Wiley
Advances in Civil Engineering
Volume 2024, Article ID 1737352, 11 pages
https://doi.org/10.1155/2024/1737352

WILEY

Research Article

Modeling Cost-Estimation Factors for Public Building Projects with Hybrid Approach in Addis Ababa

Behailu Temesgen Habe, Lucy Feleke Nigussie, and Mamaru Dessalegn Belay

Construction Engineering and Management, Institute of Technology, Faculty of Civil and Environmental Engineering, Jimma University, Jimma, Ethiopia

Correspondence should be addressed to Lucy Feleke Nigussie; lucy.nigussie@ju.edu.et

Received 9 December 2022; Revised 28 April 2024; Accepted 16 May 2024; Published 27 May 2024

Academic Editor: Timothy O. Olawumi

Assessing the most important cost-influencing factors is essential for enhancing the predictive ability of cost estimation for building construction projects. The goal of this study is to examine and design a valid cost prediction model for assessing factors that impact the cost estimation of public buildings in Addis Ababa. This research solves these issues that typically arise in predictive cost estimation models in two major processes. First, the insights of 133 professionals gathered on the 38 cost-impacting elements, and 15 top factors design, time or cost, and parties' experience were determined. The suggested hybrid approach is based on the Akaike information criterion (AIC) and principal component regression (PCR) employed, coupling a stepwise linear regression model. According to the findings of the study, principal component analysis reduced important factors to 14 and efficiently solved the problem of multicollinearity with a variance inflation factor of less than 2, while stepwise cross-validation solved the overfitting problem at the lowest AIC. The cost prediction model sorted out five factors: design completion by the public body when bids are invited; completion of the project scope definition when bids are invited; level of construction complexity; importance of project completion within budget; and subcontractor experience and capability have all been identified as the main cost-determining factors. The study's contribution is the first approach (PCR–AIC) utilized in this work to explore numerous cost-estimating components, eliminate those that were related to one another, and identify the most crucial ones that consisted of the majority of the original variables' attributes.

1. Introduction

In Ethiopia, public authorities' cost estimates for construction projects frequently diverge from those provided by the contractor and/or the designer. These variations result from various practices, goals, and procedures. According to studies, one of the causes of cost overruns in the Ethiopian construction sector is the use of inaccurate cost estimation methodologies [1]. Interest in cost-estimating techniques and cost-influencing factor evaluation has grown as well [2, 3, 4]. The independent variable with a high degree of correlation is likely to be excluded from such models [5]. However, multiple regression is frequently employed by researchers to find elements that affect and estimate a project's cost. This resulted in a limited number of factors to be included in estimating project costs, and a prediction of project costs would not be accurate. As a result, there is an urgent need to tackle some of the core problems impeding the estimation of increased performance and viability, one of which is riddled with challenges presented by the prevailing wide variations between anticipated and actual project costs as a result of the absence of effective cost estimation techniques. Although it is possible to simply eliminate one or more predictors from a model to improve it [6], whether a variable is kept or removed should be based on the underlying theory [7]. Previous research, summarized by Xiong et al. [5], identified loss of information when deleting variables and collinearity in a model. To overcome this difficulty, researchers developed three widely used methods: ridge regression (RR), partial least squares regression (PLS), and principal component regression (PCR) [8]. Moreover, researchers applied powerful machine learning techniques, including ANN and hybrid models of ANN with fuzzy logic, CBR, and GA, to improve the accuracy of the estimation when compared to other methods [9]. As a fundamental form of CBR, K-nearest neighbor (KNN) and the hybrid model combining PCA and AIC were compared by Xiong et al. [5] for "eliminating variables" and "cost estimation accuracy," and it was shown that the hybrid model (PCA–AIC) improved the predictive cost model.

The previous research did not adequately address the issue of multicollinearity and overfitting concerns for a cost-estimating tool with acceptable predictability utilizing principal component analysis. Hence, this study seeks to answer the issues of multicollinearity and overfitting without the loss of the original characteristics of the cost estimation variables to get a reliable building project cost estimate. It also contributes to the limitations of the previous study conducted by Xiong et al. [5] by testing the applicability of the approach in the context of public building construction projects. It also addresses the gaps of Chan and Park's [3] and Ganiyu and Zubairu's [10] study, which employed a forecasted cost model to identify variables affecting construction costs employing PCR; however, these works did not adequately address the problem of multicollinearity and overfitting problems for good predictability of the cost estimation tool using the principal component analysis. As a result, the cost estimation model in this study used a hybrid approach of the Akaike information criterion (AIC) and PCR methods to assess the factors that influenced cost estimation and, as an outcome, attempt to develop a unique predictive cost model for public building projects in Addis Ababa.

2. Research Design and Methodology

2.1. Research Design. For this study period, descriptive and quantitative statistical research methods were used to answer the three research questions. To address the three research objectives, analysis and research findings are conducted throughout six stages. Stage 1: establish the data sources for the public building construction projects in Addis Ababa, analyze the survey, and present the results in charts and tables. Stage 2: determine cost-influencing factors from the historical data completed in the literature review. Stage 3: construct databases from various scholars and studies in the study area based on the evaluation and referencing of tools used in the research analysis. Stage 4: identifying the most important cost-influencing factors by using principal component analysis, and Stage 5 of the research process is followed. Stage 5: develop predictive models: undertake data analysis and statistical modeling using multivariate regression analysis and establish the correlations between project costs and factors that are important to cost estimation. Stage 6: test the predictive model for overfitting and collinearity risks, and then apply the PCR–AIC hybrid approach analysis if the risks exist and develop the final model.

2.2. Study Variables. The dependent variable for the study is construction cost. The independent variables for the study were 38 parameters categorized into three groups: design-related factors, time- or cost-related factors, and project parties' experience-related factors.

2.3. Sampling Size and Techniques. As the data on the finished public buildings were gathered from those firms willing to supply the information that the data must also reflect Addis Ababa, consequently the researcher utilized a nonprobabilistic purposive sampling approach. Around 34 public building construction projects were finished in the city throughout the previous 10 years, according to Birhanu [11]. As a result, skilled experts who have taken part in projects as consultants, contractors, or employers are chosen at random. With three experts from each of the three parties, there would be a total of 300 replies from the three parties mentioned above. As a result, it is believed that the Bartlett et al. [12], formulae were used to compute the relative necessary sample size. A total of 143 specialists who had participated in building construction projects that were completed during the previous 10 years made up the sample size for expert opinion to get an original insight on 38 parameters.

