Forecasting the Final Cost of Iraqi Public School Projects Using Regression Analysis

Dr. Zeyad S. M. Khaled
College Engineering, Alnahrian University/ Baghdad.
Dr. Qais Jawad Frayyeh
Building and Construction EngineeringDep, University of Technology/ Baghdad.
Gafel Kareem Aswed
Building and Construction EngineeringDep, University of Technology/ Baghdad.
Email:gafelkareem@gmail.com

Received on:29/5/2014 & Accepted on:8/1/2015

ABSTRACT
The actual final cost of public school building projects, like other construction projects, is unknown to the owner till the final account statement is prepared. An attempt to predict the final cost of such projects before work starts, using backward elimination regression analysis technique is carried out. The study covers two story (12 classes) school projects awarded by the lowest bid system. Records of (65) school projects completed during (2007-2012) are employed to develop and verify the regression model. Based on experts' convictions, nine factors are considered to have the most significant impact on the final cost. Hence they are used as model input parameters. These factors are: awarded bid price, average bid price, estimated cost, contractor rank, resident engineer experience, project location, number of bidders, year of contracting, and contractual project duration. It was found that the developed regression model have the ability to predict the final cost (FC) for school projects, as an output, with a very good accuracy having a correlation coefficient (R) of (93%), determination coefficient (R²) of (86.5%) and average accuracy percentage of (92.02%).

Keywords: cost estimation, cost modeling, regression analysis, and school project.
INTRODUCTION

Construction projects costs are influenced by several factors. These factors are related to project characteristics, construction teams, and market conditions. When unexpected events occur during the execution phase of construction projects, their final costs are driven up. Most of such events are uncontrollable factors that increase the gap between the contract award price and the final completion cost. It is greatly important that the client should know what contingencies he must have in hand to ensure his project final completion in time. Lack of information about these factors, lack of relevant data, and weak expectations of possible circumstances to be faced by the project are the main challenges facing researchers in this essence. This research attempts to use real measurable parameters, to be in hand before the project starts, as predictors for the expected final cost of school projects.

Research Objectives

This research aims at the following objectives:

• To explore the factors that can be used to predict the final cost of school building projects before starting works.
• To raise the efficiency of estimating initial costs using data already in hand.
• To build a mathematical model using multiple regression analysis to predict cost deviation in school building projects before starting works.

Research Hypothesis

At the project start phase, it can be said that awarded bid price, average bid price, estimated cost, contractor rank, resident engineer experience, project location, number of bidders, year of contracting, and contractual project duration are good predictors to the final cost of public school building projects before starting works.

Research Justification

The reasons for carrying out this research are:

• The large number of under construction school projects accompanied with everlasting cost overrun and the ever growing demand on additional school buildings in Iraq.
• The need of knowing an accurate anticipated final cost of a construction projects before starting works, is highly essential in budgeting concerns, especially in contingency allocation.

Suitability Of Multiple Regression

The objective of most parametric costs estimating approaches is to use some historical cost data and try to find a functional relationship between changes in cost and the factors affecting the final cost. The regression technique is a statistical modeling method that can be used for analysis and prediction in different knowledge domains. Multiple regression estimation models are well established and widely used.
in cost estimation. They are effective due to their well-defined mathematical procedure, as well as being able to explain the significance of each variable and the relationships between them. Basically, regression models are intended to find the linear combination of variables which best correlates with dependent variables. The general regression equation is expressed as follows:

\[ Y = A_0 + A_1 I_1 + A_2 I_2 + \ldots + A_n I_n \]  

(1)

Where

- \( Y \) is the total estimated final result,
- \( A_0 \) is a constant estimated by regression analysis,
- \( A_1, A_2, \ldots, A_n \) are coefficients also estimated by regression analysis, given the availability of some relevant data \( I_1, I_2, \ldots, I_n \) as measured distinguishable variables that may help in estimating \( Y \) [2].

**Literature Review**

Literature review shows a variety of ways used to predict the project final cost and deviations. Many variables were used as predictors in those studies. Williams [3] concluded five mathematical models to predict the final cost of highway construction using low bids in five states in USA as independent variable. From competitive bidding of highway construction projects, the low bid price and the cost of the completed contract were obtained for each project. These models aimed at predicting the project final cost according to the low bid as the only input.

Wibowo and Wuryanti [4] studied education building projects in Indonesia to prepare early stage cost estimate models. They found that the estimated cost at earlier stages could be predicted according to the total project area.

