Regression Modelling for Prediction of Construction Cost and Duration

Article · November 2016

DOI: 10.4028/www.scientific.net/AMM.857.195

2 authors:

Nivea Thomas
Indian Institute of Technology Delhi
3 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Anu V. Thomas
TKM College of Engineering, Kollam
4 PUBLICATIONS **30** CITATIONS
SEE PROFILE

Applied Mechanics and Materials
ISSN: 1662-7482, Vol. 857, pp 195-199
doi:10.4028/www.scientific.net/AMM.857.195

Submitted: 2016-05-21
Revised: 2016-08-07
Accepted: 2016-08-19
Online: 2016-11-15

Regression Modelling for Prediction of Construction Cost and Duration

Nivea Thomas[1, a *] and Dr. Anu V. Thomas[2,b]

[1]PG Scholar, Dept. of Civil Engineering, TKM College of Engineering, Kollam, Kerala, India

[2]Associate Professor, Dept. of Civil Engineering, TKM College of Engineering, Kollam, Kerala, India

[a]niveathomas93@gmail.com, [b]anuthomastkmce@gmail.com

Keywords: Statistical regression; Modelling; Relationship; Validation.

Abstract. Construction investments are sensitive to time and cost overruns. Delay and cost escalation are considered two threats to project success. The project objective is to develop a model to predict project cost and duration based on historical data of similar projects. Statistical regression models are developed using real data of building projects. The methodology is adopted in 3 steps: a) Data collection b) Statistical analysis using Statistical Package for Social Sciences (SPSS) software c) Interpretation of results. The real data of cost and duration of 51 building projects have been collected. In statistics, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modelling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. The analysis is done using SPSS developed by IBM Corporation. The Regression models have been developed using the data collected from Noel Builders, Kakkanad, Ernakulam to predict the project cost and duration. The developed models are validated using split sample approach. The model outputs can be used by project managers in the planning phase to validate the scheduled critical path time and project budget.

Introduction

Regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Thus, it provides a good basis for estimating the cost and duration. If y is a dependent variable and $x_1, \ldots, x_k$ are independent variables then the multiple regression model provides a prediction of y from the x_i of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

where $\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ is the deterministic portion of the model and ε is the random error. We further assume that for any given values of the x_i the random error ε is normally and independently distributed.

The multiple regression model is based on the following assumptions:

1. ***Linearity***: The dependent variable y can be expressed as a linear combination of the independent variables $x_1, \ldots, x_k$.
2. ***Independence***: Observations are selected independently and randomly from the population.
3. ***Normality***: Observations are normally distributed.
4. ***Homogeneity of variances***: Observations have the same variance [1].

Data Analysis and Results

Data obtained were analyzed using SPSS software and following result were obtained :

Table 1 Descriptive Statistics of the data variables

PARAMETERS	AREA (Sq.Ft.)	ESTIMATED COST (Rs)	ESTIMATED DURATION (Days)	ACTUAL COST (Rs)	ACTUAL DURATION (Days)
Number	51	51	51	51	51
Mean	2920.18	5432651.961	329.69	5882148.627	360.88
Std. Deviation	746.554	1010927.8919	37.169	1121646.3706	36.621
Skewness	.008	.478	.151	.514	-.104
Std. Error of Skewness	.333	.333	.333	.333	.333
Kurtosis	-.806	-.383	-.747	-.426	-.568
Std. Error of Kurtosis	.656	.656	.656	.656	.656

Table 1 shows the descriptive statistics of the data variables: the mean, standard deviation, the skewness and kurtosis measures with their standard errors. The skewness and kurtosis measures should be as close to 0 as possible, in SPSS. A small variation from 0 is not a problem, as long as the measures are not too large compared to their standard errors. Skewness z-value = (Skewness measure / Standard error) and Kurtosis z-value = (Kurtosis measure / Standard error).The z-value should be between -1.96 to +1.96. Then, they are normally distributed, in terms of skewness and kurtosis [6].