2.4. Research Validity and Reliability. The test results of Cronbach's alpha achieved an overall high of 0.860, indicating the overall reliability of the research instrument for factor analysis [13]. The data were further examined using the Kaiser–Meyer–Olkin (KMO) test for sample adequacy, which yielded a significant result of 0.696.

2.5. Principal Component Analysis. In this study, the principal components—common characteristics that significantly contribute to and are significant for estimating building costs —are extracted using the MATLAB program. These variables are chosen based on eigenvalues, which are used to gauge how much of the contribution common components make to the model. According to Kaming et al. [14], which was evaluated by Ganiyu and Zubairu [10], the total number of extracted components must be fewer than or equal to the number of original elements utilized in the model. The elements that have eigenvalues larger than or equal to 1 (eigenvalue 1), according to Ganiyu and Zubairu [10], are the most important ones that affect project cost. This study used the same selection criteria for principal components as the studies that came before it [3, 10, 11].

2.6. Regression Model Estimation. Regression analysis was carried out by Chan and Park [3], utilizing a parametric estimating strategy. Accordingly, when modeling is done using the K columns of X, L and K main components are employed, and the regression equation is as follows:

$$\widehat{y} = \alpha + \beta X C_L + \in, \tag{1}$$

where $\widehat{y}$ represents the fitted values of an N-dimensional response vector, α is a constant, X is an N rows by K columns $(N \times L)$ matrix of the original variables, C_L is a K rows by L column $(K \times L)$ matrix containing the eigenvectors from the selected principal components, β is an L-dimensional vector of unknown regression parameters, and $\in$ is a random vector that meets the basic normality assumptions of $E(\in) = 0$ and $\mathrm{var}(\in) = \delta^2$.

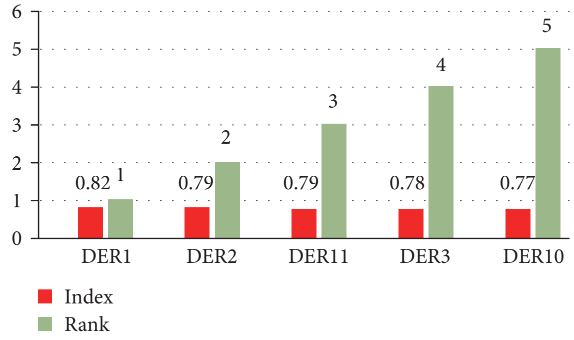

Figure 1: Top 5 ranked design-related factors (x-axis is factors, y-axis is rank and relative importance index (RII)).

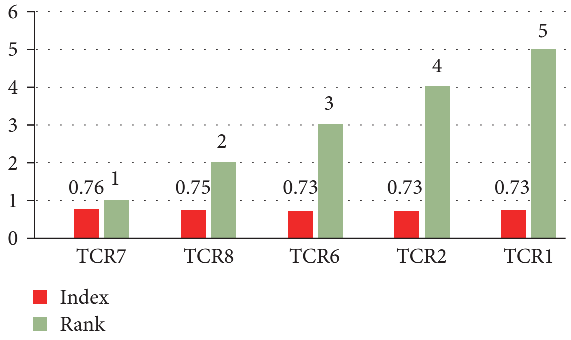

Figure 2: Top 5 ranked time-/cost-related factors (x-axis is factors, y-axis is rank and RII).

2.7. Prediction Model Selection Criteria. The AIC criterion was utilized in this work to prevent overfitting and the PCR method to address collinearity. To prevent overfitting with a relatively small sample size, this statistic measures the amount of information lost in the model fit when predictor variables are added [15], as reviewed by Xiong et al. [5].

3. Results and Discussions

3.1. Questionnaire Response Rate. For this objective, detailed questionnaires were prepared and delivered to significant stakeholders in the construction sector, including clients (project owners), contractors, consultants, and other professional organizations. To increase the scope of the analysis, a total of 179 online questionnaires were given to customers (project owners), consultants, and contractors, of which 133 were completed and 46 were invalid, yielding a rather high response rate of 74.3%. A total of 56 construction projects' costs were gathered, along with expert opinions on the cost-influencing factors from Addis Ababa's subcities various contractors, consultants, clients, and other professional institutes.

3.2. Factors that Affect Public Building Cost Estimation in Addis Ababa

3.2.1. Design-Related Factors. In this category, the "level of design complexity" has the highest respondent score and is the most important to cost estimation of public building construction projects, whereas the "presence of special issues" has the lowest score and is less important to cost estimation, according to respondents. Figure 1 illustrates the level of design complexity (DER1), construction complexity (DER2), design completion (by owner) when bids are invited (DER11), technological advancement (DER3), and project scope definition completion when bids are invited (DER10) as the top 5 design-related factors.

3.2.2. Time-/Cost-Related Factors. The time- and cost-related factors ranking their importance to cost estimation of public buildings indicated that "the consultant's level of construction sophistication" is highest, whereas "time given to the consultant to evaluate bids" had the lowest factors of their importance to cost estimation.

Figure 2 illustrates the top 5 most important factors selected by the respondents, which are the consultant's level of construction sophistication (TCR7), the owner's level of construction sophistication (TCR8), the bidding environment (TCR6), the importance of the project to be delivered (TCR2), and the importance of the project to be completed within budget (TCR1).

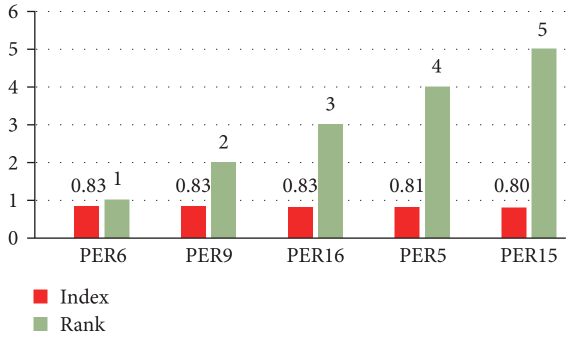

Figure 3: Top 5 ranked experience-related factors (x-axis is factors, y-axis is rank and RII).