Olatunji [5] collected data of (137) public contract projects executed between (2003) and (2007) in Nigeria. Lowest/winner bid, average bid, consultant’s cost estimate, gross floor area were the model variables to predict the final construction cost. The conducted regression model has an adjusted \( R^2 \) value of (0.949).

Mahamid and Amund [6] investigated the statistical relationship between actual and estimated cost of road construction projects. Data collected from (169) road construction projects awarded in the West Bank in Palestine over the years (2004–2008) were analyzed. The study concluded that (100%) of road projects in Palestine suffer from cost deviation. They developed a regression model with a coefficient of determination \( R^2 \) of (0.96).

Bedford [7] studied the risks of excessive competition in the Canadian public sector that award contracts solely to the low bid. It has been concluded that the bidding process is a good indicator to the final cost and possible cost escalations.

Ganiyu and Zubairu [8] developed a predictive cost model for public building projects in Nigeria using principal components regression. The study showed that the project cost basically depends on factors related to; adequacy of equipment, experience in similar projects, time allowed for bidding, level of technology, client commitment to time, repetitive work, design complexity, communications, project scope, construction complexity, and previous relationship with the client.

Mohd et al. [9] studied the historical data of (83) school projects in the Malaysian public sector. Multiple regression analysis was used to predict the effect of the lowest bid, average tender price, and the winning tender in the interpretation of the deviation in cost estimation. The regression model from mean bids showed that the project size,
number of tenders, type of schools, and location are the best-fitted predictors to explain biased estimates.

Aziz\cite{10} investigated and ranked factors perceived to affect cost variation in the Egyptian wastewater projects. It was discovered that factors such as: lowest bid procurement method, additional works, bureaucracy in bidding, tendering method, wrong method of cost estimation, and funding problems were crucial in causing cost variation, while, inaccurate cost estimation, mode of payments, unexpected ground conditions, inflation, and price fluctuation are less important.

Datacollection

The initial parameters that are intended to be used in the model were collected from the literature review of previous studies as shown in Table (1). A questionnaire form has been directed to (50) local experts in order to determine the most significant factors in predicting the final cost of school projects before it starts. Fifty questionnaires were directed to owner's representative engineers in the public sector. Thirty two respondents forms, forming (64%) of the total number of questionnaires, have successfully been submitted. The respondents were asked to select the parameters that they believe is most important in developing the mathematical model. Likert’s scale of five ordinal measures of importance (from 1 to 5) is used. An effort had been made to limit the number of variables that affects the cost model. Nine out of eleven variables were identified and analyzed as independent variables of the regression equation based on the respondents’ opinions. These variables are: \(I_1\) accepted bid price, \(I_2\) average bid price, \(I_3\) estimated cost, \(I_4\) contractor rank, \(I_5\) experience of the resident engineer (R.E), \(I_6\) location of project, \(I_7\) number of bidders, \(I_8\) year of contracting, and \(I_9\) project duration. To build the regression model a historical data is collected from (68) completed schools projects in Karbala province. These projects are executed during the years (2007-2012). By eliminating the incomplete records, the number of projects put under consideration for the final selected model became (65) projects. The projects were awarded under the lowest bid tendering system having the same design and number of classrooms.

<table>
<thead>
<tr>
<th>InfluentialFactors</th>
<th>Supporting Previous Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I_1) Accepted bid price</td>
<td>Elhag [11], Williams [3], Olatunji [5]</td>
</tr>
<tr>
<td>(I_2) Average bid price</td>
<td>Olatunji [5], (Mohd et al. [9])</td>
</tr>
<tr>
<td>(I_3) Estimated cost</td>
<td>Olatunji [5], Mahamid and Amund [6]</td>
</tr>
<tr>
<td>(I_4) Contractor rank</td>
<td>Ganiyu and Zubairu [8]</td>
</tr>
<tr>
<td>(I_5) Experience of R. E.</td>
<td>Ganiyu and Zubairu [8]</td>
</tr>
<tr>
<td>(I_6) Project location</td>
<td>Mohd et al. [9],</td>
</tr>
<tr>
<td>(I_7) Number of bidders</td>
<td>Mohd et al. [9]</td>
</tr>
<tr>
<td>(I_8) Year of contracting</td>
<td>Al-zwainy [12]</td>
</tr>
<tr>
<td>(I_9) Contractor duration</td>
<td>Elhag [11]</td>
</tr>
<tr>
<td>(I_{11}) Owner duration</td>
<td>Ayman et al. [13]</td>
</tr>
<tr>
<td>(I_{12}) Second lowest bid</td>
<td>Bordat et al. [14]</td>
</tr>
</tbody>
</table>

