Applying Regression Analysis to Predict and Classify Construction Cycle Time

M. F. Siu[1], R. Ekyalimpa[2], M. Lu[3] and S. Abourizk[4]

[1]Construction Engineering and Management, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Alberta, Canada; PH (780) 655-8532; email: siumingfungfrancis@gmail.com
[2]Construction Engineering and Management, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Alberta, Canada; PH (780) 492-3496; email: rekyalimpa@ualberta.ca
[3]Construction Engineering and Management, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Alberta, Canada; PH (780) 492-5110; email: mlu6@ualberta.ca
[4]Construction Engineering and Management, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Alberta, Canada; PH (780) 492-8096; email: abourizk@ualberta.ca

ABSTRACT

Regression techniques are commonly used for addressing complicated prediction and classification problems in civil engineering thanks to its simplicity. For a given dataset, the linear regression from the input space to the output variables can be achieved by using the "least square error" approach, which minimizes the difference between the predicted and actual outputs. The "least mean square" rule can also be used as a generic approach to deriving solutions on linear or non-linear regressions. The paper addresses the fundamental algorithms of "least square error" and "least mean square" in order to facilitate the prediction and classification of cycle times of construction operations. The classic XOR problem is selected to verify and validate their performances. A viaduct bridge was installed by launching precast girders with a mobile gantry sitting on two piers. The effectiveness of regression techniques in classifying and forecasting the cycle time of installing one span of viaduct considering the most relevant input factors in connection with operations, logistics and resources are demonstrated.

INTRODUCTION

Classification and predication play an important role to the success of planning and control of a construction project. The production rate tracking can be beneficial to forecasting project performances such as the expected activity and project completion times. Corrective actions or changes can be made to tackle any adverse impact on schedule. Yet, applications of artificial intelligence are hampered by its complexities. The operations simulation and artificial neural computing are the two main streams of research which have focused on cycle-time and production rate predictions of construction operation processes. Previous research efforts by Teicholz

(1963), Ahuja and Nandakumar (1985), Halpin (1977), Lu (2003), AbouRizk (2010) have greatly contributed to advancing production rate predictability by implementing simulation techniques. Contributions are claimed to have better planning and control during the project executions, such as estimating the project duration and cost under uncertainty, productivity improvement and potential savings in time or cost (Halpin, 1977, Pritsker et al. 1989, Sawhney and AbouRizk 1995, AbouRizk and Mohamed 2000, Lu 2003 and Tian et al. 2010).

The regression modeling, formally named as linear regression analysis, is profoundly addressed in this paper, which is the earliest prototype that heralded the later development of artificial neural networks. Linear regression techniques are widely applied in the construction field, such as Thomas and Sakarcan (1994); Sanders and Thomas (1993). The basic mechanism is to determine the output y by multiplying each input variable x and an associated weight w as shown in Eq. (1). The weights can be determined by using historical records. As linear regression evolved into neural network computing, research efforts such as Mutlu al et. (2008) used artificial neural networks to forecast the daily water flow at multiple gauge stations in the agricultural domain; Portas and AbouRizk (1997), Lu et al. (2000), embedded fuzzy logic into neural computing and implemented sophisticated probabilistic inference neural networks to estimate labor production rates, respectively. The neural network computing research has mainly focused on advancing methodologies (single- and multi-layer perceptrons, radial-basis functions and support vector machines) in terms of computing efficiency and effectiveness (Haykin, 1998).

However, sufficient knowledge of neural network computing is required on construction professionals in order for them to fully trust and harness these advanced tools instead of simply using them as "black box". This research has contributed to elucidating on fundamental regression analysis techniques underlying neural network computing, namely: (1) the "least square error" and (2) the "least mean square" algorithms, in an effort to facilitate prediction and classification applications in construction engineering. The following sections provide mathematical background, algorithmic verification using the XOR dataset, and implementation of regression techniques to predict and classify precast bridge segment erection cycle times.