3.2.3. Project Parties Experience-Related Factors. The relative importance of the ranking of project parties' experience-related factors indicates the highest factor important to cost estimation of public buildings is "Contractor‘s experience with similar size of project," whereas the lowest is "client experience with similar project," which is least important to cost estimation.

Figure 3 illustrates the top 5 most important factors selected by the respondents among the parties' experience, which include the contractor's experience with similar size of the project (PER6), communication among the project team (PER9), the adequacy of the contractor's plant and equipment (PER16), the contractor's experience with similar types of projects (PER5), and the contractor's staffing level (PER15).

3.3. Determining the Most Important Cost Influencing Factors

3.3.1. Design-Related Factors. Table 1 shows that the extracted components of the first principal component (PCDF1) explained 37.19% of the overall variance, whereas the second principal component (PCDF2) explained 10.8% of the remaining variation that the first component did not explain. The third principle (PCDF3) accounted for 7.88% of the dataset's variance, whereas the fourth principal component (PCDF4) contributed 7.24%.

Table 1: Total variance explained for design-related factors.

Factor	Initial eigenvalues			Extraction sums of squared loadings		
	Total	Percentage of variance	Cumulative	Total	Percentage of variance	Cumulative
DER1	4.71	37.19	37.19	4.71	37.19	37.19
DER2	1.41	10.80	47.99	1.41	10.80	47.99
DER3	1.02	7.88	55.87	1.02	7.88	55.87
DER4	0.98	7.24	63.11	0.98	7.24	63.11
DER5	0.90	6.54	69.65	—	—	—
DER6	0.76	5.75	75.40	—	—	—
DER7	0.70	5.22	80.62	—	—	—
DER8	0.60	4.63	85.25	—	—	—
DER9	0.55	4.20	89.46	—	—	—
DER10	0.44	3.47	92.92	—	—	—
DER11	0.34	2.82	95.74	—	—	—
DER12	0.31	2.30	98.04	—	—	—
DER13	0.26	1.96	100.00	—	—	—

Table 2: Total variance explained for time-/cost-related factors.

Factor ID	Initial eigenvalues			Extraction sum of squared loadings		
	Total	Percentage of variance	Cumulative	Total	Percentage of variance	Cumulative
TCR1	2.98	39.40	39.40	2.98	39.40	39.40
TCR2	1.68	18.20	57.61	1.68	18.20	57.61
TCR3	0.84	10.77	68.38	—	—	—
TCR4	0.70	9.95	78.33	—	—	—
TCR5	0.59	7.56	85.89	—	—	—
TCR6	0.34	5.45	91.35	—	—	—
TCR7	0.39	4.50	95.85	—	—	—
TCR8	0.48	4.15	100.00	—	—	—

3.3.2. Time Cost-Related Factors. Table 2 illustrates time- and cost-related factors. The first principal component's eigenvalue is 2.98, whereas the second component's eigenvalue is 1.68. The first principal component (PCTCF1) accounted for 39.4% of the total variance, while the second principal (PCTCF2) component explained 18.2% of what was not explained by the first component.

3.3.3. Project Parties Experience-Related Factors. Table 3 shows that the first main component eigenvalue is 4.95 for the parties' experience-related components, while the sixth component eigenvalue is 0.97. It was found that the first principal component (PCEF1) explained 27.54% of the overall variance, whereas the second principal component (PCEF2) explained 12.46% of the remaining variation not described by the first component. The third principle (PCEF3) accounted for 9.32% of the variance not explained by all preceding components; the fourth principal (PCDF4) accounted for 8.55%; the fifth principal (PCEF5) explained 7.81%, and the sixth principal (PCEF6) accounted for 5.81%.

3.4. Factor Loadings before and after Rotation. As illustrated in Tables 4, 5, and 6 that factors' loadings before rotation show no complex structures on design-related and time- and cost-related factors, whereas eight factors in the parties' experience-related factors encountered complex structures before rotation and are not found in the rotated factors. Based on previous studies, the study focused on minimizing the number of factors on which the determinants have high loading. Varimax rotation is applied to the extracted factors at each principal component.

Tables 7, 8, and 9, respectively, show the findings of the design, time/cost, and experience-related aspects of the rotation matrix. A varimax rotation is also performed on the factors to generate factor loadings that are easier to read. Rotation describes the behavior of variables under severe conditions by maximizing the loading of each variable on one of the principal components while minimizing the loading on all other factors, and it is the best factor output solution for interpreting factor analysis.

Tables 7, 8, and 9 show the factors with complex structures resolved on rotated loadings with values greater than 0.5. This confirms that rotated loadings have meaningful interpretation, as can be seen from the design-related and time- or cost-related rotated factor analysis. This iteration for factor extraction solved the complex structure after rotation, except for one factor. Furthermore, for experience-related factors, the number of original factors is 17 and the number of principal components is 6. This factor extraction iteration

Table 3: Total variance explained for parties experience-related factors.

Factor ID	Initial eigenvalues			Extraction sum of squared loadings		
	Total	Percentage of variance	Cumulative	Total	Percentage of variance	Cumulative
PER1	4.95	27.54	27.54	4.95	27.54	27.54
PER2	2.23	12.46	40.00	2.23	12.46	40.00
PER3	1.54	9.32	49.32	1.54	9.32	49.32
PER4	1.22	8.55	57.87	1.22	8.55	57.87
PER5	1.00	7.81	65.68	1.00	7.81	65.68
PER6	0.97	5.83	71.51	0.97	5.83	71.51
PER7	0.88	5.18	76.69	—	—	—
PER8	0.74	4.81	81.50	—	—	—
PER9	0.71	4.33	85.83	—	—	—
PER10	0.60	3.05	88.87	—	—	—
PER11	0.17	2.50	91.37	—	—	—
PER12	0.19	2.22	93.59	—	—	—
PER13	0.22	1.94	95.53	—	—	—
PER14	0.49	1.75	97.28	—	—	—
PER15	0.40	1.21	98.49	—	—	—
PER16	0.36	0.86	99.35	—	—	—
PER17	0.33	0.65	100.00	—	—	—

Table 4: Design-related factors before rotation.

