation. In an ideal situation, when the predictors are not correlated, all $R_i^2 = 0$ and the VIF values of all variables have a minimum value of 1. A larger R_i^2 (dependency on other predictors) leads to a larger VIF. A VIF larger than 10 is usually used to indicate significant collinearity (Neter *et al.*, 1989). However, high VIF values do not necessarily worsen the regression analysis, and the influence of other factors on the variance of regression coefficients should also be considered (O'brien, 2007).

Two questions are particularly important in collinearity diagnostics: (1) how many dimensions in the predictor space are nearly collinear; and (2) which predictors are most strongly implicated in each of those dimension (Friendly and Kwan, 2009). To address these questions, Belsley *et al*. (2005) propose a strategy involving PCA, known as Belsley collinearity diagnostics. The strategy seeks to identify collinearity by introducing two statistics: a condition index (*CI*) and coefficient variance proportion (*CVP*). *CI* is defined as $CI_{\mathrm{k}} = \sqrt{\lambda_1/\lambda_{\mathrm{k}}}$ where $\lambda_k$is the Eigenvalue in collinearity diagnostics. Belsley *et al*. (2005) recommend caution with CI>10, while Friendly and Kwan (2009) regard CI_{k}<5 as "ok", 5<CI_{k}<10 as "warning", and CI_{k}> 10 as "danger". *CVP* indicates the proportion of variance of each variable associated with each principal component as a decomposition of the coefficient variance for each dimension (Belsley *et al*., 2005; Friendly and Kwan, 2009). Caution is needed with two or more CVP_{k}>0.5.

Overfitting diagnosis

Leave one out cross validation (LOOCV) is a commonly used method in model selection to detect overfitting and compare predictive ability (Xu, 1994). Compared with the insufficient data utilization and unreliability of traditional separate holdout-set in-out sample performance, LOOCV does not waste data and has better reliability in model predictive ability comparisons (Moore, 2001). LOOCV is easy to understand in that the i^{th} model is developed by training the remaining dataset without the i^{th} case, then using this model to predict the i^{th} case, and calculating the mean error after repeating this exercise N (i.e. sample size) times with replacement. Although widely applied in model development and selection in many other scientific areas (such as chemistry) this technique is comparatively new to the construction management and economics field. For example, Cheung and Skitmore (2006b) used the technique to compare the efficiency of the storey enclosure area method and four other traditional methods in very early design stage cost forecasting.

Despite having wide application and well-known predictive properties, however, LOOCV is often criticized for being time-consuming and performs comparatively poorly in selecting linear models when compared with more classical statistical methods (Rivals and Personnaz, 1999). For this reason, LOOCV is only used here to compare regression models developed by other statistical linear model development criteria (i.e. SSE, adjusted R^2, and AIC).

AIC-PCR procedure and formulas

If a MLR has overfitting problems and the model with the lowest AIC still suffers from collinearity problems, then the AIC-PCR procedure could be a useful alternative. AIC-PCR is described in eight steps, as follows:

Step 1: Proceed with the AIC criterion stepwise regression, with a column containing the actual values of the dependent variable Y and a matrix X comprising all the independent variables to obtain a model with the lowest AIC. The logic is to add the one variable that is most critical in reducing the AIC of the model, and then repeat by adding another variable or removing an existing variable. Whichever action most helps reduce AIC is made until the lowest AIC with *k* predictors is achieved. Software such as MATLAB has an automatic command for this task. Note that, while SPSS can provide the AIC values of a stepwise regression model using Syntax programming, these models are still selected according to the SSE criterion.

Step 2: Proceed to obtain collinearity diagnostics including *VIF*, *CI,* and CVP. The variance inflation factor (VIF) of the i^{th} predictor variable, indicating its collinearity with other predictors is given by:

$$X_i = \alpha + \beta_1 X_1 + \ldots + \beta_n X_n + error\ (X_i\ is\ excluded\ in\ the\ right\ side) \qquad (2)$$

$$Total\ sum\ of\ squares\ (TSS) = \sum_1^n (X_i - \overline{X_i})^2 \qquad (3)$$

$$Explained\ sum\ of\ squares\ (ESS) = \sum_1^n (\hat{X}_i - \overline{X_i})^2 \qquad (4)$$

$$Residual\ sum\ of\ squares\ (RSS) = \sum_1^n (X_i - \hat{X}_i)^2 \qquad (5)$$

$$TSS_i = ESS_i + RSS_i \qquad (6)$$

$$R_i^2 = ESS_i / TSS_i \qquad (7)$$

$$VIF = \frac{1}{1 - R_i^2} = TSS_i / RSS_i \qquad (8)$$

where, $\overline{X_\text{i}}$ is the mean value of X_i, and $\hat{X}_\text{i}$ is the estimated value of X_i. *TSS* is the sum of the squared differences between each observation and the overall mean; *ESS* is the sum of the squared deviations between the estimated values and mean values of each variable; and *RSS* is the sum of the squared residuals. CI and CVP can be determined by applying the collinearity diagnostics command in software such as MATLAB and SPSS, or calculated using equations 2-8 as provided.

Step 3: Apply the PCA with software such as MATLAB and SPSS to transform the k correlated variables to a set of uncorrelated principal components, C_i. All components should be extracted at this stage, and they should account for 100% of the variance.

