This is the author’s version of a work that was submitted/accepted for publication in the following source:


This file was downloaded from: https://www.escholar.manchester.ac.uk

© Copyright 2006 ASCE

Reproduced in accordance with the copyright policy of the publisher.

Notice: Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source.
Predicting construction cost using multiple regression techniques

David J Lowe, Lecturer, The Manchester Centre for Civil and Construction Engineering, Project Management Division, UMIST, UK

Margaret W Emsley, Lecturer, The Manchester Centre for Civil and Construction Engineering, Project Management Division, UMIST, UK


Correspondence Address:

Dr David J Lowe
The Manchester Centre for Civil and Construction Engineering
Project Management Division
UMIST
PO Box 88
Manchester
M60 1QD
Tel: 0161 200 4643
Fax: 0161 200 4646
E-mail david.lowe@umist.ac.uk

Approximate number of words:

(Version 1: 29/01/04)
Predicting construction cost using multiple regression techniques

Abstract

This paper describes the development of linear regression models to predict the construction cost of buildings, based on 286 sets of data collected in the United Kingdom. The same data have been used to develop neural network model and one purpose of developing the regression models was to provide a benchmark against which the neural networks could be compared.

Raw cost is rejected as a suitable dependent variable and models are developed for three alternatives – cost/m², log of cost and log of cost/m². Both forward and backward stepwise regression analyses were performed to produce a total of six models.

Forty-one independent (predictor) variables were identified and classified either as project strategic, site related or design related. Five variables appeared in each of the six models: GIFA, function, duration, mechanical installations and piling, suggesting that they are the key linear cost drivers in the data.

The best regression model is the log of cost backward model which gives an $R^2$ of 0.661 and a MAPE of 19.3%. The best neural network model is one which uses all 41 variables and a voting system using 100 networks; this gives an $R^2$ value of 0.789 and a MAPE of 16.6%. However, the models developed using both techniques compare...
favourably with past research which has shown that traditional methods of cost estimation have values of MAPE typically in the order of 25%.

Key words: Forecasting, Cost estimating, Construction industry, Regression analysis

**Introduction**

Construction clients require early and accurate cost advice, prior to site acquisition and the commitment to build, to enable them to assess the feasibility of the proposed project, therefore a key task, performed by construction contract price forecasters (usually a Quantity Surveyor in the UK).

The objective of this feasibility (early) stage construction contract price forecast is to generate an indication of a project’s likely construction costs (Ashworth, 1988) to assist the client in setting a budget, predicting the tender price and managing the design so that it meets the budget. When preparing a feasibility forecast, the forecaster usually attempts to predict the winning tender, therefore, the forecast is primarily a forecast of the contractor's forecast, and is made without reference to the contractor's data and with many inherent uncertainties relating to the lack of a detailed design and possibly even a site (Raftery, 1994).

The choice of forecasting technique is determined by its ease of operation, familiarity, speed and a satisfactory degree of accuracy (Ashworth, 1988), in conjunction with the availability of design information. Skitmore and Patchell (1990) provide a comprehensive taxonomy of construction contract price forecasting techniques.
Fortune and Lees (1996) concluded that the traditional rather than the newer techniques of early cost advice are still more popular with practitioners. However, they established that as the size of the organisation increased the more likely that organisation would be to use alternative costing techniques (Fortune and Lees 1994).

Several studies have found that clients are generally dissatisfied with the initial cost advice provided by their construction professionals (RICS, 1984; Ellis and Turner, 1986; and Procter et al. 1993). Further, the RICS (1991) concluded that there is a need to provide more accurate and robust forecasts of construction costs. While it is widely held that a perfect estimate is not possible and even the best possible estimate will always contain a number of key risks, the goal of the forecaster is a practicable level of accuracy (Smith, 1995). Ashworth and Skitmore (1983) maintain that a vital consideration with any method of estimating is the accuracy by which anticipated costs can be predicted. Further, "... any improvement in prediction will be welcomed because existing methods are poor when judged by this criterion".

Both Ashworth and Skitmore's (1983) and Ogunlana and Thorpe (1987) conclude that "... a suitable accuracy of forecasting in the early stages would be of the order of 15% to 20% cv, improving perhaps to around 13% to 18% cv at detailed design stage immediately prior to receiving tenders". Although, Birnie (1993) concluded that the quantity surveyors' cost prediction ability may not be as good as they believed it to be.