ID	Loadings before rotation				h^2
	PCDF1	PCDF2	PCDF3	PCDF4	
DER1	0.61	0.09	0.47	0.13	0.62
DER2	0.60	0.15	0.45	0.03	0.58
DER3	0.58	0.34	−0.13	0.29	0.55
DER4	0.59	−0.05	0.17	0.28	0.45
DER5	0.63	0.31	0.01	−0.08	0.50
DER6	0.37	−0.60	0.15	0.31	0.62
DER7	0.80	0.11	0.00	0.02	0.66
DER8	0.44	−0.73	0.08	−0.31	0.82
DER9	0.74	−0.22	0.04	−0.19	0.63
DER10	0.69	−0.01	−0.25	−0.39	0.70
DER11	0.64	0.13	−0.29	−0.40	0.67
DER12	0.61	0.26	−0.26	0.21	0.55
DER13	0.36	−0.36	−0.55	0.46	0.77

h, sum of squared loadings for the factors.

Table 5: Time- and cost-related factors before rotation.

Factor ID	Loadings before rotation		h^2
	PCTCF1	PCTCF2	
TCR1	0.80	0.07	0.64
TCR2	0.71	0.24	0.50
TCR3	0.70	0.10	0.48
TCR4	0.65	−0.00	0.42
TCR5	0.79	0.10	0.62
TCR6	0.55	−0.46	0.30
TCR7	0.15	−0.83	0.02
TCR8	0.00	−0.83	0.00

h, sum of squared loadings for the factors.

Table 6: Experience-related factors before rotation.

Factor ID	Loadings before rotation						h^2
	PCEF1	PCEF2	PCEF3	PCEF4	PCEF5	PCEF6	
PER1	−0.49	−0.00	0.55	0.07	0.21	−0.51	0.86
PER2	−0.47	−0.30	0.50	−0.13	0.05	−0.12	0.59
PER3	−0.57	−0.32	0.37	0.30	0.05	0.13	0.66
PER4	−0.30	−0.18	0.55	−0.18	−0.08	0.62	0.84
PER5	−0.59	−0.29	−0.01	0.23	0.06	−0.02	0.49
PER6	−0.57	−0.39	−0.07	−0.25	0.18	0.14	0.60
PER7	−0.25	−0.37	−0.18	−0.72	−0.01	−0.08	0.75
PER8	−0.71	−0.28	−0.20	−0.05	−0.20	−0.09	0.69
PER9	−0.62	−0.04	−0.04	−0.16	−0.21	−0.35	0.58
PER10	−0.55	0.25	−0.08	−0.22	−0.56	0.09	0.75
PER11	−0.39	0.66	0.19	0.07	−0.41	−0.05	0.80
PER12	−0.60	0.52	0.01	−0.09	0.31	0.01	0.74
PER13	−0.50	0.61	−0.01	−0.19	0.31	0.29	0.85
PER14	−0.69	0.53	−0.13	0.12	0.15	−0.03	0.82
PER15	−0.63	−0.15	−0.32	0.20	−0.08	0.03	0.57
PER16	−0.57	−0.31	−0.26	0.50	−0.10	0.19	0.79
PER17	−0.42	−0.09	−0.48	−0.06	0.32	−0.00	0.53

h, represents the proportion of total variance explained by each principal component in the principal component analysis (PCA).

solved the complex structure of eight factors. From this, it is concluded that the number of complex structures among the variables increased with an increase in the number of original variables and an associated increase in the number of principal components. In this finding, factor rotation remains useful for the removal of complex structures from original variables; hence, it is similar to the finding of [16]. Finally, factors with the highest loading factor under each principal component are selected from the rotated loadings for predicting the cost model.

Table 7: Rotated loadings of design-related factors.

ID	Rotated loadings			
	PCDF1	PCDF2	PCDF3	PCDF4
DESR1	0.13	−1.56E-01	0.76	−0.01
DESR2	0.20	−0.11	0.72	−0.08
DESR3	0.30	0.21	0.47	0.44
DESR4	0.12	−0.21	0.56	0.29
DESR5	0.50	0.10	0.48	0.11
DESR6	−0.13	−0.66	0.26	0.31
DESR7	0.51	−0.12	0.55	0.27
DESR8	0.31	−0.85	0.03	−0.04
DESR9	0.54	−0.44	0.38	0.10
DESR10	0.79	−0.20	0.16	0.10
DESR11	0.80	−0.05	0.14	0.09
DESR12	0.41	0.14	0.36	0.49
DESR13	0.10	−0.29	−0.08	0.82

Table 8: Rotated loadings time-/cost-related factors.

ID	Rotated loadings	
	PCTCF1	PCTCF2
TCR1	0.797	−0.057
TCR2	0.736	0.130
TCR3	0.704	−0.006
TCR4	0.640	−0.106
TCR5	0.795	−0.026
TCR6	0.468	−0.544
TCR7	0.015	−0.839
TCR8	−0.131	−0.824

Table 9: Rotated experience-related factor loadings.

Factor ID	Rotated loadings					
	PCEF1	PCEF2	PCEF3	PCEF4	PCEF5	PCEF6
PER1	−0.07	0.19	0.90	0.03	−0.05	−0.02
PER2	−0.15	−0.01	0.63	−0.22	−0.05	0.34
PER3	−0.51	0.04	0.46	0.08	0.02	0.42
PER4	−0.05	0.06	0.11	−0.08	−0.08	0.90
PER5	−0.61	0.09	0.30	−0.09	−0.01	0.10
PER6	−0.43	0.15	0.17	−0.54	0.08	0.26
PER7	0.00	−0.05	0.04	−0.86	−0.08	0.04
PER8	−0.62	0.06	0.19	−0.39	−0.32	0.01
PER9	−0.31	0.13	0.38	−0.33	−0.45	−0.14
PER10	−0.20	0.20	−0.05	−0.19	−0.78	0.13
PER11	0.04	0.40	0.14	0.31	−0.72	0.03
PER12	−0.11	0.81	0.20	−0.05	−0.16	0.00
PER13	0.01	0.89	−0.04	−0.06	−0.14	0.18
PER14	−0.32	0.77	0.14	0.08	−0.29	−0.11
PER15	−6.97E−01	0.18	0.03	−0.14	−0.18	−0.05
PER16	−0.88	0.01	−0.02	0.09	−0.06	0.09
PER17	−0.44	0.36	−0.09	−0.36	0.16	−0.21