Modelformulation

Previous studies showed different methods used to study the relation between the final cost and factors believed to influence that final cost. In this research a back
elimination regression technique is adopted to analyze historical cost data in order to provide a powerful model to assist budgeting and cost estimating before work starts. The Statistical Package for Social Science SPSS and MS Excel are used to develop a suitable model. In order to remove a linear trend from the data, transformations by taking the natural logs of some of the variables are applied. Then a simple linear model is developed using in each run, the natural log of the final project cost (FC) as the dependent variable, the natural logs of: accepted bid price ($I_1$), average bid price ($I_2$), estimated cost ($I_3$), and the untransformed parameters of: contractor rank ($I_4$), experience of R.E. ($I_5$), location of project ($I_6$), number of bidders ($I_7$), year of contracting ($I_8$), contractor duration ($I_9$) as independent variables. SPSS (version 20) is used for data analysis. Backward elimination technique is adopted to develop the regression model as in Tables (2) and (3). The procedure of this technique is to enter all nine variables in the model equation first, then sequentially remove the variable with the smallest partial correlation with the dependent variable in each run.

### Table (2): Summary of Analysis Results

<table>
<thead>
<tr>
<th>FC-Model</th>
<th>R</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>Standard Error</th>
<th>Residual Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.930a</td>
<td>0.865</td>
<td>0.841</td>
<td>0.1232519</td>
<td>0.015</td>
</tr>
<tr>
<td>2</td>
<td>0.930b</td>
<td>0.865</td>
<td>0.844</td>
<td>0.1220738</td>
<td>0.015</td>
</tr>
<tr>
<td>3</td>
<td>0.930c</td>
<td>0.865</td>
<td>0.847</td>
<td>0.1209313</td>
<td>0.015</td>
</tr>
<tr>
<td>4</td>
<td>0.930d</td>
<td>0.865</td>
<td>0.850</td>
<td>0.1198686</td>
<td>0.014</td>
</tr>
<tr>
<td>5</td>
<td>0.929e</td>
<td>0.862</td>
<td>0.850</td>
<td>0.1199525</td>
<td>0.014</td>
</tr>
<tr>
<td>6</td>
<td>0.927f</td>
<td>0.860</td>
<td>0.850</td>
<td>0.1199546</td>
<td>0.014</td>
</tr>
<tr>
<td>7</td>
<td>0.926g</td>
<td>0.857</td>
<td>0.849</td>
<td>0.1201290</td>
<td>0.014</td>
</tr>
</tbody>
</table>

Where
- a. Predictors: (Constant), $I_9$, $I_4$, $I_6$, $I_7$, $I_5$, $I_2$, $I_8$, $I_3$, $I_1$
- b. Predictors: (Constant), $I_9$, $I_4$, $I_6$, $I_7$, $I_5$, $I_8$, $I_3$, $I_1$
- c. Predictors: (Constant), $I_8$, $I_4$, $I_6$, $I_7$, $I_5$, $I_3$, $I_1$
- d. Predictors: (Constant), $I_9$, $I_4$, $I_7$, $I_5$, $I_3$, $I_1$ (the best model)
- e. Predictors: (Constant), $I_4$, $I_7$, $I_5$, $I_3$, $I_1$
- f. Predictors: (Constant), $I_4$, $I_7$, $I_3$, $I_1$
- g. Predictors: (Constant), $I_7$, $I_3$, $I_1$
- h. Dependent Variable: FC