Step 4: Compute the standardized dependent variable, the p standardized independent variables, and the values of the p principal components respectively in preparation for establishing p standardized PCR equations:

$$Y^{'} = (Y - \overline{Y})/S_Y \tag{9}$$

$$X_i{'} = \frac{X_i - \overline{X_i}}{S_{X_i}} \; (i = 1,....,k) \tag{10}$$

$$C_j = a_{1j}X_1^{'} + a_{2j}X_2^{'} + ... + a_{kj}X_k{'} \; (i = 1,....,k; j = 1,....,p) \tag{11}$$

where, $Y^{'}$ denotes the standardized dependent variable; Y the dependent variable; S_Y the standard deviation of the dependent variable; $\overline{Y}$ the mean of the dependent variable; $X_\mathrm{i}^{'}$ the ith standardized independent variable; X_i the ith independent variable; $\overline{X_\mathrm{i}}$ the mean of the ith independent variable; $S_{\mathrm{X_i}}$the standard deviation of the ith independent variable, C_j the jth principal component and a_ij the coefficient of the principal component matrix (the matrix consists of C_j and X_i).

Step 5: Proceed with the AIC criterion stepwise regression of principal components to estimate $Y^{'}$ in MATLAB and, if not all principal components are significant, select the lowest AIC regression equation, as in:

$$\hat{y}^{'} = \sum B_j^{'} \ C_j(\mathrm{j} = 1, \dots, q \leq p) \tag{12}$$

Step 6: Transform Eq. (12) with Eq. (11) to obtain:

$$\hat{y}^{'} = \sum b_i^{'} \ X_i^{'} \ \ (i = 1, \dots k) \tag{13}$$

where $\hat{y}^{'}$ is the estimate and $b_{\mathrm{i}}^{'}$ the i^{th} standardized coefficient of the standardized linear regression equation (Liu *et al.*, 2003).

Step 7: Calculate the regression coefficients and constant, and transform the standardized linear regression equation into a general linear regression equation

$$b_i = b_i^{'} \ (\frac{L_{yy}}{L_{x_i x_i}})^{1/2} \tag{14}$$

$$b_0 = \bar{Y} - \sum b_i \bar{X}_i \ (i = 1, \dots, \mathrm{k}) \tag{15}$$

$$\hat{y} = b_0 + \sum b_i X_i \ (i = 1, \dots, \mathrm{k}) \tag{16}$$

where, b_{i} is the regression coefficient of the i^{th} variable; L_{yy} the sum of squares of dependent variable *Y*; $L_{\mathrm{x_i x_i}}$ the sum of squares of the i^{th} independent variable X_{i}; and b_0 is the constant of the new linear model.

Step 8: Calculate the mean squared error (MSE) of the final model from

$$MSE = \frac{1}{n} \Sigma_1^n (Y_i - \hat{Y}_i)^2 \tag{17}$$

Experimental evaluation of the hybrid approach

This section introduces the detailed evaluation of the approach through an application in estimating construction preliminaries. In addition to the proposed approach, such alternative approaches as ANN and SVM are also applied to gauge the relative usefulness of the approach.

Target cost-preliminaries

The elemental cost analysis technique has been widely used by consultant quantity surveyors for decades as a base for their early stage predictions. According to a survey conducted by Scotos and Lowe (2011), The Standard Form of Cost Analysis (SFCA) developed by the *Building Cost Information Service* (BCIS) of the *Royal Institution of Chartered Surveyors* (RICS) is the most popular model. In the SFCA, construction costs comprise eight components: substructure, superstructure, internal finishes, fittings, services, external works, contingencies, and preliminaries. The quantity of the items involved in most of these cost components are relatively straightforward to determine given the level of information typically available at an early project stage, and the value of contingencies is decided by the clients/consultants prior to tendering with specific consideration of risks.

The CIOB's *Code of Estimating Practice* defines preliminaries as *the cost of administering a project and providing general plant, site staff, facilities, and site based services and other items not included in prior rates*. Their actual allocation by the contractor is influenced by a complex combination of past and recent experience on the part of the contractor, the current workload of the contractor, market conditions, and project characteristics often determined by the contractor, such as contract duration (Akintoye, 2000; Tah *et al*., 1994). The cost of subcontractor rework is another expenditure component generally included as a component of the preliminaries (Love and Li, 2000). Preliminaries are regarded as a key competitive component not determined until the tendering stage. Since the mark up strategy for preliminaries is different to that of other cost components, it can be used to achieve an unbalanced tender that significantly improves the cash flow of a contractor (Kaka, 1996). On this basis, the preliminaries can be particularly problematic to estimate using standard cost estimating techniques, and for that reason, the preliminaries is used as the targeted cost estimate in this research.

Sample projects

The sample cases comprise 204 UK school building projects completed during 2000-2012, selected from a large commercial cost database. Project differences due to geographical location, construction year, and rate of inflation are addressed by rebasing all prices to the same date (fourth quarter, 2012) and location (Greater London district), using the BCIS Construction Price Index. Some important characteristics of the sample cases are presented in Tables 1 and 2. The left hand column provides the categories, or cost drivers. These are blank in places for some projects due to a lack of complete information concerning such features as building height and contract type.

Please Insert Table 1 here

Please Insert Table 2 here

Framework of elemental cost items and sample descriptions

The 2012 4th edition of the *Elemental Standard Form of Cost Analysis* is used as the reference to construct the analysis framework. The elemental cost items in the first column of Table 3 refer to the variables used for model development in the MLR process, and those in the second code column refer to the selected variables in the PCR process.

