4th Construction Specialty Conference
4e Conférence spécialisée sur la construction

Montréal, Québec
May 29 to June 1, 2013 / *29 mai au 1 juin 2013*

Development of Initial Costs Forecasting Models for School Buildings Using Multiple Linear Regression

O.S. Alshamrani[1], S. Alkass[2], and K. Galal[3]
[1] Department of Building Engineering, College of Architecture and Planning, University of Dammam, Dammam, Kingdom of Saudi Arabia
[2&3] Department of Building, Civil and Environmental Engineering, Concordia University, Montréal, Québec, Canada

Abstract: The recent economic crises affected the whole world economic which enforced some countries to declare their bankruptcy such as Argentina and Island. Government of Canada spends yearly billions of dollars to construct and run school buildings which represent the major domain and largest footprint of public sector. Since that the initial costs proved to have the highest impact of the life cycle costs in school buildings in Canada, and since that these initial costs are affected by structure and envelope types of school, Linear regression models are developed to enable school boards to predict the initial costs of new conventional school buildings associated with applying various structure and envelope types. Seven multi-regression models are developed in this paper to predict the square footage initial costs for different structure and envelope types, covering steel, concrete, and wood structures, in various combinations for new conventional school buildings. The RS Means is used in this study to predict the construction costs and to create 420 data points to be used in the models. Each model is developed to predict the specified structure and envelope type with regards to correlated predictor variables that include: school area (square foot), number of floors (which ranges from1-4), and school level, which ranges from 1-3: elementary school (1), middle school (2), and high school (3). The development of the models is performed throughout three major stages; preliminary diagnostics on data quality, the models development process, and models validation. The developed models will help the school boards to predict the initial costs and to select the economical alternative.

1. Introduction

Most school buildings in the United States and Canada were built during three general time periods: the 1920s and 30s, the 1950s and 60s, and the 1980s to the present. In the 1930s, schools were built as a part of government work projects. After World War II, the first baby boom required the number of school buildings to expand from the late 1950s until the early 1970s. School buildings again needed to be built from the late 1980s to the present, due to the second baby boom as baby boomers had children of their own, and since many school buildings had been changed to other uses (Maciha, 2000).

Government of Canada spends billions of dollars annually to construct new school buildings and run the existing ones. For instance, in Ontario alone, In January 2013, the Ministry of Education announced about $700 million to support school boards' capital priority school projects needed in the next three years. These projects include building new schools to address enrolment growth, to support schools, to replace schools in poor condition and to support school consolidations (OMOE, 2013).

These Initial Costs contain the total ownership costs related to the initial development of a project. Some of these costs include construction costs, fee costs, and other costs such as real estate, site, and professional services (Dell'Isola 2003).

The literatures show that the initial costs of school buildings represent the major single cost of the life cycle costs. It represents about 43% -54.7% of the whole Life Cycle Costs (LCC) for elementary and high school buildings, respectively (Flangan, 1983). Since the initial costs represent the highest cost and since that government of Canada spends lots of money in construction of new school buildings, this paper develop initial costs forecasting models to assist schools boards in their estimation and selection of new school buildings based on their structure and envelope types.

2. Methodology

The RS Means is used in this study to predict the construction costs by identifying some significance parameters. Several input parameters are defined to calculate the initial costs, including school level, school area, floor height, number of floors, structure type, envelope type, city, and year of construction, The description possibilities of each parameter include;

- School level: Elementary, middle, and high school
- School area: 45000, 75000, 125000, 175000, and 250000ft²
- Number of floors: 1, 2, 3, and 4 floors
- Floor height: 13.1 ft
- Structure type: Steel frame, wood frame, and concrete frame
- Envelope type: steel studs, wood studs, concrete brick, masonry wall, and precast concrete panels.
- City: Montreal, Canada
- Year of construction: 2011

After identifying the parameters, the model is applied to estimate the construction costs for a new school building. The output of the RS Means is presented in a detailed table that has a breakdown of the component cost used to develop the base building cost. The breakdown cost components include:
Substructure: foundations, slab on grade, basement excavation and walls.
Super structure (Shell): floor construction, roof construction, exterior walls, windows, doors, roof coverings and roof openings.
Services: elevators and lifts, plumbing fixtures, domestic water distribution, rain water drainage, energy supply, cooling systems, sprinklers, standpipes, electrical services/distribution, lighting and branch wiring, communications and security, and other electrical systems.
Interiors: partitions, interior doors, fittings, stair construction, wall finishes, floor finishes, ceiling finishes.
Equipment and furnishings: institutional equipment, HVAC, and other equipment.
Contractor fees: general conditions, overhead, contingency, and profits.
Architecture fee: design, drawing, and supervision.