Regression analysis and neural networks are two of the modelling techniques, identified by Newton (1991), which have been used to develop models to estimate the
cost of buildings. However, for the most part, these models rely on the use of historic (but recent) cost data. In the United Kingdom, the Building Cost Information Service (BCIS) provides details of construction projects and their associated tender prices. While cost advisors may use these data to advise on the cost of a building, based on the cost of a similar comparable project adjusted to reflect any differences, this does not enable any general truths about the relationships that exist between the cost and significant predictor variables or provide any models which can predict or forecast construction costs. Further, these data reflect the lowest bid price tendered, not the final cost to the client.

Early examples of the use of regression analysis as a forecasting tool are provided by McCaffer (1975) and McCaffer et al. (1984), while a more recent application of the technique to early cost estimating is provided by Trost and Oberlender (2003). Further a review the application of regression analysis to construction price forecasting is presented by Skitmore and Patchell (1990). Likewise, Elhag and Boussabaine (2001; 2002) modelled tender price estimation using artificial neural networks, while Emsley et al. (2002) applied a neural network approach to the prediction of total construction costs.

**Research Objectives**

The initial impetus for the research arose from the paucity of data available that can provide reliable information about the relative costs of using different procurement routes, based on the final cost of construction. However, it soon became apparent that
the type of procurement route could not be modelled in isolated from the other cost significant variables that influence the cost of a building. (Harding et al., 1999).

The main aim of this paper is to describe the development of a robust regression cost model to predict the construction cost (final account) of a building. While such a model would be valuable in itself, developing a model fulfils two other purposes:

- it provides a useful benchmark against which neural network models could be measured; and
- it assists in identifying those variables that demonstrated a strong linear relationship with cost.

The model developed uses data collected from 286 United Kingdom construction projects. As well as developing a regression model, the data have also been used to develop neural network models (Emsley et al., 2002) and a brief comparison between the performances of the two modelling techniques is made.

**Input and output variables**

The data collected were divided into independent input variables and dependent output variables (Emsley et al., 2002).

Two independent variables were identified – construction cost (final account) and client costs (incorporating professional fees and the internal costs incurred by the client), which may be combined to give the total cost to the client. However, the regression models described in this paper are based on construction cost only.
An extensive literature review identified numerous predictor (input) variables. These were finally reduced to 41 variables, which it was believed would be known at the early estimating stage (the stage at which the models are intended to be used). The variables were categorised as shown in Table 1. The influence of time and geographical location were accommodated through the use of the BCIS indices by adjusting all the data sets (projects) to a common location and base date. The cost of activities such as external works and demolition were removed from the final account figure as it was thought that these activities would be extremely difficult to model, as they are unique to specific sites.

Potential models/predictors

Linear regression analysis has, in the past, been performed by using the raw cost as the dependent variable. However, there are a number of assumptions implicit in this choice of variable which must be addressed.

1. The standard deviation in the error associated with the dependent variable (cost) remains constant throughout the domain.
2. This error is normally distributed.
3. The effect of any variable is always expressed in terms of a fixed cost increase or decrease, irrespective of project size or type.

At least two of these assumptions are open to question.
The first is that the standard deviation of error remains constant. That is to say, the cost of a small project can vary by the same monetary amount as a large project. This is highly unlikely to be the case and is much more likely to be proportional to the size or cost of the building. Because regression modelling minimises the squares of the errors, it will be inherently biased towards minimising the errors for very large projects, where the errors are greatest. It is therefore unlikely to be a good predictor of the cost of smaller projects. Given that the costs of projects for which data have been collected vary between £36,000 and £15.8M, the influences of errors in the cost of the largest projects are several orders of magnitude more than those of the smallest projects, so the effect will be pronounced.

The second assumption which is questioned is that the effects of any variable are expressed as a fixed cost change. If, for example, the specification of the floor finishes changed to one of a higher cost, the cost of the building would be expected to rise. However, the cost of a small building would not be expected to rise by the same amount as the cost of a large building and it is more likely that the cost would rise either as a proportion of the building size or as a proportion of the building cost, much the same as with the error.