### Table (3): Model Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.(p-Value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>2.032</td>
<td>1.235</td>
<td>1.646</td>
<td>0.106</td>
</tr>
<tr>
<td>$I_1$</td>
<td>0.344</td>
<td>0.147</td>
<td>0.358</td>
<td>2.343</td>
</tr>
<tr>
<td>$I_2$</td>
<td>0.569</td>
<td>0.147</td>
<td>0.360</td>
<td>8.860</td>
</tr>
<tr>
<td>$I_3$</td>
<td>-0.025</td>
<td>0.020</td>
<td>-0.067</td>
<td>-1.283</td>
</tr>
<tr>
<td>$I_4$</td>
<td>0.004</td>
<td>0.004</td>
<td>0.059</td>
<td>1.111</td>
</tr>
<tr>
<td>$I_5$</td>
<td>-0.028</td>
<td>0.013</td>
<td>-0.109</td>
<td>-2.096</td>
</tr>
<tr>
<td>$I_6$</td>
<td>0.00022</td>
<td>0.000</td>
<td>0.065</td>
<td>1.037</td>
</tr>
</tbody>
</table>
Forecasting the Final Cost of Iraqi Public School Projects Using Regression Analysis

Resulted Equation

After applying multipled regression analysis on the historical data of the whole (60) school projects, the resulted final construction cost estimation equation is:

\[ FC = 2.032 + 0.344I_1 + 0.569I_3 - 0.023I_4 + 0.004I_5 - 0.028I_7 + 0.00022I_9 \cdots (2) \]

This model is chosen based on the smallest Standard Error of Estimate which is (0.1198686) and the Residual Mean Square which is (0.014).

According to Tabachnick and Fidell[1] advice, the relative importance of the independent variables is assessed by examining their respective standardized coefficients i.e. Beta values in Table (3). Predictors with higher standardized coefficients such as: ln accepted bid price (I_1), ln estimated cost (I_3) and number of bidders (I_7) are more important to the regression equation than those with lower values such as contractor rank (I_2), experience of R.E. (I_5) and contractor duration (I_9). Therefore values of I_1, I_3, and I_7 indicate a highly significant regression fit. It can be concluded that I_1, I_3, and I_7 contribute significantly to the regression model. The small constants of I_5 and I_9 in the model equation refer to the small effect of experience of R.E and the contractor duration. The exclusion of the average bid price (I_2), location (I_6) and the year of contracting (I_8) parameters is because of their insignificance.

Multi-Collinearity Assessment

To assess multi-collinearity among the variables, tolerances and variance inflation factors (VIF) are examined as shown in Table (4). Tolerance refers to the proportion of the variance of that variable that is not accounted for by other predictors in the model and is calculated using the formula (1–R^2) for each variable. The range of tolerances is from (0) i.e. perfect collinearity, to (1) i.e. no collinearity. A tolerance with values less than (0.1) typically indicates a multi-collinearity problem. Variance inflation factor (VIF) is another index for the diagnostic of multi-collinearity which is just the inverse of the tolerance value. The high value of (VIF) for a variable indicates that there is a strong association between that variable and other remaining predictors [1]. Variables that have high tolerances will definitely have small variance inflation factors. A variance inflation factor in excess of (10) indicates a multi-collinearity problem [15]. Since the final cost model predictors have tolerances and (VIF) values that do not violate the aforementioned criteria, therefore, multi-collinearity is not a serious problem in this analysis.

<table>
<thead>
<tr>
<th>Model</th>
<th>VIF</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>I_1</td>
<td>9.180</td>
<td>.109</td>
</tr>
<tr>
<td>I_3</td>
<td>8.276</td>
<td>.121</td>
</tr>
<tr>
<td>I_4</td>
<td>1.070</td>
<td>.934</td>
</tr>
<tr>
<td>I_5</td>
<td>1.091</td>
<td>.916</td>
</tr>
<tr>
<td>I_7</td>
<td>1.069</td>
<td>.936</td>
</tr>
<tr>
<td>I_9</td>
<td>1.538</td>
<td>.650</td>
</tr>
</tbody>
</table>
Model Validation

One of the most important steps in developing a cost model is to test its accuracy and validity. This process is also referred to as the model validation. It involves testing and evaluating the developed model with some test or validation data. The validation data should be some representative data from the targeted population but haven't been used in the development of the model. In this study, the validation data is extracted from the same historical data file but for five randomly selected additional projects. They are not a part of the (60) projects used in the development of the model. The predicted cost of these five projects computed using the model equation are compared with real cost records and the results of this comparison is shown in Table (5). It is evident now that the model performed very well. Its predictions deviate only by (–8.702%) to (6.513%).