Here the Skewness and Kurtosis z-values are: Area (0.024 and -1.23); Estimated Cost (1.435 and -0.584); Estimated Duration (0.453 and -1.138); Actual Cost (1.543 and -0.649) and Actual Duration (-0.312 and -0.866) respectively. Hence they are normally distributed.

Regression Model for Actual Cost

In the regression analysis, the Actual cost is set as the (*Y*) dependent variable. The assigned independent variables (*X*) are: project area, Estimated cost, and Estimated duration. Since Estimated duration is not statistically significant (Significance value = 0.294 > 0.05) for the Actual cost regression model, it is not taken as an independent variable.

Table 2 Regression Model for Actual cost

Variables	Unstandardized Coefficients		Standardized Coefficients	Sig.
	B	Std. Error	Beta	
(Constant)	-264240.249	37849.883		<.0001
AREA (Sq.Ft.)	-146.963	23.370	-.098	<.0001
ESTIMATED COST (Rs)	1.210	.017	1.091	<.0001

Equation of the Multiple Regression Analysis for the Actual Cost from Table 2 is given by,

1. ACTUAL COST = -146.963 (AREA) + 1.210 (ESTIMATED COST) – 264240.249

The standardized coefficients reveal estimated cost to have a greater influence on actual cost than the other independent variables.

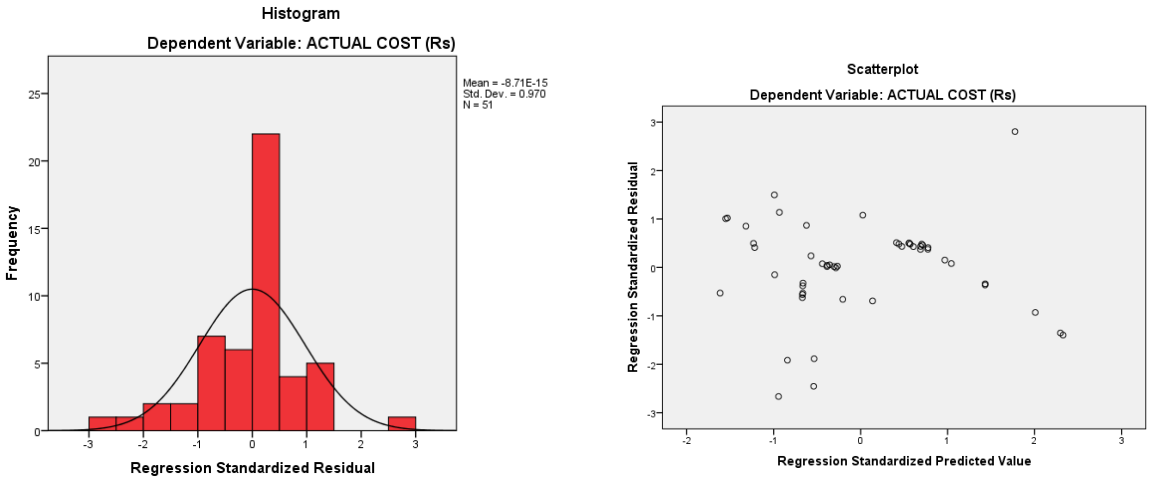

Fig. 1 Histogram of Residuals for the Actual cost regression model and Plot of Regression Standardized Residuals against the Standardized Predicted Values for the Actual cost Regression model

Fig. 1 shows the histogram of the residuals for the developed model. The histogram shows that the residuals are normally distributed and it also presents a plot of the regression standardized residuals against the regression standardized predicted values. The points are approximately randomly distributed in the plot and hence indicate that the assumption of homoscedasticity or equality of variances are met [6].