These criticisms raise questions as to the meaningfulness of models produced by using raw cost as the predictor for a linear regression model. Therefore three other possible models are proposed - predicting the natural logarithm of the cost, the cost/m² and the natural logarithm of cost/m².
Log of building cost

In order to address the problem of the large differences in the cost of building, a common solution would be to model the log of the cost. This assumes that the log of cost is normally distributed around a mean value, which corresponds to the regression surface. In terms of the actual cost, this means that variation in cost is expected to be proportional to the expected (or mean) cost of the project. That distribution, which corresponds to a normal distribution in respect of the log of cost, is one whose mean is the project cost, and whose standard deviation is a fixed proportion of project cost. When the normal distribution is converted to raw cost for any project, it is a slightly positively skewed normal distribution, such that the peak of the probability density function is slightly less than the mean.

The skewed nature of this function could, it might be argued, be a better representation of the possible variation in project cost than a true (unskewed) probability distribution as generally there is more scope for the project costs to be much higher than expected rather than much lower.

A further property of this model is that it assumes that a change in any variable within the model will cause a proportionate change in cost.

Cost/m$^2$

The cost per m$^2$ is the cost predictor most used by quantity surveyors. The reason for this is that it provides a measure of cost that is essentially independent of building
size. If this value were to be used in a regression model, then it assumes that any variation in project cost is proportionate to the size of the building (rather than the cost). This may seem to be an unrealistic solution, as projects which are of a higher specification (and hence a higher cost/m$^2$) might be expected to show correspondingly higher variations in cost.

However, it has the added advantage of removing the understood linear part of the relationship between GIFA and project cost from the model. This should allow the modelling to focus on other, less understood influences on project cost.

The model also assumes that any changes in a variable will produce a fixed change in the cost/m$^2$ of the model. In other words, the change in cost associated with a change in a variable is proportional to the size of the building.

**Log of cost/m$^2$**

The log of the cost/m$^2$ makes the same assumptions as the log of cost that variations in project cost are proportionate to the expected cost. However, it also provides a variable which is devoid of the linear relationship between cost and size, in the same way as the cost/m$^2$ output. While this makes little difference to regression models, as will be shown later, it could be useful in the neural network modelling. The correlation of the log of GIFA with the log of cost/m$^2$ is much less than with the log of cost; this could stop the neural network being swamped with learning one relationship which is much more significant than any others.
Rejecting the raw cost as a predictor

While the arguments raised against the use of raw cost are valid ones, the inappropriate nature of the raw cost can be further demonstrated by comparing the results of a simple forward stepwise regression to those of the other three variables. However, it is important to establish a means of comparison which is representative of the predictive capacity of the model. The easiest way of comparing the models would be to compare the $R^2$ values for each model, each in terms of its own predictive variable. However, this would be invalid, as can be demonstrated by a simple example.

Consider a regression model where the cost $y$ is predicted by multiplying the gross internal floor area $g$ by a constant $b$, the mean cost/m$^2$ of all $n$ projects in the data set.

The matrix equation for this model is as follows.

$$Y = bG + E$$

Where $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ is the matrix of actual costs, $G = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{bmatrix}$ is the matrix of floor areas, and $E = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$ is the matrix of error not explained by the model.
The $R^2$ value can be determined as the proportion of the variation about the mean cost explained by the gross floor area to the total variation in the mean cost. The value for the data set under analysis is 0.78.

Equation 1 can then be translated into a cost/m$^2$ equation as follows:

Define $G^{-1}$ to be the left side inverse matrix of $G$:

$$
G^{-1}Y = bG^{-1}G + G^{-1}E
$$

$$
Y_2 = b_l + E_2
$$

Where $Y_2$ is the $1\times n$ matrix of cost/m$^2$, $y_i/g_i$, $l$ is the $1\times n$ matrix all of whose terms are 1, and $E_2$ is the $1\times n$ error matrix in the new model, such that the error terms are $e_i/g_i$.