3.5. Selection of Factors/Variables for Predicting Cost Model. According to the PCA analysis, the total of 38 factors from the original model was reduced to 19 factors, out of which eight design-related factors are found to be important and explain 62% of the total variation in the data; hence, the eight coefficients were selected for the cost estimation model. Similarly, time- and cost-related two factors are found important and explain 67% of the total variation. The parties' experience-related factors selected are nine factors that are found to be important and explain 70% of the total variation. These 19 factors shall also be passed for further analysis for the principal component selection for the final model, and in this analysis, the first 19 factors together explain 66% of the variation; hence, it is decided that that's good enough. Then, for later analysis, it would only keep those 19 factors. However, the selected factors for the model shall have commonalities above the cumulative variances; this further reduces the 19 factors in the next section. The principal component analysis result selected that 19 variables are important for cost estimation from the three groups of variables. Among these, the highest loading value at each principal component selects 12 variables from the three groups of factors.

3.6. Develop Predictive Cost Models

3.6.1. Selecting Principal Components for Regression. This study criterion for selecting principal components for a regression model follows similarly to these previous studies, selecting the principal component whose eigenvalues and percentage variance are greater than the average eigenvalues and the percentage cumulative variance of the factor, respectively. This analysis further reduced the 19 factors to 14 that are included for regression analysis in the final model. Among the 14 variables obtained by PCA, six factors were identified by RII as top most rated factors, which include design completion (by owner) when bids are invited, project scope definition completion when bids are invited, level of design complexity, level of construction complexity, importance for the project to be completed within budget, and adequacy of contractor plant and equipment. Accordingly, the cumulative percentage variances of design, time/cost, and parties experience-related factors are 62%, 58%, and 70%, respectively, whereas the average eigenvalues and specific variances of the components are 2.03, 2.33, and 1.98, respectively. The extraction of principal components that represent the highest variation in the data was completed in the previous section, which sorted 19 factors for further analysis to include in the final model. In this section, the most significant principal component shall be selected to be used in the model estimation.

Using the PCA factor analysis statistical package, among the 13 original design-related factors four principal components are selected so that the cumulative variance explains 62% of the variation. Using the [3] selection criterion based on the significance of the contribution of the principal component and compared with the average eigenvalue (2.03), the

Table 10: Categorized variables of the principal components.

Factor ID	Given name	PCDF1	PCDF3	PCTCF1	PCEF1	PCEF3
DSER11	x1	0.80	0.14	—	—	—
DESR10	x2	0.79	0.16	—	—	—
DESR9	x3	0.54	0.38	—	—	—
DESR1	x4	0.13	0.76	—	—	—
DESR2	x5	0.20	0.72	—	—	—
DESR7	x6	0.51	0.55	—	—	—
TCR1	x7	—	—	0.80	—	—
TCR5	x8	—	—	0.80	—	—
PER16	x9	—	—	—	−0.88	0.01
PER8	x10	—	—	—	−0.62	0.06
PER4	x11	—	—	—	−0.05	0.06
PER13	x12	—	—	—	0.01	0.89
PER12	x13	—	—	—	−0.11	0.81
PER14	x14	—	—	—	−0.32	0.77

two principal components (PCDF1 and PCDF3) are selected for the estimation of the regression model. The total percentage variance for the two principal components is 69%. Likewise, for the time- and cost-related factors, there are eight factors and one principal component extracted that explained 58% of the variation. The eigenvalue for the component is 2.33; hence, the component (PCTCF1) selected t explains 67% of the total variance. Concerning the parties' experience related to the 17 original factors, six principal components were selected that explained 70% of the total variance. The average eigenvalue of these factors is 1.98; likewise, two principal components (PCEF1 and PCEF3) that explain 80% of the total variance are to be included in the regression model.

3.6.2. Grouping of Input Variables. The basis for selecting significant principal components for regression grouped into 14 variables out of the 19 originals. The components that were grouped into five variables have new headers for clarity. Design factors include PCDF1 (project scope definition) and PCDF3 (project complexity), time/cost factors include PCTCF1 (project cost and time performance), parties' experience factors include PCFE1 (parties' experience) and PCFE3 (parties' commitment to time, cost, and quality), and design factors include PCTCF1 (project cost and time performance).

3.6.3. Independent Variables. The use of principal component analysis has reduced the original 38 factors to 19 variables. These were further reduced to 14, which are also extracted into five new variables and chosen to be the independent variables of this study, as shown in Table 10. The sample statistics of all the independent variables are presented in Table 11.

3.6.4. Regression Model Estimation for Cost Prediction. The initial and final project costs of 56 construction projects are available from the data collection. Based on the cost of variation specified in the General Conditions of Contracts of Ethiopia, Public Procurement Agency (PPA), 2011, as well as the adequacy of interpretation and creation of a cost prediction model, these projects were grouped into seven categories. The classification based on the final cost amount resulted in four models, which are as follows:

(i) Model 1: project final cost from Birr 300 million–3 billion.
(ii) Model 2: project final cost from Birr 120 million up to 280 million.
(iii) Model 3: project final cost from Birr 50 million up to 120 million.
(iv) Model 4: project final cost from Birr 1 million up to 50 million.

Accordingly, the regression model for the cost variation is classified into three models:

(i) Model 5: the projects with more than 25% of cost variation.
(ii) Model 6: the projects with up to 25% cost variation.
(iii) Model 7: the projects with 0%−30% cost variation.

The seven models were created in MATLAB using the Create a stepwise linear regression model of project cost as the dependent variable and the XCL matrix as the independent variables (14 reduced sets of cost influencing factors). The results revealed that the six models created for different project scopes in terms of project amount and cost of variation were generated as fitted linear models, whereas the model created based on the projects list showed variation up to 28% and was found not to be fitted linear models. Although the results of the six models were fitted to the linear model, it is necessary to validate the models based on the provided criteria, which will be utilized to identify the predictive models by testing and validating the models.