After estimating the breakdown component costs, they are added to calculate the subtotal cost which is then added to the contractor’s and architecture fees. The total building cost is then estimated a square foot cost.

Computing of the initial costs in this study is performed by applying different scenarios to build a correlation between the input parameters and the total square foot base cost. Each structure and envelope type is estimated for different school levels at specific area sizes and number of floors, resulting in 21 tested scenarios. Each area size is applied on a different number of floors resulting in 20 various tested scenarios as. Four hundred and twenty (420) construction cost estimating scenarios result from the combination of the complete range of input parameters for new school buildings in Montreal. Seven cost scenarios are grouped together and evaluated according to the structure and envelope types of a new school building in order to select the most economically viable alternative.

2.1 Data Preparation for Modeling

The input parameters (independent variables) for the initial costs are gathered from the RS Means. Some of these variables are normalized, such as the city, year of construction, and floor height. The other parameters, such as structure and exposure type, school level, number of floor, and school area are variables and have a significant effect on the initial costs. These factors are investigated in this paper to develop their correlation to the resulted initial costs (dependent factor). The computed initial costs from RS Means include about 420 data points. Eighty percent of this data, (336 points) are used to build the initial cost prediction models for conventional school buildings. Twenty percent of the data (84 points) are randomly picked and excluded from the analysis to be used for model validation. The data is sorted based on structure and envelope type in order to be used in developing the prediction model.

2.2 Model Development Process

The main aim of the model development is to find correlations between the predictors and the response variables. The multiple linear regression technique was utilized to address the correlation and to develop prediction models for each structure and exposure type. Regression model development methodology consists of three major stages; preliminary diagnostics on data quality, the model development process, and model validation, as shown in Figure 1. The preliminary data checks include two steps: determining any possible relationship and interaction of data, and performing the best subset regression analysis. The next stage is the model development process which has four major steps: building the regression model, testing basic factors, performing residual analysis, and selecting the model for validation. The final stage in the model development process is performing the validation. Each step in the various development process stages can be illustrated as follows:

2.3 Preliminary Data Diagnostics

2.3.1 Addressing Correlations and Interactions

The first step in the preliminary checks on data is to detect and address any existing multi-colinearity or possible interactions in the predictor variables of the developed models. The matrix scatter plot is simulated for all predictor variables vs. the response factor to detect the correlation. Scatter plot representation is significant in detecting the linearity of data or any other correlation between predictors.

2.3.2 Best Subset Analysis

The next step in the preliminary diagnosing of data is to perform best subsets’ regression analysis. This test identifies the best possible combination of predictors with regards to the highest R^2 and R^2 (adjusted) values and the lowest error and variation values. Hence, the best-fit regression models that can be developed with the specified number of variables are determined using best subset regression analysis.

3 Regression Models Development

After detecting the correlation and identifying the best data subset, seven regression models are developed out of the best data set using RS Means. These models are developed to enable school boards to predict the initial costs of new conventional school buildings associated with applying various structure and envelope types. These regression models are built to be best-fitted to the data at hand, and also to be simple and easily applied by decision makers on school boards.

The computed data is stored in Microsoft Excel due to the versatility of spreadsheet analysis. The Minitab® statistical software package was selected for developing the various regression models. The corresponding data for the deformed variables is installed in Minitab® for regression analysis.

The output from Minitab® consists of constructed regression equations with an estimate of regression coefficients “β_k” for the analyzed data.