It can be seen that the error term $E_2$ fully explains the variation in cost about the mean cost/m$^2$, $b$. Therefore the $R^2$ value for this model is 0, and the $R^2$ values of the two models are not the same. This does not, however, indicate that the performance of the cost model is superior to that of the cost/m$^2$ model. The two models are algebraically identical and therefore make the same predictions of project cost. What has changed between the two models is the variable by which those predictions are measured. Therefore, it is essential that $R^2$ values be compared in terms of the same variable.
In comparing the cost model, three measures are used: the $R^2$ on cost, the $R^2$ on cost/m², and the Mean Absolute Percentage Error (MAPE). The $R^2$ on cost value is the value of $R^2$ obtained when each model’s prediction is expressed as the raw cost of the project. Similarly, the $R^2$ on cost/m² is the value obtained when predictions are expressed as cost/m². The MAPE is the mean deviation from the actual value, expressed as a proportion of the project cost. Values for these three measures for the data set in question are shown in Table 2.

The value associated with cost appears to be very favourable in terms of $R^2$ on cost. However, the other two values show it not to be a very useful model. Its effectiveness at predicting the cost/m² is very low, and the average error in the predicted cost of a project is over 65%. The values of the other three models are much better than this and comparable with each other. Thus the hypothesis that the raw cost model will not, generally, predict the costs of projects very well has been shown to be true. Moreover, it is also possible to show that the first assumption, that the standard deviation of error is constant, is false by considering the scatter plot of error against cost, as shown in Figure 1.

As this figure clearly shows, the error in project cost rises with the project cost, suggesting that the error is either approximately proportional to the cost of the project or, because of the high correlation between cost and size, proportional to the size of...
the project. Therefore the raw project cost must be rejected as a suitable predictor for a regression model.

**Predictive models**

In order to create a predictive regression model, two methods were attempted. The first was a simple forward stepwise regression modelling technique. This involved including/excluding variables one step at a time. At each step all the variables in the model were checked for their values of $T$, and rejected from the model if the value of $T$ lead to a significance of greater than 0.1. Provided no variable was excluded from the model, the values of $T$ which each variable not in the model would have if it were to be included were evaluated. The variable whose value was greatest was added to the model, provided its significance was less than 0.05.

One problem with a forward stepwise regression is that a variable that correlates well with a number of cost significant variables may be included ahead of those variables, because it appears to encapsulate those variables. If this encapsulating variable has a higher significance than the individual variables themselves then the variable will be included first. When other variables are considered for addition to the model, some of the information contained in them will already be present in the model, which will make them appear less significant than they really are.

One possible way of circumventing this problem is to perform a backwards modelling technique. This technique begins with a regression model which contains all the variables. The values of $T$ are evaluated for each variable and the variable with the
lowest value is removed from the model, provided the value of T is not significant at the 95% confidence level. The process halts when all the variable remaining in the model have values of T which are significant at 95% confidence level.

Therefore both forwards and backwards modelling were performed. Three models were generated for each method, to predict cost/m², log of cost/m² and log of cost.

**Summary of models found**

A total of 6 regression models were developed. The number of variables in the model varied considerably. The smallest number of variables used was 8, in the forward stepwise log of cost model. The largest was 14, in both the log of cost and log of cost/m² backward models. Throughout the models a total of 19 different variables were used. However, there were two variables for which the both the logarithmic form and the unfactored form were used in different models: function and duration. The variables are shown in Table 3. They are ranked by the number of times they appear in the models, and then by the mean value of T for inclusion in those models. Note that function and duration are broken down into their log and raw values to allow comparison of the two representations. Note also that the variables which appear in only one or two models are sorted together as those that appear in two appear in the two log of cost backward models. These models are actually identical, as will be discussed.

<<<< Insert Table 3 about here >>>>
Before looking at the results in any detail, it is important to note that the log of the cost/m² and log of cost backward models are identical. The coefficient matrices are identical apart from the coefficient which corresponds to the log of the GIFA. The coefficient for the cost/m² model is –0.07, and the coefficient for the log of cost model 0.93.