Model 1: According to R^2 and modified R^2, the eight factors account for 98% and 94.7% of the total variation of the project cost, respectively. At the 1% significance level, the F-ratio test suggests that the cumulative impact of the seven factors is very significant. The regression model may be quantitatively defined as given in Equation (2):

$$\begin{aligned}\text{Project cost}(Y) = {} & 209.25 + 4{,}914.7\text{x}1 + 3{,}071.7\text{x}2 \\ & - 1{,}101.3\text{x}4 - 826.81\text{x}5 + 1{,}055.5\text{x}7 \\ & + 3{,}105\text{x}10 - 1{,}009.8\text{x}13 + 1{,}411.4\text{x}14.\end{aligned} \tag{2}$$

Number of observations: 14, error degrees of freedom: 5; root mean squared error: 196; R-squared: 0.98, adjusted R-squared: 0.947; F-statistic vs. constant model: 30, p-value $= 0.000825$.

Model 2: As indicated by the original R^2 and adjusted R^2, these variables account for 97.6% and 92.2% of the total variance of the project cost, respectively. At the 1% significance level, the F-ratio test suggests that the combined effect of the seven factors is very significant. The regression model may be quantitatively defined as stated in Equation (3):

Table 11: Cost prediction variables sample statistics.

ID	Mean	SD	Variance	Median	Skewness	Kurtosis	Jarqua–Bera	Probability
DER11	4.02	0.98	0.95	4.00	−0.95	0.69	22.77	0.00
DER10	3.89	0.99	0.98	4.00	−0.77	0.26	13.59	0.00
DER9	3.81	1.00	1.01	4.00	−0.63	−0.05	8.82	0.01
DER1	4.12	0.84	0.70	4.00	−0.71	−0.09	11.12	0.00
DER2	4.03	0.90	0.81	4.00	2.06	15.41	1,409.38	0.00
DER7	3.92	1.04	1.09	4.00	−1.38	4.09	135.07	0.00
TCR1	3.66	1.03	1.07	4.00	−0.50	−0.24	5.84	0.05
TCR5	3.62	1.09	1.20	4.00	−0.51	−0.32	6.32	0.04
PER16	4.13	0.89	0.79	4.00	−1.24	2.05	57.58	0.00
PER8	3.96	0.95	0.90	4.00	−0.68	−0.15	10.40	0.01
PER4	3.74	0.87	0.75	3.00	−0.13	0.07	0.42	0.81
PER13	3.87	1.06	1.12	4.00	−0.56	−0.74	10.04	0.01
PER12	3.85	1.11	1.23	4.00	−0.55	−0.79	10.10	0.01
PER14	3.80	1.11	1.24	4.00	−0.56	−0.66	9.33	0.01

Table 12: VIF values from MATLAB.

ID	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
VIF	1.59	1.55	1.58	1.62	1.44	1.78	1.66	1.23	1.64	1.49	1.79	1.99	1.24	1.48

$$\begin{aligned}\text{Project cost}(Y) = {} & 147.8 + 173.08\text{x1} + 148.4\text{x2} \\ & + 161.56\text{x3} + 118.38\text{x5} + 144.58\text{x7} \\ & + 78.51\text{x8} + 52.13\text{x9} + 262.75\text{x10} \\ & - 143.56\text{x13}.\end{aligned} \tag{3}$$

Number of observations: 14, error degrees of freedom: 4; root mean squared error: 15; R-squared: 0.976, adjusted R-squared: 0.922; F-statistic vs. constant model: 18, p-value = 0.00678.

Model 3: At the 1% significance level, the F-ratio test suggests that the combined effect of the seven factors is very significant. The regression model may be quantitatively defined as stated in Equation (4):

$$\begin{aligned}\text{Project cost}(Y) = {} & 55.69 + 50.13\text{x1} + 58.82\text{x2} + 141.59\text{x3} \\ & + 33.11\text{x4} + 50.71\text{x5} + 64.94\text{x7} \\ & + 59.7\text{x10} - 58.62\text{x13}.\end{aligned} \tag{4}$$

Number of observations: 14, error degrees of freedom: 6; root mean squared error: 8.05; R-squared: 0.962, adjusted R-squared: 0.917; F-statistic vs. constant model: 21.5, p-value = 0.000754.

Model 4: The f-ratio test suggests that the combined effect of the seven factors is highly significant at the 3% significance level. The regression model may be quantitatively defined as stated in Equation (5):

$$\begin{aligned}\text{Project cost}(Y) = {} & -1.3 + 52.78\text{x1} + 15.55\text{x2} + 66.15\text{x3} \\ & + 48.93\text{x5} + 51.911\text{x6} + 50.34\text{x7} + 27.91\text{x9} \\ & + 64.47\text{x10} + 75.52\text{x11} + 44.21\text{x12}.\end{aligned} \tag{5}$$

Number of observations: 14, error degrees of freedom: 3; root mean squared error: 4.31; R-squared: 0.979, adjusted R-squared: 0.908; F-statistic vs. constant model: 13.8, p-value = 0.0266.

3.6.5. Collinearity Diagnosis. The VIF values revealed that the predictor variables in the dataset have a multicollinearity issue. Table 12 demonstrates that the VIF values for all values are less than 2. This demonstrates that the variables did not experience the multicollinearity issue. VIFs greater than 10 are typically used to indicate considerable collinearity [5].

3.6.6. Overfitting Diagnostics. Table 13 illustrates the model estimation initial iteration result, which shows the values overfitted with corresponding F-statistics, showing a misleading p-value that is smaller than the 5% level of significance. Stepwise regression cross-validation indicated that the overfitting problem at the lowest AIC was solved at the 1% level of significance, hence it is concluded that there is no need to proceed with the AIC PCR model (Table 14).

3.6.7. Model Selection Comparison. The diagnoses of overfitting and collinearity showed that this dataset has neither overfitting nor collinearity problems; hence, it is concluded that the AIC–PCR solved the problem in this case. Therefore, the final predicted model is selected based on the AIC and SSE model selection criteria for comparison. Among the seven models, Model 1 is used to carry out the model comparison; accordingly, two models, Model A and Model B are generated based on stepwise regression analysis in the MATLAB software. Model A is generated for the lowest SSE, and Model B is generated for the lowest AIC values. The root mean square value is used to compare the models [5] used mean

Table 13: Four models iteration and corresponding AIC values and iterations.