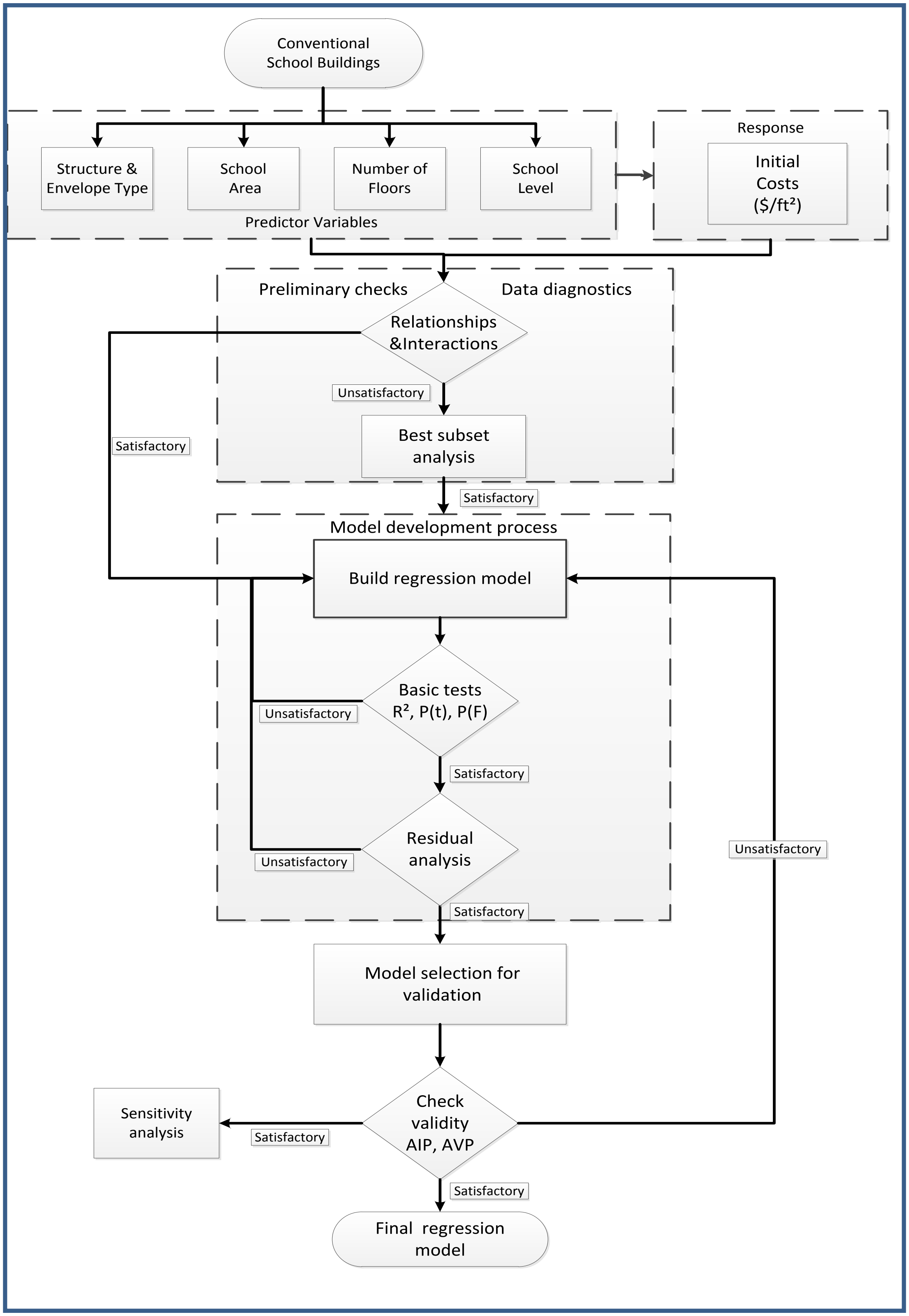


Figure 1: Regression model development process

3.1 Preliminary Tests for Model Adequacy

The preliminary tests of the regression models include: coefficient of multiple determinations (R^2), a regression relation test (F), and a (t) test for each regression parameter's coefficient "β_k". The R^2 value measures the predictor variables' variance, or the fitting of data in correlation to "initial costs" (response variable), while the R^2 (adjusted) accounts for the number of predictors in the model. Both values should indicate that the model fits the data well.