Please Insert Table 3 here

Experimental results

Four models were developed in the model development phase by stepwise regression modelling under the criteria SSE (model 1), adjusted *R* square (model 2), AIC (model 3) and by directly entering all the predictor variables (model 4), as presented in Table 4. The overfitting and collinearity diagnoses follow.

In this case, using LOOCV to detect overfitting involved developing 204x4=816 sub-models to evaluate their predictive ability. The $\mathrm{MSE}_{\mathrm{LOOCV}}$ values for each model are presented in parentheses in Table 4. Although the MSE values of all four models are comparable, the model overfitting varies greatly. For example, Model 2 has the lowest MSE and highest adjusted R^2 but its predictive ability is weaker, while Model 3, developed under the AIC, has the lowest $\mathrm{MSE}_{\mathrm{LOOCV}}$. As the collinearity diagnoses for the four models indicate, collinearity is a common problem, with more than one *VIF* larger than 10, *CI* larger than 10, and *CVP* larger than 0.5. That is, even the overfitting-reduced Model 3 suffers from collinearity. Although both MATLAB and SPSS can help handle such diagnoses, MATLAB programming is selected here for its ability to visualize the collinearity diagnostics (Friendly and Kwan, 2009). Figure 1 provides a visual illustration of the diagnostics for Model 3. For clarity, only the principal components with *CI* values larger than five are shown.

Please Insert Table 4 here

Please Insert Figure 1 here

Figure 1: Collinearity diagnostics for Model 3.

According to diagnoses of overfitting and collinearity, this dataset is suitable for testing the AIC-PCR approach as it displays both features. Following the AIC-PCR approach, Model 3, with the lowest overfitting and best predictability, is obtained by applying the PCR under AIC. Table 4 gives the MSE results of the four models for comparison with Model 1 representing a stepwise regression under SSE; PCR, representing the traditional PCR approach developed under SSE; Model 3 representing the stepwise regression under AIC; and AIC-PCR representing the approach proposed in this study. The results show that the AIC-PCR approach not only avoids overfitting (by applying the AIC criterion) and collinearity (by applying PCR) but also improves predictability (with 7.73% less MSE than the default stepwise regression model) and accuracy (with 1.74% less MSE than the traditional PCR approach using the SSE criteria).

Comparisons with other methods

The AIC-PCR approach is compared with other data mining techniques of ANN, PCA-SVR, and K-Nearest Neighbour (KNN) as a basic type of CBR, to see how well it performs. Two absolute evaluation criteria of root mean squared error (RMSE) and mean absolute error (MAE), and a scaled indicator of mean absolute percentage error (MAPE), are used in common with similar studies (e.g. Dang and Le-Hoai; 2018; Kim, 2005), where

$$RMSE = \sqrt{\tfrac{1}{n}\Sigma_1^n (Y_i - \hat{Y}_i)^2} \qquad (18)$$

$$MAE = \tfrac{1}{n}\Sigma_1^n |\, Y_i - \hat{Y}_i| \qquad (19)$$

$$MAPE = \tfrac{1}{n}\Sigma_1^n \left|\frac{Y_i - \hat{Y}}{Y_i}\right| \qquad (20)$$

Please Insert Table 5 here

The results presented in Table 5 confirm the effectiveness of AIC-PCR, its error rate being the same or lower than the other four methods whichever criteria are used.

Proposed estimation process for practical applications

A proposed estimation process, incorporating the research design and AIC-PCR model development steps, for practical applications is illustrated in Figure 2. The main feature of this process is that, once the model is constructed, the overfitting and collinearity tests are applied and adjustments made as necessary by the AIC-PCR hybrid approach. Ideally, this would involve several models for comparison by the LOOCV method to obtain their likely performance and hence the selection of the best method to use.

Please Insert Figure 2 here

Figure 2: Estimation modelling framework and AIC-PCR procedure.

Implications

The study enables the application of the AIC-PCR approach to deal with overfitting and collinearity problems. For the first time, the developed approach is applied to model the cost of preliminaries using a significant sample of building construction projects. An acceptable level of predictability is obtained in comparison with other alternatives. The estimation modelling framework and AIC-PCR procedure together offer a practical and effective paradigm for researchers and practitioners to predict even the most complex elements of a construction project cost estimate.

Limitations

Some limitations should be acknowledged. Firstly, the sample tested in the study is from construction projects. Whilst the experimental application presented should be taken cautiously in wider generalization, it does demonstrate the capability of the hybrid approach in avoiding overfitting and collinearity problems and gaining accurate estimates. Additionally, there is not yet a standard cut-off study for determining the appropriate linear level needed for applying the proposed approach or such other approaches as SVM. From a practical perspective, the overall approach would benefit from simplification given practitioners may be unfamiliar with the likes of Matlab and SPSS. However, the proposed approach is no more difficult to apply than other cost models commonly built using ANN, LBR or SVM, for example. In spite of these limitations, the research proposes an estimation framework incorporating the AIC-PCR approach that offers interesting and prospective ground for further research and practice. Future studies could benefit from testing the applicability of the approach in other contexts such as to other estimate components, other building types, other/smaller/larger data sources, and other aspects of construction management and economics research entirely.