Table 3 Correlation Matrix for Actual cost

PEARSON CORRELATION	ACTUAL COST (Rs)	AREA (Sq.Ft.)	ESTIMATED COST (Rs)	ESTIMATED DURATION (Days)
ACTUAL COST (Rs)	1.000	.929	.999	.898
AREA (Sq.Ft.)	.929	1.000	.942	.722
ESTIMATED COST (Rs)	.999	.942	1.000	.890
ESTIMATED DURATION (Days)	.898	.722	.890	1.000

The correlation coefficient matrix of Table 3 depicts the linear relationship between each two variables. The coefficient of correlation is a value between 0 and 1. Additionally, a value closer to +1 indicate a strong relationship, a value of zero indicate no relationship among the two variables.
R square is the "percent of variance explained" by the model. It indicates the proportion of the variance in the dependent variable that is predictable from the independent variable [6]. Here 99.9% of the variance can be explained by the model . This may be due to the homogeneous region of study.

Regression Model for Actual Duration

In the regression analysis, the Actual duration is set as the (*Y*) dependent variable. The assigned independent variables (*X*) are: project area, Estimated cost, and Estimated duration. Since Estimated cost is not statistically significant (Significance value = 0.716 > 0.05) for the Actual duration regression model, it is not taken as an independent variable.

Table 4 Regression Model for Actual duration

Variables	Unstandardized Coefficients		Standardized Coefficients	Sig.
	B	Std. Error	Beta	
(Constant)	68.887	9.106		<.0001
AREA (Sq.Ft.)	.012	.002	.248	<.0001
ESTIMATED DURATION (Days)	.778	.037	.789	<.0001

Equation of the Multiple Regression Analysis for the Actual Duration from Table 4 is given by,

2. ACTUAL DURATION = 0.012 (AREA) + 0.778 (ESTIMATED DURATION) + 68.887

The standardized coefficients reveal estimated duration to have a greater influence on actual duration than the other independent variables.

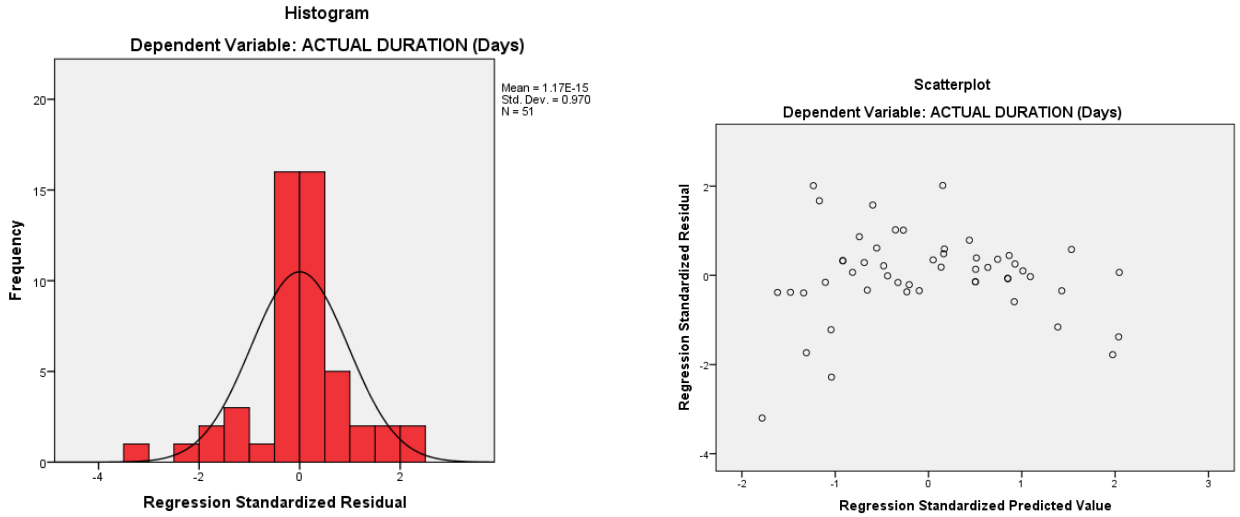

Fig. 2 Histogram of Residuals for the Actual duration regression model and Plot of Regression Standardized Residuals against the Standardized Predicted Values for the Actual duration Regression model

Fig. 2 shows the histogram of the residuals for the developed model. The histogram shows that the residuals are normally distributed and it also presents a plot of the regression standardized residuals against the regression standardized predicted values. The points are approximately randomly distributed in the plot and hence indicate that the assumption of homoscedasticity or equality of variances are met [6].