The second test is the regression relation test (F). To determine P (F) for the whole model, a hypothesis test is applied. The assumption of the null hypothesis (H_0) is that all regression coefficients, β_0, $\beta_1 \ldots \beta_{p-1}$ are zero i.e. $\beta_0 = \beta_1 = \beta_{p-1} = 0$. The assumption of the alternate hypothesis (H_a) is that not all coefficients are equal to zero. If the p-value (statistical significance) is 0.00, it means that the null hypothesis is rejected. This hypothesis proves that the estimated model is significant at an α - level of 0.05, indicating that at least one coefficient in the developed regression model is not equal to zero.

The third test is to verify if all of the predictors are significantly corresponded to the response variable or not. "t-tests" are performed individually to determine the validity of regression coefficient, and are performed separately for β_0, $\beta_1 \ldots \beta_{p-1}$ in a similar fashion. In the case of β_0, the null hypothesis (H_0) of the t-test assumes that $\beta_0 = 0$; while the alternative hypothesis (Ha) assumes that $\beta_0 \neq 0$.

3.2 Residual Analysis

After diagnosing the coefficients and bases satisfactorily, the next step is to analyze the residuals and their patterns. Checking the normality of error is performed to verify the linearity correlation assumptions. Normal probability and frequency is represented in a plotted graph in the developed models in order to perform the residual analysis.

3.3 Testing the Regression Model's Validity

The first step in the validation is to compare the actual observation with the predicted values for the validation data for each developed model. This validation is performed using the excluded 20% data points and plotted to compare the prediction model with the observed data in hand. The mathematical validation method is performed using the average validity and invalidity percent. Average invalidity and validity percent is computed in this study for data validation using the following formulae (Zayed and Halpin 2005):

$$AIP = \frac{\sum_{i=1}^{n} \left| 1 - \left(\frac{E_i}{C_i} \right) \right|}{n}$$

(Equation 1) (Zayed and Halpin 2005)

and $$AVP = 1 - AIP$$ (Equation 2)

where AVP is the average validity percent, AIP is the average invalidity percent, E_i is the Predicted Value, C_i is the Actual Value, and n is the number of observations. The AIP value varies from 0 to 1.

4. Analysis and Results

Seven multi-regression models are developed in this study to predict the square footage initial costs for new conventional school buildings. Each model is developed to predict the specified structure and envelope type with regards to correlated predictor variables that include: school area (square foot), number of floors (which ranges from1-4), and school level, which ranges from 1-3: elementary school (1), middle school (2), and high school (3). The process of model development is applied to the entire range of prediction models and can be explained below:

4.1 Wood Structure with Concrete Brick Walls Model (WC)

4.1.1 Correlation Tests

The first step is to test the linearity of the data by detecting the possible correlation from the obtained scatter plot matrix and the correlation matrix with the transformed Y´ variable; these plots are presented in Figure 2.