Conclusions

The main aim of this study is to build an alternative approach to deal with the preponderance of overfitting and collinearity problems that typically occur in construction research. A hybrid AIC-PCR method is developed and tested using cost data for the Preliminaries of 204 construction projects. The tests show that the hybrid approach not only reduces the risk of overfitting and collinearity, but also results in better predictability compared with the commonly used stepwise regression models and traditional PCR approach under the SSE criterion. The study also validates its applicability by comparison with other conventional methods including ANN, CBR, and SVM. The proposed approach provides a promising alternative for equivalent situations where overfitting and collinearity can be problematic, especially when the linear form offers a reasonable approximation to describe the relationships between the independent and dependent variables.

Acknowledgements

TBA

References

Aamodt, A. and Plaza, E. 1994, "Case-based reasoning: Foundational issues, methodological variations, and system approaches", *AI Communications*, Vol. 7, pp. 39-59.

Ahmed, V., Aziz, Z., Tezel, A. and Riaz, Z. 2018, "Challenges and drivers for data mining in the AEC sector", *Engineering, Construction and Architectural Management*, Vol. 25, No. 11, pp. 1436-1453.

Aibinu, A., Dassanayake, D., Chan, T. and Thangaraj, R , 2015, "Cost estimation for electric light and power elements during building design: A neural network approach", *Engineering, Construction and Architectural Management*, Vol. 22, No. 2, pp. 190-213,

Akaike, H. 1974, "A new look at the statistical model identification", Automatic Control, *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, pp. 716-723.

Akintoye, A. 2000, "Analysis of factors influencing project cost estimating practice", *Construction Management and Economics*, Vol. 18, No. 1, pp. 77-89.

An, S.-H., Kim, G.-H. and Kang, K.-I. 2007a, "A case-based reasoning cost estimating model using experience by analytic hierarchy process". *Building and Environment*, Vol. 42, No. 7, pp. 2573-2579.

An, S.-H., Park, U.-Y., Kang, K.-I., Cho, M.-Y. and Cho, H.-H. 2007b, “Application of support vector machines in assessing conceptual cost estimates”. *Journal of Computing in Civil Engineering*, Vol. 21, No. 4, pp. 259-264.

Andersen, C.M. and Bro, R. 2010, “Variable selection in regression - a tutorial”. *Journal of Chemometrics*, Vol. 24, No. 11-12, pp. 728-737.

Belsley, D.A., Kuh, E. and Welsch, R.E. 2005, *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.

Burnham, K.P. and Anderson, D.R. 2002, *Model selection and multimodel inference: a practical information-theoretic approach*. Springer.

Cheung, F.K, and Skitmore, M**.** 2006a**.** **“**A modified storey enclosure model.” *Construction Management and Economics* **24**(4) 391-406.

Cheung, F.K. and Skitmore, M. 2006b, “Application of cross validation techniques for modelling construction costs during the very early design stage”. *Building and Environment*, Vol. 41, No. 12, pp. 1973-1990.

Cheung, S.O., Wong and P.S.P., Fung, A.S., Coffey, W. 2006, “Predicting project performance through neural networks”. *International Journal of Project Management*, Vol. 24, No. 3, pp. 207-215.

Chou, J.-S. and Tseng, H.-C. 2011, “Establishing expert system for prediction based on the project-oriented data warehouse”. *Expert Systems with Applications*, Vol. 38, No. 1, pp. 640-651.

Farrar, D.E. and Glauber, R.R. 1967. “Multicollinearity in regression analysis: the problem revisited”. *The Review of Economic and Statistics*, Vol. 49, No. 1, pp. 92-107.

Flyvbjerg, B., Bruzelius, N. and Rothengatter, W. 2003, *Megaprojects and risk: An anatomy of ambition*. Cambridge University Press.

Friendly, M. and Kwan, E. 2009, “Where's Waldo? Visualizing collinearity diagnostics”. *The American Statistician*, Vol. 63, No. 1, pp. 56-65.

Guerrero, M.A., Villacampa, Y. and Montoyo, A. 2014, “Modeling construction time in Spanish building projects”. *International Journal of Project Management*, Vol. 32, No. 5, 861-873.

Hassim, S., Muniandy, R., Alias, A. and Abdullah, P. 2018, “Construction tender price estimation standardization (TPES) in Malaysia: Modeling using fuzzy neural network”. *Engineering, Construction and Architectural Management*, Vol. 25, No. 3, pp. 443-457.

Hatamleh, M., Hiyassat, M., Sweis, G. and Sweis, R. 2018, “Factors affecting the accuracy of cost estimate: case of Jordan”. *Engineering, Construction and Architectural Management*, Vol. 25, No. 1, pp. 113-131.

Johnson, J.B. and Omland, K.S. 2004, “Model selection in ecology and evolution”. *Trends in Ecology & Evolution*, Vol. 19, No. 2, pp. 101-108.

Kaka, A.P. 1996, "Towards more flexible and accurate cash flow forecasting". *Construction Management and Economics*, Vol. 14, No. 1, pp. 35-44.

Kim, G.-H., An, S.-H. and Kang, K.-I. 2004, "Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning". *Building and Environment*, Vol. 39, No. 10, pp. 1235-1242.

Kim, K.J. and Kim, K. 2010, "Preliminary cost estimation model using case-based reasoning and genetic algorithms". *Journal of Computing in Civil Engineering*, Vol. 24, No. 6, pp. 499-505.