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots \tag{1}$$

LEAST SQUARE ERROR APPROACH

The least square error approach is the earlier form used for numerical regression and prediction. The general equation of linear regression model is in the form of Eq. (1). The weight parameters (w_n) could be optimized and stabilized by analyzing available input data (x_n). The analytical matrix-approach transforms Eq. (1) in forms of vector as given in Eqs. (2) and (3). The error is defined as the difference between the actual output y and the model's output (summation of xw.) By least square adjustment techniques, the error of the system can be minimized by taking its partial derivatives with respect to w_n which are set as zero to derive optimal solutions, as shown in Eqs. (4) to (6). The Eqs. (7) to (9) show a system of rearranged equations which can be expressed in a compact matrix form as Eq. (10). The weight w parameters can be easily calculated by matrix manipulations only involving

parameters of input x and output y. It is noteworthy that w_0 in Eq. (1) is the error (noise or disturbance) term commonly defined in applied statistics. This term is essential to account for any unobserved random variable that presents noise to the linear relationship between input x and output y. According to Harrell (2001), the error term should be statistically independent and identically distributed, approximately normally distributed and have a common variance. To prove the effect of unobserved random variables on the linear regression model, Eq. (1) can be interpreted as Eq. (11). The linear regression system explicitly considers the inputs from 1 to p while factoring in both observed inputs (1 to p) and unobserved inputs (p to ∞). The indispensable bias term maintains the integrity of regression techniques resulting in a unique solution. In contrast, the "least mean square rule" approach utilizes iterative procedures to stabilize the weights for each input variable along with the bias, eventually minimizing errors on model outputs.

$$W = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix};\ X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \tag{2}$$

$$y = f(X,W) = X^T W \tag{3}$$

$$\frac{\partial \sum \text{error}^2}{\partial w_0} = \sum 2(y_i - (w_0 + w_1 x_{i1} + w_2 x_{i2}))(-1) = 0 \tag{4}$$

$$\frac{\partial \sum \text{error}^2}{\partial w_1} = \sum 2(y_i - (w_0 + w_1 x_{i1} + w_2 x_{i2}))(-x_{i1}) = 0 \tag{5}$$

$$\frac{\partial \sum \text{error}^2}{\partial w_2} = \sum 2(y_i - (w_0 + w_1 x_{i1} + w_2 x_{i2}))(-x_{i2}) = 0 \tag{6}$$

$$\sum y_i = n w_0 + w_1 \sum x_{i1} + w_2 \sum x_{i2} \tag{7}$$

$$\sum x_{i1} y_i = w_0 \sum x_{i1} + w_1 \sum x_{i1} x_{i1} + w_2 \sum x_{i2} x_{i1} \tag{8}$$

$$\sum x_{i2} y_i = w_0 \sum x_{i2} + w_1 \sum x_{i2} x_{i1} + w_2 \sum x_{i2} x_{i2} \tag{9}$$

$$W = (X^T X)^{-1} X^T Y \tag{10}$$

$$y_i = \sum_{j=1}^{\infty} w_j x_{ij} = \sum_{j=1}^{p} w_j x_{ij} + \sum_{j=p}^{\infty} w_j x_{ij} = \sum_{j=1}^{p} w_j x_{ij} + w_0 \tag{11}$$

LEAST MEAN SQUARE APPROACH

The least mean square rule expresses the partial derivative in terms of vector $\bar{G}$ defined as the system error in linear regression analysis, as given in Eq. (12). The least mean square rule represents a generic approach to optimize the weights by continuously seeking stabilized value of the bias w_0 expressed as Eq. (13). The objective is to progressively evaluate w by applying iterative procedures aimed at minimizing the system error, as given in Eqs. (14) to (17). $Y(i)$ in Eq. (14) denotes the output. The error in relation to each input pattern can be computed by subtracting the desired output $d(i)$ and $Y(i)$ as Eq. (15). W in Eq. (16) and w_0 in Eq. (17) can be updated by using the learning rate parameter η until the values are steady, when the optimum solution of $W(i)$ is reached. The range of η can be estimated by the

eigenvalue of the correlation matrix $\boldsymbol{X}$. Generally, more time is needed to search the optimum solution if the step-size chosen is too small, while no solution can be reached if the value is too large due to increased chances of divergence. As the data may be highly linearly-non-separable, pre-processing data is essential to transform the data from being linearly-non-separable to being linearly-separable.