Model name	Lowest AIC	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Iteration 6	Iteration 7	Iteration 8	Iteration 9	Iteration 10
Model 1	191.15	214.63	213.3	212.2	212.69	209.74	206	203.01	204.69	206.58	202.26
Model 2	117.97	146.54	147.67	149.2	151.18	153.16	154.79	155.52	157.44	150.62	141.05
Model 3	99.4	122.88	121.09	122.52	124.31	125.55	127.43	125.55	120	121.27	122.16
Model 4	81.05	98.14	99.55	99.07	100.88	102.86	102.83	100.28	100.97	102.17	95.58

Table 14: Four models iteration and corresponding AIC values iterations 11–21.

Model name	Iteration 11	Iteration 12	Iteration 13	Iteration 14	Iteration 15	Iteration 16	Iteration 17	Iteration 18	Iteration 19	Iteration 20	Iteration 21
Model 1	204.01	205.37	inf	inf	inf	194.44	193.85	191.85	192.59	191.15	193.56
Model 2	142.95	136.28	inf	inf	inf	119.94	120.3	142.39	140.51	138.75	117.97
Model 3	121.27	117.38	118.67	117.38	100.8	102.63	100.8	99.4	102.27	—	—
Model 4	97.06	82.99	81.25	81.05	82.54	81.05	85.22	85	92.03	91.11	89.41

Table 15: The approximate models R^2 values.

Models	R^2	Adj R^2	Explained (%)	Model R^2 (%)	Model adj R^2 (%)
Model 1	0.98	0.947	55	56	59
Model 2	0.976	0.922	55	56	61
Model 3	0.962	0.917	55	57	62
Model 4	0.979	0.908	55	56	62

squared error (MSE) to select the best models. Model A has a RMSE of 371, whereas Model B has a RMSE of 196. The lowest RMSE value indicates the lowest AIC value of a model. The findings of the model under SSE demonstrated that the 12 factors deemed essential for cost estimation account for 98.5% and 81% of the entire variation of the project cost, respectively, as evidenced by the original R^2 and amended R^2. The F-ratio test shows that the total impact of the seven variables has a level of significance of 31% at the 5% significance level. The model under the AIC criterion revealed that the eight factors determined to be relevant for the cost estimate account for 98% and 94.7% of the total variation of the project cost, as evidenced by the original R^2 and modified R^2, respectively. At the 0.00% significance level, the F-ratio test suggests that the combined effect of the seven factors is very significant. As a result, it is determined that the AIC selection criteria, when compared to the SSE, is the best-fit model for cost estimation of public building construction projects at the 1% level of significance, given that both approaches employed PCA for factor analysis.

3.6.8. Testing and Validating Models. The prediction model findings revealed that the impact of the number of factors on the cost of the project varied with the volume of the final project. The results also showed that the predicted model best explained the actual project cost rather than the initial cost. This was confirmed by Models 5–7 which were developed taking into account high-cost and low-cost variation of projects, but the models were rejected because of the highest percentage error estimation report when compared to previous studies. To test the validity of each of the six models, the predicted values of the project cost were computed using the mean average percentage error of estimation (MAPE) and found to be 12.8%, 2.6%, 4.4%, 14.26%, 850%, and 282%, respectively. In their investigation, Chan and Park [3] discovered that the average percentage error of the estimate was roughly 13%. This analysis projected better error for four models and two models (Models 5 and 7) from the selected projects based on cost variance indicated the largest error of estimation when compared to previous studies by Chan and Park [3] and Xiong et al. [5]. To test and validate the models, mean absolute percentage error (MAPE) is employed to determine the predictive ability of the models. Table 15 shows that the five principal components chosen for the prediction model explain: 45%, 47%, 68%, 47%, and 68% of the total variation. The average variation explained by the five principal components is 55%:

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}. \tag{6}$$

The R^2 and adjusted R^2 values for Model 1 are approximately 56% and 59%, respectively, for Model 2, 56% and 61%, for Model 3, 57% and 62%, and for Model 4, 56% and 62%. Model 6 was rejected because it did not fit in linear regression, whereas Models 5 and 7 were rejected due to the

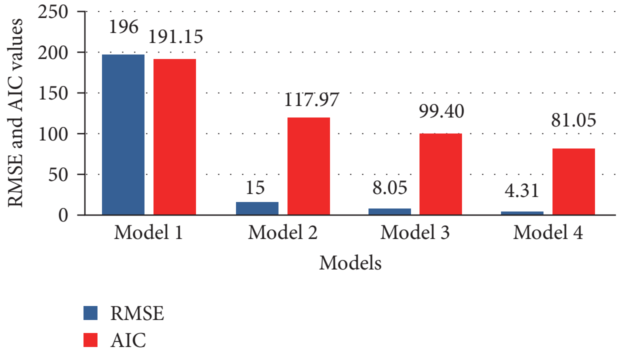

Figure 4: The AIC and root mean squared error of the four models.

highest percentage of error reported. These four results contrasted favorably with previous research on cost estimation and prediction models, as indicated by published (as reviewed by Ganiyu and Zubairu [10]) values of R^2 of 20.8% [17], and 41% [3] and 20% [10], and 58.6% using neural networks [18]. As a result, the findings of this study are consistent with those of Emsley et al.'s [18] study. The study's results were better than those of [3, 5] estimates of 13% and 41.9%, respectively, according to the MAPE, which showed that the three models were 12.8%, 2.6%, and 4.4%, respectively.

Figure 4 illustrates that, even though Model 1 has a very high RMSE value, the other three models demonstrated a smooth increase. Model 4 has the lowest RMSE and AIC values, as well as the best-adjusted R^2, accounting for 62% of variance with 10 factors set among the 38 original sets of components, but its MAPE result reveals that its predictive power is poorer when compared to the other models. As a consequence of the breadth of this study, the results showed that the most predictive model could not be chosen among the four models since it required further extensive investigation. Consequently, the most essential factors that happened in all four models were found to be very important for the cost estimation of public buildings in Addis Ababa. Among the 14 factors identified by PCA, five were represented in the four models: design completion (by owner) when bids are invited, project scope definition completion when bids are invited, level of construction complexity, importance for the project to be completed within budget, and subcontractor experience and capability. The validity of the study's findings is demonstrated by the fact that all of these important factors—design (three factors), time/cost (one factor), and project party experience (one factor)—were among the top 5, except for subcontractor experience and capability, which was rated among the top 6 but included nonetheless because it was one of the factors chosen by the regression model.