Kim, S.-Y., Choi, J.-W., Kim, G.-H. and Kang, K.-I. 2005, "Comparing cost prediction methods for apartment housing projects: CBR versus ANN". *Journal of Asian Architecture and Building Engineering*, Vol. 4, No. 1, pp. 113-120.

Kullback, S. and Leibler, R.A. 1951, "On information and sufficiency". *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79-86.

Li, H., Shen, Q. and Love, P.E. 2005, "Cost modelling of office buildings in Hong Kong: an exploratory study". *Facilities*, Vol. 23, No. 9/10, pp. 438-452.

Liu, R., Kuang, J., Gong, Q. and Hou, X. 2003, "Principal component regression analysis with SPSS". *Computer Methods and Programs in Biomedicine*, Vol. 71, No. 2, pp. 141-147.

Love, P.E. and Li, H. 2000, "Quantifying the causes and costs of rework in construction". *Construction Management and Economics*, Vol. 18, No. 4, pp. 479-490.

Lowe, D.J., Emsley, M.W. and Harding, A. 2006, "Predicting construction cost using multiple regression techniques". *Journal of Construction Engineering and Management*, Vol. 132, No. 7, pp. 750-758.

Masrom, M.A., Skitmore, M. and Bridge, A. 2013, "Determinants of contractor satisfaction". *Construction Management and Economics*, Vol. 31, No. 7, pp. 761-779.

Merrow, E.W., Chapel, S.W. and Worthing, C. 1979, *Review of cost estimation in new technologies: implications for energy process plants*. RAND Corp., Santa Monica, CA (USA).

Min, J. H., and Lee, Y. C. 2005, "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters". *Expert Systems with Applications*, Vol. 28, No. 4, pp. 603-614.

Moore, A.W. 2001, *Cross-validation for detecting and preventing overfitting*. School of Computer Science Carnegie Mellon University.

Movahedian Attar, A., Khanzadi, M., Dabirian, S. and Kalhor, E. 2013, "Forecasting contractor's deviation from the client objectives in prequalification model using support vector regression". *International Journal of Project Management*, Vol. 31, No. 6, pp. 924-936.

Næs, T. and Martens, H. 1988, "Principal component regression in NIR analysis: viewpoints, background details and selection of components". *Journal of Chemometrics*, Vol. 2, pp. 155-167.

Neter, J., Wasserman, W. and Kutner, M.H. 1989, *Applied linear regression models*. Irwin, Homewood, Ill.

Newton, S. 1991, "An agenda for cost modelling research". *Construction Management and Economics*, Vol. 9, No. 2, pp. 97-112.

Ng, S.T, Skitmore, M., Wong, K.F. 2008. "Using genetic algorithms and linear regression analysis for private housing demand forecasts". *Building and Environment*, Vol. 43, No. 6, pp. 1171-1184.

Ngoc, D.C. and Long, L.H. 2018. "Revisiting storey enclosure method for early estimation of structural building construction cost". *Engineering, Construction and Architectural Management*, Vol. 25, No. 7, pp. 877-895.

O'brien, R.M. 2007, "A caution regarding rules of thumb for variance inflation factors". *Quality & Quantity*, Vol. 41, No. 5, pp. 673-690.

Oberlender, G.D. and Trost, S.M. 2001, "Predicting accuracy of early cost estimates based on estimate quality". *Journal of Construction Engineering and Management*, Vol. 127, No. 3, pp. 173-182.

Ökmen, Ö. and Öztaş, A. 2015 "Scenario based evaluation of a cost risk model through sensitivity analysis". *Engineering, Construction and Architectural Management*, Vol. 22, No. 4, pp. 403-423.

Posada, D. and Buckley, T.R. 2004, "Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests". *Systematic Biology*, Vol. 53, No. 5, pp. 793-808.

Rivals, I. and Personnaz, L. 1999, "On cross validation for model selection". *Neural Computation*, Vol. 11, No. 4, pp. 863-870.

Salah, A. and Moselhi, O. 2015, "Contingency modelling for construction projects using fuzzy-set theory". *Engineering, Construction and Architectural Management*, Vol. 22, No. 2, pp. 214-241.

Skitmore, M. 1985, *The influence of professional expertise in construction price forecast*. University of Salford, Salford, UK.

Skitmore, M. 1987, "The effect of project information on the accuracy of building price forecasts", in Brandon Peter S. (Ed.) *Building Cost Modelling and Computers*, E and F.N. Spon Ltd., London, pp. 327-336.

Skitmore, M. and Marston, V. 1999, *Cost modelling*. E&FN Spon, London

Skitmore, M. and Patchell, B. 1990, "Developments in contract price forecasting and bidding techniques". In Brandon, Peter (Ed.) *Quantity Surveying Techniques: New Directions*. BSP Professional Books, Oxford, pp. 75-120.

Skitmore, M., Stradling, S., Tuohy, A. and Mkwezalamba, H. 1990, *The accuracy of construction price forecasts*. University of Salford, Salford.

Son, H., Kim, C. and Kim, C. 2012, "Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables". *Automation in Construction*, Vol. 27, pp. 60-66.

Soutos, M. and Lowe, D. 2011, "Elemental cost estimating: current UK practice and procedure". *Journal of Financial Management of Property and Construction*, Vol. 16, No. 2, pp. 147-162.