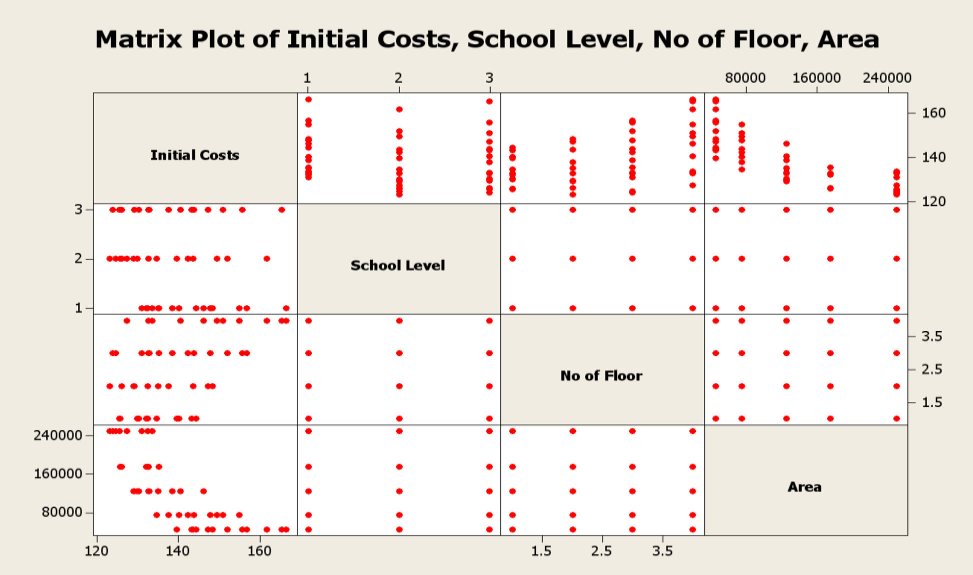


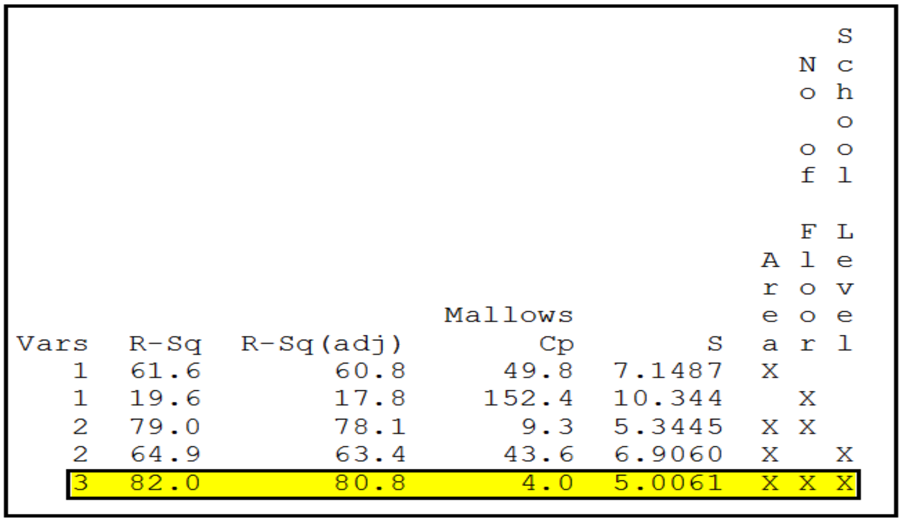
Figure 2: scatter matrix plot for regression model parameters

The plots’ representation shows that the data is constant and distributed evenly across the graph without forming any pattern. All of these plots indicate that each of the predictor variables is nearly linearly associated with the response variable and so the plots are considered satisfactory.

4.1.2 Best Subset Analysis

The output of subset regression analysis generates various regression models in each line, as shown in table 6.6. In the WC regression model, the highest values of R^2 and R^2 (adjusted) are recorded at 82.0% and 80.8% respectively, while the lowest values of C_p and standard deviation (S) are recorded at 4.0 and 5.0, respectively as shown in table 1. The result of the best subset analysis proves that all predictors are significant and should be combined and included in the developed regression model. This combination of variables is proven to be the best case in all seven of the developed regression models.

Table 1: Best subset analysis result using Minitab



```
                                              S
                                           N  c
                                           o  h
                                              o
                                           o  o
                                           f  l

                                           F  L
                                        A  l  e
                                        r  o  v
                       Mallows          e  o  e
Vars  R-Sq  R-Sq(adj)       Cp       S  a  r  l
   1  61.6       60.8     49.8  7.1487  X
   1  19.6       17.8    152.4  10.344     X
   2  79.0       78.1      9.3  5.3445  X  X
   2  64.9       63.4     43.6  6.9060  X     X
   3  82.0       80.8      4.0  5.0061  X  X  X
```

$$IC(WC) = 148 - 0.000123 \times Area + 4.17 \times No.of\ floor - 2.39 \times School\ level$$

(Equation 3)

Where:
IC: Predicted initial costs,
WC: Wood structure with concrete brick walls
Area: School area in square feet
Number of floors: Ranges between 1-4 floors
School level: elementary school (1), middle (2), and high school (3)