Table 5 Correlation Matrix for Actual duration

PEARSON CORRELATION	ACTUAL DURATION (Days)	AREA (Sq.Ft.)	ESTIMATED COST (Rs)	ESTIMATED DURATION (Days)
ACTUAL DURATION (Days)	1.000	.818	.937	.969
AREA (Sq.Ft.)	.818	1.000	.942	.722
ESTIMATED COST (Rs)	.937	.942	1.000	.890
ESTIMATED DURATION (Days)	.969	.722	.890	1.000

The correlation coefficient matrix of Table 5 depicts the linear relationship between each two variables. The coefficient of correlation is a value between 0 and 1. Additionally, a value closer to +1 indicate a strong relationship, a value of zero indicate no relationship among the two variables.

R square is 96.8% which is the “percent of variance explained” by the model.

Model Validation

The regression model developed was validated by the split sample approach. The data was split into two parts, with 75% of the data randomly chosen for estimating the regression model and the remaining data used for validating the model. To measure the overall predictive fit, adjusted R square is calculated [1]. The significant variables in the model estimated with the split sample were the same based on the entire dataset. In case of actual cost, the adjusted R square value is 0.999 for the split sample and the entire dataset and for actual duration, the adjusted R square is 0.964 and 0.968 for the split sample and entire dataset respectively. These values are very close and thereby the regression model is validated. The regression models are also shown below.

Regression model for Actual cost in split sample approach:

3. ACTUAL COST = -153.110 (AREA) + 1.215 (ESTIMATED COST) – 274263.417

Regression model for Actual duration in split sample approach:

4. ACTUAL DURATION = 0.010 (AREA) + 0.799 (ESTIMATED DURATION) + 67.673

The regression models in split sample is similar to that of the entire dataset. The actual cost and duration of the remaining 25% of the dataset can be calculated using these models.

Conclusion

- This project utilizes a real time approach in estimating project cost and duration.
- Regression model equations for actual cost and actual duration are obtained which can be used by the project managers.
- The model results predict the amount of time and money that should be budgeted to the project. Critical path durations and budgeted costs should be increased or decreased up to the regression models outputs.
- The project outcome has realized a realistic schedule with budgeted cost and duration that can incorporate lessons learnt from similar history projects. The findings of the present study are limited to the data collected from 51 building projects of the same construction firm.

References

[1] Hair J. F., W. C. Black, B. J. Babin, R. E. Anderson and R. L. Tatham, Multivariate Data Analysis, Pearson Education, Sixth ed., (2011)

[2] Hammad A., M. A. Ali, G. J. Sweis and A. Bashir, Prediction Model for Construction Cost and Duration in Jordan, Jordan Journal of Civil Engineering, 2008, 2, pp. 250-266.

[3] Hammad A., M. A. Ali, G. J. Sweis and R. J. Sweis, Statistical Analysis on the Cost and Duration of Public Building Projects, Journal of Management in Engineering, 2010, 26, pp. 105-112.

[4] Hwang S., Dynamic Regression Models for Prediction of Construction costs, Journal of Construction Engineering and Management, 2009, 135, pp. 360-367.

[5] Koushki P. A., K. Al-Rashid and N. Kartam, Delays and cost increases in the construction of private residential projects in Kuwait, Construction Management and Economics, 2005, 23, pp. 285-294.

[6] Malhotra N. K. and S. Dash, Marketing Research: An Applied Orientation, Pearson Education, Sixth ed., (2010)