Stoy, C. and Schalcher, H.-R. 2007, "Residential building projects: building cost indicators and drivers". *Journal of Construction Engineering and Management*, Vol. 133, No. 2, pp. 139-145.

Sullivan, J. and Joyce, P. 2005. "Model selection in phylogenetics". Annual Review of Ecology, *Evolution, and Systematics*, Vol. 36, pp. 445-466.

Tah, J.H.M., Thorpe, A. and McCaffer, R. 1994, "A survey of indirect cost estimating in practice". *Construction Management and Economics*, Vol. 12, No. 1, pp. 31-36.

Vapnik, V. 2000. *The nature of statistical learning theory*. Springer Science & Business Media.

Vigneau, E., Devaux, M., Qannari, E. and Robert, P. 1997, "Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration". *Journal of Chemometrics*, Vol. 11, No. 3, pp. 239-249.

Williams, T., Lakshminarayanan, S. and Sackrowitz, H. 2005, "Analyzing bidding statistics to predict completed project cost", *Proceedings of 2005 International Conference on Computing in Civil Engineering*, pp. 1-10.

Williams, T.P. 2003, "Predicting final cost for competitively bid construction projects using regression models". *International Journal of Project Management*, Vol. 21, No. 8, pp. 593-599.

Xiao, X., Wang, F., Li, H., and Skitmore, M. 2018, "Modelling the stochastic dependence underlying construction cost and duration". *Journal of Civil Engineering and Management*, Vol. 24, No. 6, pp. 444-456

Xiong, B., Skitmore, M. and Xia, B. 2015, "A critical review of structural equation modeling applications in construction research". *Automation in Construction*, Vol. 49 Part A, pp. 59-70.

Xu, L.D. 1994. "Case based reasoning". *Potentials*, *IEEE*, Vol. 13, No. 5, pp. 10-13.