Table 2: Statistical diagnostic of the WC model

Predictor	Coefficient	SE Coef.	T	P
Constant	148.494	2.902	51.18	0.000
Area	-0.00012269	0.00001009	-12.16	0.000
No of Floor	4.1745	0.6469	6.45	0.000
School Level	-2.3922	0.8861	-2.70	0.010

S = 5.00609 , R-Sq = 82.0% , R-Sq(adj) = 80.8%

Table 3: Analysis of Variance of WC model

Source	DF	SS	MS	F	P
Regression	3	5018.2	1672.7	66.75	0.000
Residual Error	44	1102.7	25.1		
Total	47	6120.9			

4.1.3 Tests of Model Adequacy

The developed model shows that a positive correlation is detected and that the number of floors variable is linked to the response variable (initial costs). On the other hand, a negative relationship is correlated between school area and school level with the predicted initial costs.

1. **Test of R^2 and R^2 (adjusted)**

The result of the preliminary tests shows that R^2 and R^2 (adjusted) values are recorded at 82.0% and 80.8%, respectively. The R^2 value indicates that the predictor variables explain 82.0% of the variance in the response variable (initial costs) for the WC model. The R^2 (adjusted) value is a modification of R^2 that adjusts for the number of explanatory terms in a model. The standard deviation of data (S) is recorded at 5.00. These R^2 values indicate that the data fits well in the built model.

2. **t-tests**

This test is performed to test if all predictors are significantly correlated to the response variable. The p-values for the estimated coefficients for predictors “School area” and “No. of floors” are 0.000 as presented in table 2. Similarly, the p-value for predictor “School level” is 0.010. As a result, the null hypothesis is rejected and the alternative hypothesis is accepted. This indicates that the predictors are significantly correlated to the response variable “initial costs” at an α - level of 0.1.

3. **F-Test**

The p-value (statistical significance) in the analysis of variance is 0.000 as shown in table 3. The null hypothesis is thus rejected. This shows that the estimated model is significant at an α - level of 0.05. Consequently, at least one coefficient in the developed regression model is not equal to zero.

4. **Residual Analysis (Normality of Errors)**

The normal probability plot indicates that error terms are approximately normally distributed. Minor departures from normality are observed as presented in graph 3; the normal probability plot and the histogram of residuals plot. These departures are considered as unusual possible outliers. R^2 values and other statistical parameters could be improved by eliminating these outliers; however, the model would not be the best representation of the real world data in hand. The result of the residual analysis is satisfactory since a few minor departures from normality do not indicate any serious problems (Kutner et al 2005).