<table>
<thead>
<tr>
<th>FC (observed) in IQD</th>
<th>FC (predicted) in IQD</th>
<th>*Deviation %</th>
</tr>
</thead>
<tbody>
<tr>
<td>688482184.5</td>
<td>733325333.7</td>
<td>6.513335</td>
</tr>
<tr>
<td>1158046071</td>
<td>1057263708</td>
<td>-8.70279</td>
</tr>
<tr>
<td>832546804.6</td>
<td>760502532.7</td>
<td>-8.65348</td>
</tr>
<tr>
<td>1852866989</td>
<td>1877430643</td>
<td>1.325711</td>
</tr>
<tr>
<td>1594777396</td>
<td>1473269621</td>
<td>-7.61911</td>
</tr>
</tbody>
</table>

*Deviation % = \( \frac{(Predicted \ cost - Observed \ cost)}{Observed \ cost} \times 100 \) [16].

To assess the validity of the derived equation of the model for the final cost of school project (FC), the natural logarithm (Ln) of predicted values of (FC) are plotted against the natural logarithm (Ln) of real values for validation data set as shown in Figure (1).

![Figure (1): Comparison of Predicted and Actual Final Costs](image)

The coefficient of determination (R²) is found to be (97%), therefore it can be concluded that this model shows a good agreement with the actual measurements. It is finally clear that this model for school building projects in Iraq has the
generalization capability for any data set used within the range of data used in the development of the (MR) model.

**Model Evaluation**

The statistical measures used to measure the performance of the models included[17]:

**i.** Mean Percentage Error (MPE) which is one of the most important measures of accuracy of a proposed model. It is the mean of the absolute percentage differences between the predicted and the actual values:

\[
MPE = \left( \frac{1}{n} \sum_{j=1}^{n} \left( \frac{A-E}{A} \right) \right) \times 100
\]

Where:
- \( A \) = actual value
- \( E \) = estimated value (predicted value)
- \( n \) = total number of cases (5 for validation)

**ii.** Root Mean Squared Error (RMSE)

\[
RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (E-A)^2}
\] ...

**iii.** Mean Absolute Percentage Error (MAPE),

\[
MAPE = \frac{1}{n} \sum_{j=1}^{n} \left| \frac{A-E}{A} \right| \times 100
\]

Average accuracy percentage (AA %):

\[
AA\% = 100\% - MAPE
\]

**iv.** The Coefficient of Determination \( R^2 \)

**v.** The Coefficient of Correlation \( R \) is a measure that is used to determine the relative correlation and the goodness-of-fit between the predicted and observed data and show how well the model outputsmatch the target value.

These statistical measures of the regression Model (FC) in equation (2) are presented in Table (6).

<table>
<thead>
<tr>
<th>Table (6): Statistical Measures Results For Regression model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>MPE</td>
</tr>
<tr>
<td>RMSE</td>
</tr>
<tr>
<td>MAPE</td>
</tr>
<tr>
<td>AA%</td>
</tr>
<tr>
<td>( R^2 )</td>
</tr>
<tr>
<td>( R )</td>
</tr>
</tbody>
</table>

The (MAPE) and (AA) generated by (MR) model of (FC) are found to be (7.97%) and (92.02%) respectively. The \( R^2 \) value is (86.5%) which indicates that the most variability in the total cost is explained by the terms in the model. Therefore, it can be concluded that this (MR) model of (FC) shows good agreement with the actual measurements.

**Conclusions**

...
1. Backward elimination techniques is used to develop the regression model on the historical data of the school projects, resulting in the final construction cost prediction equation with ($R^2 = 86.5\%$).
2. The most significant model parameters are: accepted bid price ($I_1$), estimated cost ($I_3$), Contractor rank ($I_4$), resident engineer experience in years ($I_5$), number of bidders ($I_7$), and contractor duration ($I_9$).
3. The developed model shows a good agreement with the actual measurements based on the accuracy measurements.
4. It can easily calculate the expected cost deviation which is the difference between contractual sums and predicted final costs obtained from the developed regression model.
5. Other future work estimating parameters like cost per unit area, unit volume, classroom, or pupil can be calculated.

**Recommendations**

1. Data for another design of schools can be studied to further confirm the relationship between the independent parameters and final cost.
2. The developed model can be checked for applicability on other typical type of school projects such as (14), (16) and (18) classes.
3. Attention must be given to the documentation and feedback of data in order to achieve efficient and effective updated information and to audit a different accounting and financial procedures.

**REFERENCES**

[7]. Bedford, Thomas, “Analysis of the Low-Bid Award System in Public Sector Construction Procurement”, MSc. thesis in Civil Engineering, 2009, University of Toronto.