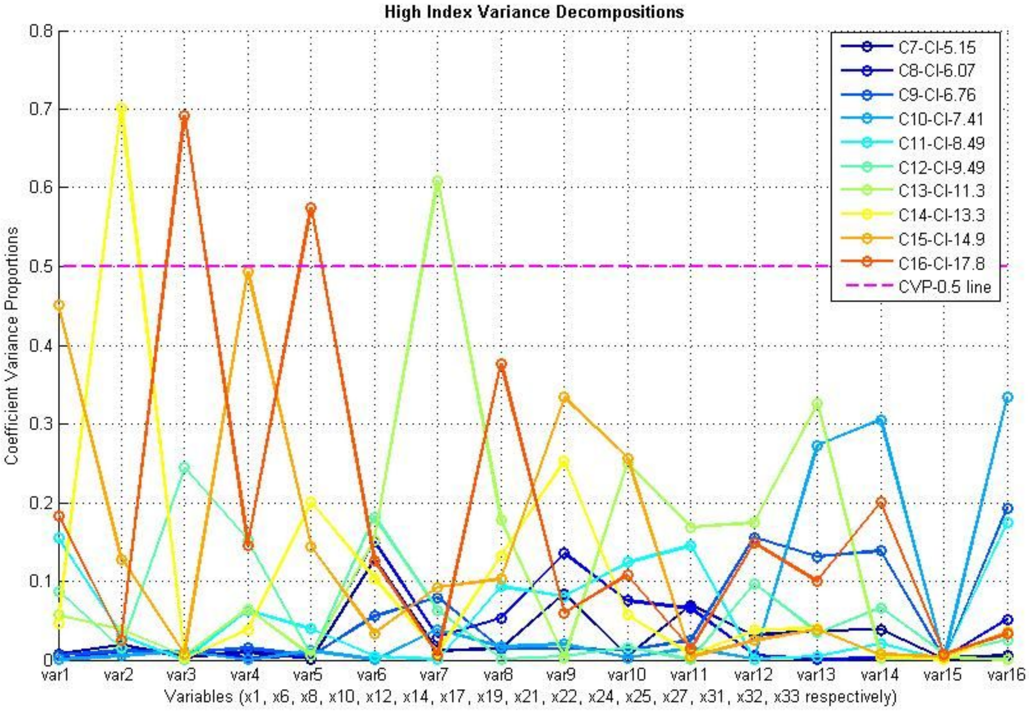

Figure 1: Collinearity diagnostics of Model 3

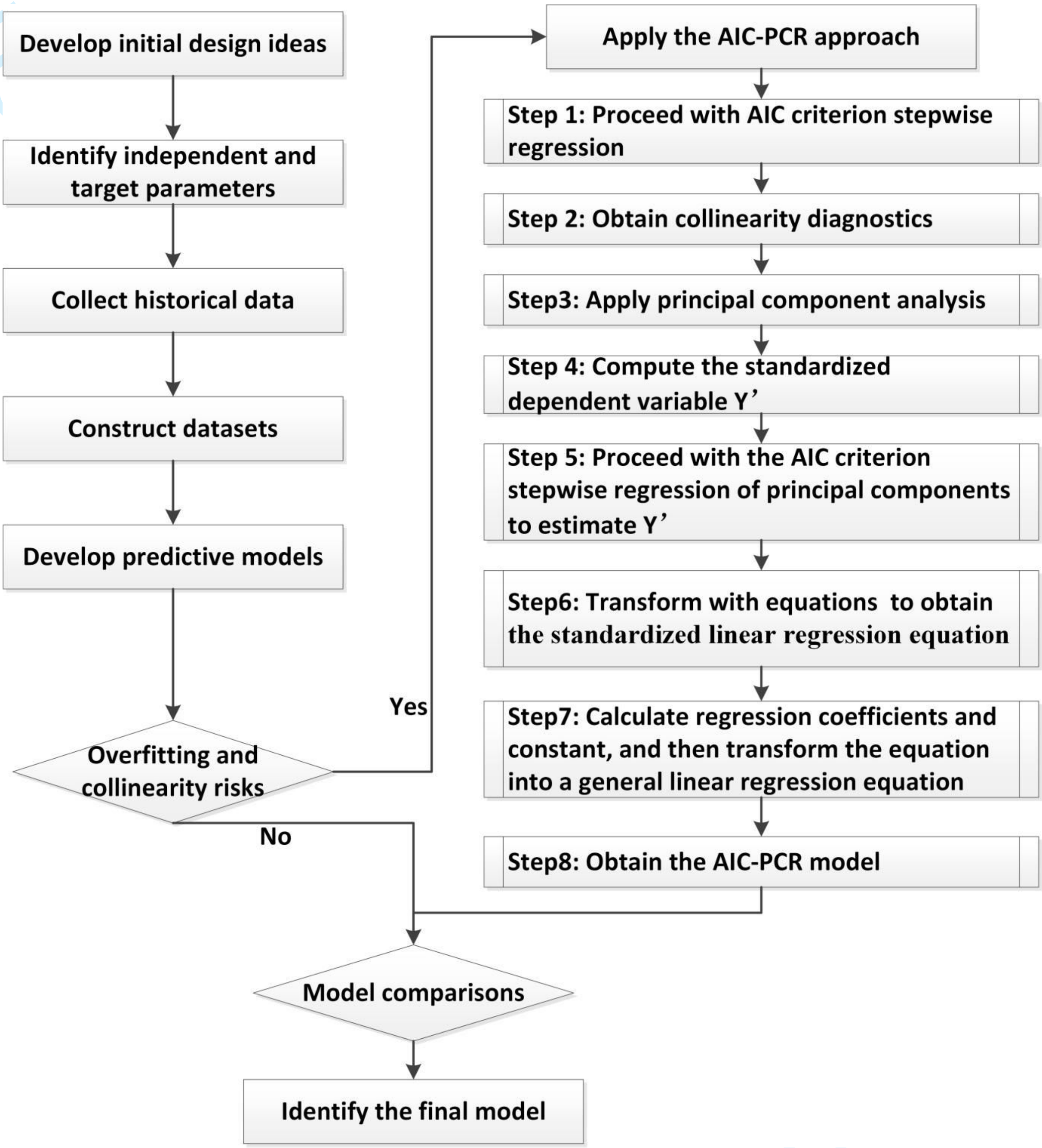
Develop initial design ideas
Identify independent and target parameters
Collect historical data
Construct datasets
Develop predictive models
Overfitting and collinearity risks
Yes
No
Apply the AIC-PCR approach
Step 1: Proceed with AIC criterion stepwise regression
Step 2: Obtain collinearity diagnostics
Step3: Apply principal component analysis
Step 4: Compute the standardized dependent variable Y'
Step 5: Proceed with the AIC criterion stepwise regression of principal components to estimate Y'
Step6: Transform with equations to obtain the standardized linear regression equation
Step7: Calculate regression coefficients and constant, and then transform the equation into a general linear regression equation
Step8: Obtain the AIC-PCR model
Model comparisons
Identify the final model

Figure 2: Estimation modelling framework and the AIC-PCR procedure

Table 1. Sample descriptions - part 1.

Category	Type	Frequency	Percentage
Building function	Primary schools	86	42.26%
	Secondary schools	64	31.4%
	Nursery schools	29	14. 2%
	Special schools	13	6.4%
	Sixth form/tertiary colleges	12	5.9%
Structure	Brick construction	70	34.3%
	Steel framed	115	56.4%
	Timber framed	15	7.4%
	Concrete framed	3	1.5%
	Unspecified	1	0.5%
Selection method	Selected competition	164	80.4%
	Open competition	10	4.9%
	Design and build - competitive	12	5.9%
	Two stage tendering	12	5.9%
	Unspecified	6	2.9%

Table 2: Sample descriptions - part 2.

Category	Description
Gross floor area (m2)	Range from 64 to 19670; *M*=1471.64; *SD*=2209.57
Stories	Range from 1 to 4; *M*=1.41; *SD*=0.60
Schedule (months)	Unspecified: 87
	Remaining: range from 5 to 32; *M*=10.83; *SD*=4.27
Ground condition	Unspecified: 23
	Bad(1)-Moderate(3)-Good(5); *M*=3.74; *SD*=1.42
Work space	Unspecified: 21
	Highly restricted(=1)-Restricted(=3)-Unrestricted(=5); *M*=3.92; *SD*=1.24
Site access	Unspecified: 17
	Highly restricted(1)-Restricted(3)-Unrestricted(5); *M*=3.71; *SD*=1.20
Market condition	Unspecified: 49
	Low competitive(1)-Less competitive(2)-Average(3)-Competitive(4)-Highly competitive(5); *M*=3.99; *SD*=0.79
Air Conditioning	Yes=1; No=0; 26 cases are 1; 178 cases are 0
Preliminaries (£)	Range from 0 (1 case) to 3,391,713; *M*=318,448; *SD*=443,345
Preliminaries/GFA	Range from 0 (1 case) to 702; *M*=251; *SD*=116

Table 3: Elemental cost items framework.