$$\vec{G} = \frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_o}\frac{\partial y_o}{\partial w_{ij}} = -(d_o - y_o)y_i \tag{12}$$

$$W(\text{new}) = W(\text{old}) - \eta\frac{\partial E}{\partial w_{ij}} \tag{13}$$

$$Y(i) = W(i) \times X(i) + w_0(i) \tag{14}$$

$$Err(i) = d(i) - Y(i) \tag{15}$$

$$W(i+1) = W(i) + \eta \times Err(i) \times X(i) \tag{16}$$

$$w_0(i+1) = w_0(i) + \eta \times Err(i) \tag{17}$$

DATA PRE-PROCESSING

Data can be difficult to be classified if the input data of the problem are characteristic of high non-linearity. The XOR (eXclusive OR) problem is used to illustrate the "non-linearity" problem. The inputs and outputs are shown in Table 1. Figure 1 graphically plots the four corresponding points; no straight line can be drawn to cluster the points having the same outputs. Thus, direct classification based on linear regression is not feasible owing to the non-linearity in data. To tackle the problem, the raw data is firstly normalized. K-means clustering technique is employed to evaluate the distances between the input pattern and the center of a cluster r_i . Note that the K-means clustering algorithm is a special case of self-organizing maps (Haykin, 1998). The P-nearest neighbor algorithm is then applied to determine the sigma σ of the Gaussian function for each cluster. Finally, the data is transformed by using the Gaussian function as shown in Eq. (18).

The K-means clustering and P-nearest neighbor algorithms are implemented considering two clusters on the XOR dataset. The two clustering centers coordinates are determined as (0, 0), (1, 1) and σ as 1.414. The transformed inputs are thus obtained by data pre-processing, ready for regression analysis. A straight line can separate the four points with into respective classifications. Figure 2 shows the linearly-separable transformed inputs. Both least square error and least mean square rule approaches were applied, achieving the same results. w_1 and w_2 are evaluated both as 20.438 and w_0 is –31.834. A multivariate regression equation can be expressed in Eq. (19) to represent the complete prediction model, where x_1 and x_2 are the two input parameters.

$$x_i = e^{\frac{-r_j^2}{2\sigma^2}} \tag{18}$$

$$Y_{output} = 20.438\, e^{\frac{[(X_1-0)^2+(X_2-0)^2]}{-2(1.414)^2}} + 20.438\, e^{\frac{[(X_1-1)^2+(X_2-1)^2]}{-2(1.414)^2}} - 31.834 \tag{19}$$

Table 1. XOR Inputs and Outputs

Inputs Before Transformation			Inputs After Transformation		
x_1	x_2	y	x_1	x_2	y
0	0	1	1.000	0.607	1
0	1	0	0.779	0.779	0
1	0	0	0.779	0.779	0
1	1	1	0.607	1.000	1

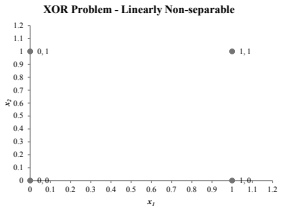

Figure 1. Linear Non-separable

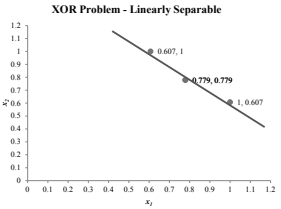

Figure 2. Linear Separable

BRIDGE SPAN ERECTION CASE STUDY

A bridge construction project is used to demonstrate the effectiveness in implementing cycle-time prediction and classification by using regression analysis techniques. The bridge is a new artery linking Hong Kong and Shenzhen, China, and consists of 227 post-tensioned spans of viaduct. A typical span is made up of 14 precast segmental box girders (12m×2.5m×2.8m of each). The stepping girder precast installation method was used to accelerate the viaduct construction process (Chan and Lu, 2009). The precast segments were fabricated near Shenzhen and hauled to the site for installation. The site was too congested to keep all segments in the convenient proximity of the site crew. As an alternative, the precast segments were partially stored in a remote storage area and transported to the working span by trailer trucks, without any intermediate storage or buffer.

In order to assist the contractor in deciding on how many precast segments to be placed at the remote storage area and how far away to locate the remote storage area while maintaining the target one-span erection cycle time of four and half working days, four input factors relevant to site operations and logistics planning were identified and assessed, namely: (1) the number of trailer trucks rented for hauling segments (the site only considered the options of two trailer trucks or three), (2) one-batch or two-batch precast segment delivery modes (14 segments can be delivered either in one batch on the night before installation operations starts or in two batches, which means the first batch of seven segments would be delivered on the night before installation starts and the second batch delivered on the following night), (3) the percentage of the total number of segments on one span to be placed in the remote storage area and (4) the haul duration for a trailer truck to transit from the remote storage area to the working span. Table 2 shows the 30 cycle-time records resulting from simulations, which define 30 different scenarios.