4. Conclusion and Recommendation

This study intended to create a cost prediction model for public building projects in Addis Ababa utilizing PCA–AIC criteria and stepwise multiple linear regression models created in MATLAB software. The results of the study led to the following conclusions: The results showed that five out of 14 factors were recognized by the cost prediction model fitted to the four models, and those factors represented by all the models were chosen as being highly essential. These factors include design completion by the owner when bids are invited; completion of the project scope definition when bids are invited, level of construction complexity, importance of project completion within budget, contractor and subcontractor experience, and capability have all been identified as the main cost-determining factors. The found variables were among the top 6 rated in the study and were relevant for the cost estimation of public building construction projects in Addis Ababa at the 5% level of significance. The study discovered four cost-estimating models in the study region for projecting the building cost of projects for different project cost categories. In comparison to prior research, the results of four prediction models showed improved R^2 and adjusted R^2 values in the final model, which identified 14 factors that explained 55% of the total variation. The study also concluded that the AIC selection criteria are the best-fit model with the lowest RMSE value for cost estimation of public building construction projects at the 1% level of significance, as opposed to the SSE, which had a higher RMSE value as long as both methods used PCA for factor analysis. The study concluded that the estimated model enhanced the study's outcomes. This is supported by higher R^2 values observed in comparison to earlier investigations. In addition, the average percentage error of estimation has been reduced by the three models to 12.8%, 2.6%, and 4.4%, as compared to 13% indicated by prior research. The study was able to construct a predictive cost model utilizing the 14 factors that have a substantial influence on project cost, which represented 55% of the model. Although the PCA has reduced the enormous number of components to a modest and crucial number in comparison to earlier research, the study's number of factors with distinct cost categories has shown a varied set of criteria typically considered crucial to cost estimation. This study introduces a novel method (PCR–AIC) for identifying the most important factors for estimating the cost of public building construction projects in Addis Ababa. By applying this method, the study eliminates the redundant factors and retains the essential ones that capture the majority of the original variables' attributes. This study's contribution is relevant for future similar projects in this location, as they can use the identified factors to estimate the cost at an early stage more accurately and efficiently. However, to arrive at a highly trustworthy prediction model, future research should conduct a more extensive analysis based on the procurement method, project complexity, and location. Future research could also investigate the model's applicability utilizing the nonlinear technique with other estimating components, which would be a significant contribution to this work.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] N. Tadesse and A. Dinku, “Conceptual cost estimation of road projects in Ethiopia using neural networks,” 2017.

[2] A. Puckett, “Development of a parametric cost estimating model for university of alaska system renovation construction projects item type report,” 2016, http://hdl.handle.net/11122/7793.

[3] S. L. Chan and M. Park, “Project cost estimation using principal component regression,” *Construction Management and Economics*, vol. 23, no. 3, pp. 295–304, 2005.

[4] A A CE International, *Skills & Knowledge of Cost Engineering: A Product of the Education Board of AACE International Paperback*, CreateSpace Independent Publishing, 5th edition, 2011.

[5] B. Xiong, S. Newton, V. Li, M. Skitmore, and B. Xia, “Hybrid approach to reducing estimating overfitting and collinearity,” *Engineering, Construction and Architectural Management*, vol. 26, no. 10, pp. 2170–2185, 2019.

[6] R. M. O’Brien, “A caution regarding rules of thumb for variance inflation factors,” *Quality & Quantity*, vol. 41, no. 5, pp. 673–690, 2007.

[7] C. M. Andersen and R. Bro, “Variable selection in regression —a tutorial,” *Journal of Chemometrics*, vol. 24, no. 11-12, pp. 728–737, 2010.

[8] R. X. Liu, J. Kuang, Q. Gong, and X. L. Hou, “Principal component regression analysis with spss,” *Computer Methods and Programs in Biomedicine*, vol. 71, no. 2, pp. 141–147, 2003.

[9] S. T. Hashemi, O. M. Ebadati, and H. Kaur, “Cost estimation and prediction in construction projects: a systematic review on machine learning techniques,” *SN Applied Sciences*, vol. 2, no. 10, 2020.

[10] B. O. Ganiyu and I. Zubairu, “Project cost prediction model using principal component regression for public building projects in Nigeria,” 2010, http://www.journalbp.co.cc.

[11] D. Birhanu, “Addis ababa institute of technology school of civil and environmental engineering model for estimating construction duration of public building projects in Addis: construction technology and management (sponsor: Ethiopian roads authority),” *C*, 2020.

[12] J. E. Bartlett, J. W. Kotrlik, and C. C. Higgins, “Organizational research: determining organizational research: determining appropriate sample size in survey research appropriate sample size in survey research,” 2001.

[13] A. Field, *Discovering statistics using SPSS*, Sage Publications, Inc, Thousand Oaks, CA, US, 2nd edition, 2005.

[14] P. F. Kaming, P. O. Olomolaiye, G. D. Holt, and F. C. Harris, “Factors influencing construction time and cost overruns on high-rise projects in Indonesia,” *Construction Management and Economics*, vol. 15, no. 1, pp. 83–94, 2010.

[15] D. Posada, T. R. Buckley, and J. Thorne, “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, 2004.

[16] R. Dobgegah, D.-G. Owusu-Manu, and K. Omoteso, “A principal component analysis of project management construction industry competencies for the Ghanaian,” *Construction Economics and Building*, vol. 11, no. 1, pp. 26–40, 2011.

[17] M. Skitmore, S. Stradling, A. Tuohy, and H. Mkwezalamba, *The Accuracy of Construction Price Forecasts*, The University of Salford, 1990.

[18] M. W. Emsley, D. J. Lowe, A. R. Duff, A. Harding, and A. Hickson, “Data modelling and the application of a neural network approach to the prediction of total construction costs,” *Construction Management and Economics*, vol. 20, no. 6, pp. 465–472, 2002.