Elemental cost items	Code	Elemental cost items	Code
1 Substructure	x_1	5F Space heating and air treatment	x_{19}
2A Frame	x_2	5G Ventilating systems	x_{20}
2B Upper floors	x_3	5H Electrical installations	x_{21}
2C Roof	x_4	5I Gas installations	x_{22}
2D Stairs	x_5	5J Lift and conveyor installations	x_{23}
2E External walls	x_6	5K Protective installations	x_{24}
2F Windows and external doors	x_7	5L Communications installations	x_{25}
2G Internal walls and partitions	x_8	5M Special installations	x_{26}
2H Internal doors	x_9	5N Builder's work in connection	x_{27}
2 Superstructure		5O Builder's profit and attendance	x_{28}
3A Wall finishes	x_{10}	5 Services	
3B Floor finishes	x_{11}	6A Site works	x_{29}
3C Ceiling finishes	x_{12}	6B Drainage	x_{30}
3 Internal finishes		6C External services	x_{31}
4 Fittings	x_{13}	6D Minor building works	x_{32}
5A Sanitary appliances	x_{14}	6 External works	
5B Services equipment	x_{15}	7 Contingencies	x_{33}
5C Disposal installations	x_{16}	8 Preliminaries	y
5D Water installations	x_{17}		
5E Heat source	x_{18}		

Table 4: Developed regression models.

Models	Equations
1	$\hat{y} = 25220.000 + 0.396x_1 - 0.373x_6 - 0.915x_8 - 1.381x_{10} + 2.391x_{12} + 1.069x_{14} + 2.621x_{17} + 0.604x_{19} + 0.760x_{21} - 4.892x_{22} + 0.690x_{24} + 1.497x_{25} - 1.586x_{27} - 1.299x_{31} + 0.135x_{32}$
2	$\hat{y} = 27296.000 + 0.322x_1 + 0.276x_2 + 0.103x_4 - 0.390x_6 - 0.920x_8 - 0.602x_9 - 1.558x_{10} + 0.292x_{11} + 2.413x_{12} + 0.837x_{14} - 0.511x_{15} + 3.050x_{17} + 0.528x_{19} + 0.725x_{21} - 7.454x_{22} + 0.382x_{24} + 1.336x_{25} - 1.692x_{27} - 3.903x_{28} + 0.082x_{29} - 1.354x_{31} + 0.118x_{32} + 0.186x_{33}$
3	$\hat{y} = 22722.000 + 0.418x_1 - 0.385x_6 - 0.885x_8 - 1.364x_{10} + 2.407x_{12} + 0.937x_{14} + 2.748x_{17} + 0.576x_{19} + 0.738x_{21} - 5.893x_{22} + 0.568x_{24} + 1.453x_{25} - 1.665x_{27} - 1.396x_{31} + 0.136x_{32} + 0.200x_{33}$
4	$\hat{y}$ $= 36079.000 + 0.239x_1 + 0.313x_2 + 0.397x_3 + 0.132x_4 - 0.148x_5 - 0.402x_6 - 0.069x_7 - 0.943x_8 - 0.58$ $x_9 - 1.468x_{10} + 0.245x_{11} + 2.210x_{12} - 0.051x_{13} + 1.151x_{14} - 0.613x_{15} + 0.088x_{16} + 2.908x_{17} - 0.440x_{18} + 0.539x_{19} + 0.205x_{20} + 0.743x_{21} - 7.191x_{22} + 0.665x_{23} + 0.375x_{24} + 1.239x_{25} + 0.186x_{26} - 1.129x_{27} - 4.189x_{28} + 0.081x_{29} + 0.015x_{30} - 1.400x_{31} + 0.119x_{32} + 0.159x_{33}$
PCR	$\hat{y} = 28152.762 + 0.366x_1 - 0.390x_6 - 0.804x_8 - 1.531x_{10} + 2.439x_{12} + 1.320x_{14} + 2.463x_{17} + 0.607x_{18} + 0.753x_{21} - 4.395x_{22} + 0.444x_{24} + 1.612x_{25} - 1.764x_{27} - 1.241x_{31} + 0.110x_{32}$
AIC-PCR	$\hat{y}$ $= 26801.134 + 0.430\ x_1 - 0.387\ x_6 - 0.857\ x_8 - 1.314\ x_{10} + 2.426\ x_{12} + 1.010\ x_{14} + 2.643\ x_{17} + 0.587\ x$ $+ 0.718\ x_{21} - 6.337\ x_{22} + 0.473\ x_{24} + 1.700\ x_{25} - 1.881\ X_{27} - 1.419\ x_{31} + 0.102\ x_{32} + 0.132\ x_{33}$

Table 5: Comparison of results for Application 1, price estimating.

Models	RMSE	MAE	MAPE
AIC-PCR	114061.7251	74954.31441	0.419
PCR	115064.3467	76398.0302	0.425
ANN	194183.5763	89780.4847	0.623
KNN	260417.2453	128660.2191	0.419
PCA-SVR	211449.8781	103757.67	0.621