Table 2. Viaduct Installation Cycle-Time Records for Regression Analysis

Rec ID	No. of Tracker	Delivery Batch	Segment at RSA (%)	Duration to RSA	Desired Install Hours	*Predicted Install Hour*	Desired Prod Class	*Predicted Prod Class*
1	2	1	0.00	0.00	103.61	-	0	0
2	3	1	0.50	0.50	104.76	-	0	0
3	3	1	0.29	0.33	104.76	-	0	0
4	3	1	0.50	0.33	104.78	-	0	0
5	3	1	1.00	0.50	105.78	-	0	0
6	3	1	1.00	0.33	105.78	-	0	0
7	3	1	0.29	0.75	108.38	-	1	**0**
8	3	1	0.50	0.75	109.36	-	1	1
9	2	1	0.50	0.50	111.51	-	1	1
10	3	1	1.00	0.75	112.05	-	1	1
11	3	1	0.71	0.75	112.41	-	1	1
12	2	1	0.29	0.75	112.72	-	1	1
13	2	1	1.00	0.50	114.15	-	1	1
14	2	1	0.50	0.75	115.70	-	1	1
15	2	1	0.71	0.75	116.47	-	2	2
16	2	1	1.00	0.75	116.61	-	2	2
17	2	2	0.29	0.50	116.67	-	2	2
18	2	2	1.00	0.33	116.70	-	2	2
19	2	2	1.00	0.75	116.71	-	2	2
20	2	2	0.00	0.00	116.74	-	2	2
21	2	2	0.57	0.50	116.74	-	2	2
22	2	2	0.57	0.75	116.74	-	2	2
23	3	2	0.57	0.50	116.74	-	2	2
24	3	2	0.57	0.75	116.74	-	2	2
25	3	1	0.29	0.50	104.76	-	0	0
26	3	1	0.71	0.50	104.89	-	0	0
27	3	1	0.71	0.33	104.89	-	0	0
28	3	1	0.00	0.00	105.77	-	0	0
29	2	1	0.71	0.33	106.00	-	0	0
30	2	1	0.29	0.50	108.47	-	1	**0**
31	2	2	0.29	0.75	N/A	118.118	2	2
32	3	2	1.00	0.50	N/A	114.425	2	2
33	2	2	1.00	0.50	N/A	118.010	2	2
34	3	2	1.00	0.75	N/A	114.763	2	2
35	2	1	0.29	0.33	N/A	104.030	0	0
36	2	1	0.50	0.33	N/A	105.373	0	0
37	2	1	1.00	0.33	N/A	110.371	1	1
38	2	1	0.71	0.50	N/A	111.135	1	1
39	2	2	0.29	0.33	N/A	115.589	2	2
40	2	2	0.57	0.33	N/A	116.276	2	2

To facilitate the implementation of the algorithms, data normalization, K-means clustering analysis and P-nearest neighbor algorithm were programmed to normalize the non-linear data. Both least square error and least mean square algorithms were coded. It is noteworthy that the number of cluster centers was estimated before executing the K-means clustering algorithm. Cluster center initialization is controlled by a random seed in the computer program. One proposed solution is to evaluate the root mean square error (RMSE) by using *n* desired and computed outputs given one particular random seed, as shown in Eq. (20). The random seed leading to the smallest RMSE will then be used in the regression analysis. The accuracy of predictions significantly depends on the selection of number of cluster centers, and the quantity and quality of available datasets. In general, the larger size of the dataset, the higher quality of the dataset, the higher

accuracy the predictions. The smallest root mean square error generated is 1.582 hour (Figure 3) by trying 100 random seeds on 30 records (n = 30). Ten "unseen" scenarios were predicted by using the derived multivariate regression equation ("predicted install hour" column with ID 31 to 40 in Table 2).

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(x_{\text{computed,i}} - x_{\text{desired,i}}\right)^2}{n}} \tag{20}$$

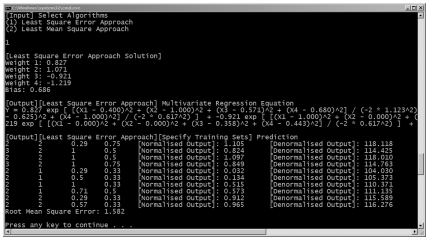

Figure 3. Cycle Time Prediction

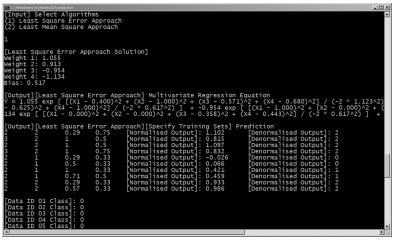

Figure 4. Productivity Classification

Productivity classification is beneficial to decision making during project execution. The contractor defined three classes based on span installation hours (less than 108 hours is "high", between 108 hours and 116 hours is "medium" and longer than 116 hours is "low"). The 30 data set (ID 1 to 30) were chosen to determine the weights of the regression equation. Figure 4 shows the classification results with RMSE 0.327. Only two out of thirty records (ID 7 and 30) are incorrectly classified ("predicted prod class" differing from "desired prod class"). Input patterns can be generalized based on classification results: In order to achieve high productivity, segments must be delivered in one batch, less than 29% segments are stored at remote storage area, and the trailer truck transit time must be within half an hour.