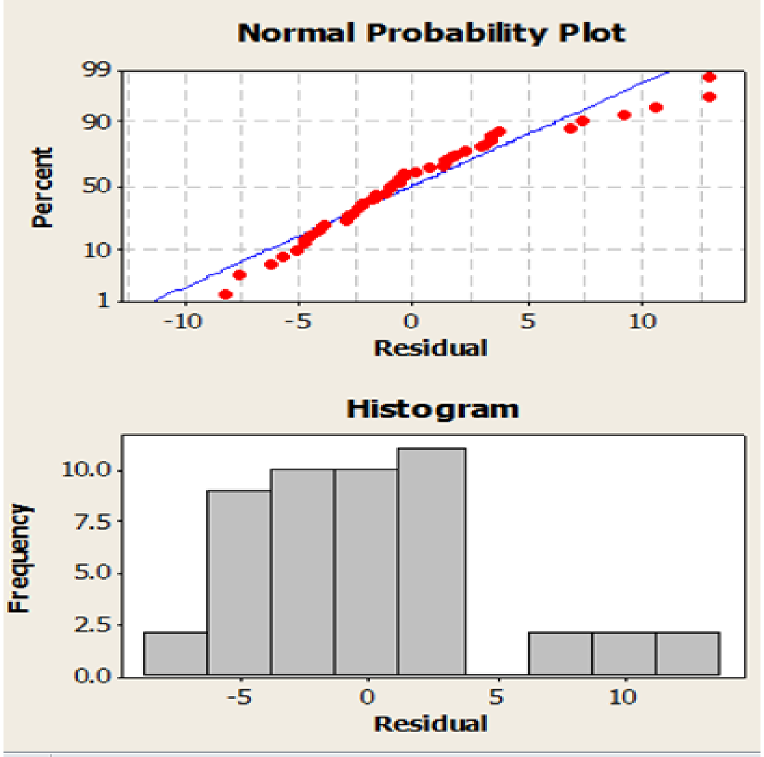
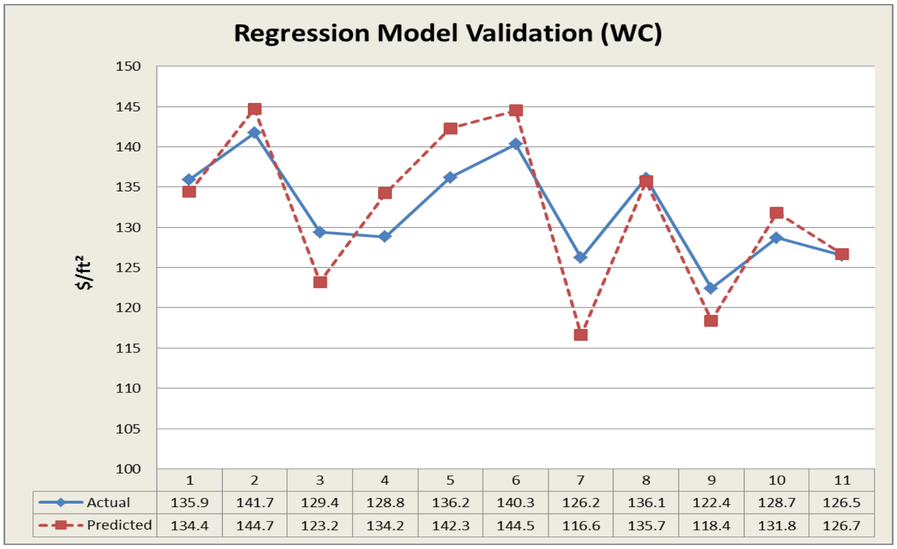


Figure 3: Normal probability plot and histogram resulted from Residual analysis

4.1.4 WC Model Validation

1. **Plot Validation Method**

Figure 4 presents the plot validation method for the actual observation vs. predicted output plot. This representation indicates that the predicted values are scattered around the actual values for the response variable. Therefore, the first validation test's results are considered to be satisfactory.



	1	2	3	4	5	6	7	8	9	10	11
Actual	135.9	141.7	129.4	128.8	136.2	140.3	126.2	136.1	122.4	128.7	126.5
Predicted	134.4	144.7	123.2	134.2	142.3	144.5	116.6	135.7	118.4	131.8	126.7

Figure 4: Plot validation for the WC regression model

2. **Mathematical Validation**

$$AIP = \frac{\sum_{i=1}^{n} \left| 1 - \left(\frac{E_i}{C_i} \right) \right|}{n} = AIP = \frac{0.3339}{11} = 0.0303 \quad , \quad AVP = 1 - AIP = 0.9696$$

The value of validation indicates that the predicted model is almost 96.9% accurate. The final validation results can be considered to be more than satisfactory because the WC model explains about 96.9% of the variation in the validation data.

4.2 Other Developed Models and Validations

Table 4 presents the other developed models for the various alternatives, their indicators such as coefficients of determination, variances, and validations. The R^2 values are found to be ranged 71.3%-79.6% for the various alternatives, R^2 adj. values are varied between (69.3%-78.2%), and standard deviation values are varied (4.05-10.5). As for the validation, the average values of invalidity are recorded between (2.2-4.2%) while the average of validity values are recorded between (95%-97.8%) as presented in table 4.