Table 3: Data Manipulations of the Predicted Productivity Classes

Prod Class	No. of Tracker	Delivery Batch	Segment at RSA (%)	Duration to RSA
0 (Class 1 – High)	2 to 3	1	0.00 – 1.00	0.00 – 0.75
1 (Class 2 – Med)	2 to 3	1	0.29 – 1.00	0.50 – 0.75
2 (Class 3 – Low)	2 to 3	1 to 2	0.00 – 1.00	0.00 – 0.75

CONCLUSION

Regression techniques are effective to analyze construction operations in terms of cycle-time prediction and productivity classification. The fundamentals of the "least square error" and the "least mean square" algorithms are contrasted and clarified. Both techniques are verified and validated by using the classic XOR problem and a bridge construction project. Also, the "best fit" classification model is obtained through assigning the number of cluster centers by trial and error. Analytical methods or guidance on how to set up the clusters given a particular dataset can be further generalized.

REFERENCES

AbouRizk, S. 2010. "The role of simulation in construction engineering and management." *Journal of Construction Engineering and Management*, ASCE, 136 (10), 1140–1153.

AbouRizk, S. and Mohamed, Y. 2000. "Simphony-an integrated environment for construction simulation." *Proceedings of Winter Simulation Conference, Orlando*, FL , USA. 1907–1914.

Ahuja, N. T. H., and Nandakumar, V. 1985. "Simulation model to forecast project completion time." *Journal of Construction Engineering and Management*, ASCE, 111(4), 325–342.

Chan, W. H. and Lu, M. 2009 *Artificial intelligence-integrated construction simulation method*, VDM Verlag, Saarbrücken, Germany.

Halpin, D. W. 1977. "CYCLONE - A Method for Modeling Job Site Processes." *J. Constr. Div.*, ASCE. 103(3), 489–499.

Harrell F. E. 2001. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, Springer.

Haykin, S. 1998. *Neural Networks: A Comprehensive Foundation*, Prentice Hall.

Thomas, R. and Sakarcan, A. 1994. "Forecasting Labor Productivity Using Factor Model." *Journal of Construction Engineering and Management*, 120(1), 228–239.

Teicholz, P. 1963. *A Simulation Approach to the selection of Construction Equipment*. Technical Report No. 26, The Construction Institute, Stanford University.

Tian, X., Mohamed, Y. and AbouRizk, S. 2010. "Simulation-based aggregate planning of batch plant operations". *Canadian Journal of Civil Engineering*, 37(10), 1277–1288.

Lu, M. (2003), "Simplified Discrete-Event Simulation Approach for Construction Simulation", *Journal of Construction Engineering and Management*, ASCE, 129(5), 537–546.

Lu, M., AbouRizk, S. and Herman, U. 2000. "Estimating Construction Productivity using Probability Inference Neural Network". *Journal of Computing in Civil Engineering*, 14(4), 241–248.

Portas, J. and AbouRizk, S. 1997. "A Neural Network Model for Estimating Construction Productivity." *Journal of Construction Engineering and Management*, ASCE, 123(4), 399–410.

Pritsker, A., Sigal, C. and Hammesfahr, R. 1989. *SLAM II Network Models for Decision Support*. Prentice-Hall, Englewood Cliffs, N. J.

Mutlu, E, Chaubey, I., Hexmoor, H. and Bajwa, S. G. 2008. "Comparison of artificial neural network models for hydrologic predictions at multiple gauging stations in an agricultural watershed", *Hydrological Processes*, 22, 5097.

Sanders, S. R. and Thomas, H. R. 1993. "Masonry productivity forecasting model", *Journal of Construction Engineering and Management*, ASCE, 119(1), 163–179.

Sawhney, A. and AbouRizk, S. 1995. "HSM–Simulation-based Project Planning Method for Construction Projects." *Journal of Construction Engineering and Management*, ASCE, 121(3), 297–303.