Table 4: Developed Models, Indicators and Validations

System	Developed Regression Forecasting Models for Initial Costs				
Wood Struct.-Wood Walls (WW)	$145 - 0.000095 \times Area + 1.92 \times No.of\ floors - 2.4 \times School\ level$				
	R^2	**R^2 adj**	S	AIP	AVP
	78.4%	77%	4.05	2.6%	97.4%
Steel Struct. - Steel Wall (SS)	$159 - 0.000081 \times Area + 2.72 \times No.of\ floors - 2.38 \times School\ level$				
	R^2	**R^2 adj**	S	AIP	AVP
	73.2 %	71.4%	4.33	2.2%	97.8%
Steel Struct. -Exterior Brick I (SC)	$164 - 0.000115 \times Area + 5.48 \times No.of\ floors - 3.38 \times School\ level$				
	R^2	**R^2 adj**	S	AIP	AVP
	79.6%	78.2%	5.58	2.7%	97.3%
Steel Strct. Wood Stud Walls (SW)	$160 - 0.000087 \times Area + 3.16 \times No.of\ floors - 2.29 \times School\ level$				
	R^2	**R^2 adj**	S	AIP	AVP
	74.1%	72.4%	4.49	2.6%	97.4%
Concrete Structure - Cavity Walls (CM)	$182 - 0.000158 \times Area + 5.14 \times No.of\ floor - 2.72 \times School\ level$				
	R^2	**R^2 adj**	S	AIP	AVP
	71.3%	69.3%	8.61	3.3%	96.7%
Concrete Struc. - Precast walls (CC)	$188 - 0.000208 \times Area + 8.64 \times No.of\ floors - 2.24 \times School\ level$				
	R^2	**R^2 adj**	S	AIP	AVP
	75.9%	74.2%	10.5	4.2%	95.8%

5. Summary

Developing of initial costs Forecasting Models using regression is done to assist decision makers on school boards to predict and then select the best structure and exposure types of new school buildings based on their economic performance. These developed mathematical models can be manually used within North American Cities after adjusting the city index (Montreal) and the time index (2011) to estimate the initial costs for the various systems for elementary, middle, and high school buildings. The models adequacy tests for R^2 values are found to be ranged 71.3%-82.0%, which indicate that these models very good models and their linear regression lines explain very well the variances in the response variables (initial costs). The regression models indicate that there is a positive correlation between the number of floors variable and the response variable (initial costs). At the same time, a negative correlation is identified between initial costs and school area as well as school level. Finally, validation results can be considered to be more than satisfactory because the developed models explain about 95.8%-97.8% of the variation in the validation data. These models are the best representation of the real world data in hand.

References

Alshamrani, O, “Evaluation of School Buildings Using Sustainability Measures and Life-Cycle Costing Technique”, July 2012, Doctor of Philosophy Thesis in Civil Engineering, Department of Building, Civil, and Environmental Engineering, Concordia University, Montreal, Quebec, Canada.

Dell’Isola, A. J., and Kirk, S. J., “Life Cycle Costing for Facilities”, First Edition, 2003, by Reed Construction Data Inc, United States of America, ISBN 0-87629-720-5.

Flangan, R., Norman G., and Furbur, J. D., “Life Cycle Costing for Construction”, 1983, Report by Quantity Surveyors Division of the RICS, Department of Construction Management, University of Reading.

Kutner, M. Neter, J. Nachtsheim, C & Wasserman, W., “Applied Linear Statistical Models”, 2005, 5th Edition, McGraw-Hill Companies Inc. USA

Maciha, J. C., “Preventive Maintenance Guidelines for School Facilities”, 2000, RS Means, Reed construction data, Inc. Construction Publishers and Consultants, USA, ISBN 978-0-87629-579-3.

Minitab Statistical Software Features, “Minitab Software for Statistics, Process Improvement ,Minitab” 2011, available online at: http://www.minitab.com/en- /minitab/features/>, accessed [21 Apr 2011].

Ontario Ministry of Education, “Capital Investments – Improving Ontario's Schools”, 2013, available online at: http://www.edu.gov.on.ca/, accessed [4 Feb 2013].

Reed Construction Data Inc., RSMeans, “RSMeans Construction Square Foot Cost calculator”, 2011, available online at: http://rsmeans.reedconstructiondata.com/Means CostWorks.aspx, Accessed on [11 FEB 2011].

Zayed, T & Halpin, D, “Productivity and Cost Regression Models for Pile Construction”, ASCE Journal of Construction engineering and Management, Volume 131, No 7, July 2005.