Considering the following can show the fact that the two models are equal. In a linear regression model, if \([b_0, b_1, \ldots, b_n]\) are the coefficients for a model with n variables \([v_0, v_1, \ldots, v_n]\), then the output \(y\) is defined as follows:

\[
y = \sum_{k=0}^{n} b_k v_k + e
\]

where \(e\) is the error in the prediction. Therefore, if \(v_0\) is the log of GIFA, and \(y\) the log of the cost/m², then the equation becomes:

\[
\log(GIFA) = \sum_{k=0}^{n} b_k v_k + e
\]

\[
\log(cost) = b_0 v_0 + \log(GIFA) + \sum_{k=0}^{n} b_k v_k + e
\]

\[
\log(cost) = (b_0 + 1)v_0 + \sum_{k=0}^{n} b_k v_k + e
\]

Thus, if the coefficient matrix of the log of cost/m² model is identical to the log of cost model, with the exception that the coefficient corresponding to the log of GIFA is exactly 1.00 higher than the same coefficient in the log of the cost/m² model, then the costs predicted by both models will be the same. As this is the case, their properties will be identical. This is observed from the fact that the models contain the same variables, with the same coefficients and values of T (except for the log of GIFA), and the same variables are excluded from the model in the same order.
The forward models, however, do not show such a correlation. This is because the models are only identical when the log of GIFA is included as a variable. When GIFA is not included the models are not the same. In the log of cost model, there is a very strong linear relationship between log of cost and the log of GIFA. However, in the log of cost/m² model, the understood linear portion of the relationship has been accounted for by dividing the cost by the GIFA. For this reason the correlation of GIFA with the log of cost/m² is not as strong as it is with the log of cost. Thus, when the forward stepwise algorithm is employed, the log of GIFA is the first variable to be added to the log of cost model, whereas it is the 6th variable to be added to the log of cost/m² model. Up until the point where GIFA is added, the models are different, so different variables will be entered to the regression model than might have been had GIFA already been included.

However, when using the backwards technique with the log of cost/m² model, the initial value of T for the log of GIFA is 1.55, much higher than many other variables. After the first exclusion (the log of envelope), the value of T rises to 1.98, which is significant with 95% confidence. With subsequent exclusions, the value of T continues to rise, so the variable will not be excluded from the model. Thus, the log of cost and log of cost/m² models will both contain GIFA, and will therefore be identical.

**Significance of variables**

On the whole, the backwards selection techniques yielded models with more variables than the forward techniques. The cost/m² and both log of cost models had 11 and 14
variables respectively. This means that by using backwards selection it is possible to 
extract more significant variables than using forward selection. One possible 
éxplanation for this is that there are a number of variables that, while not necessarily 
exerting a significant influence on cost in themselves, do correlate well with a number 
of cost significant variables that do. Thus, if this variable is included, it is possible 
that the influence of a number of other cost significant variables is also implicitly 
taken into account by the inclusion of this variable.

This effect can be demonstrated by considering the representation of the building 
function variable. Building function correlates well with many variables relating to 
the internal specification of the building, as shown in Table 4

<<<Insert Table 4 about here >>>>

Building function implicitly dictates the general level of internal specification, as can 
be seen from the correlation coefficients in the table. The building function may also 
contain other information, which is not encoded into other cost significant variables. 
Building function correlates very highly with the cost/m² and the log of cost. This 
means that is will be included in a forward selection model. Therefore, when it comes 
to including the level of a finish or mechanical and electrical (M & E) specification, 
some of the data contained in these variables will be implicitly coded in the form of 
the building function variable. This will reduce the apparent influence of these 
variables, and hence their significance. Building function is probably the most 
extreme example of this type of variable, but nevertheless others may exist, and care
should be taken when modelling to ensure, as far as possible, that this effect does not occur.

**Significant Predictor Variables**

Five variables appear in all 6 models: GIFA, function, duration, mechanical installations and piling. This suggests that these are the key linear cost drivers in the data.

There were also four variables which appeared 5 times: internal wall finishes, frame, site access and protective installations. Protective installations was omitted from the log of cost forward selection model, and all the others were omitted from the cost/m² backwards selection model. The fact that these variables, which otherwise appear to be highly important, are omitted from a single model is worth closer attention.

Three of the variables are omitted from the backwards selection model of cost/m². However, this model also includes four other variables not included in any of the other models, namely: substructure, special installations, external walls and floor finishes. There are two possible explanations of this inclusion. The first is that these variables affect the cost as a proportion of the size of the building and they correlate highly with some of the variables in the forward regression model. The fact that these variables have already been included in the forward search models reduces the significance of these variables. The converse explanation is that the correlation between the variables included and those three variables excluded, and which are present in the other models, has caused them to be rejected.
The correlations between the 3 variables excluded and the 4 included are shown in Table 5. Correlations which are greater than 0.1 are shown in bold type.

The first thing to note is that internal wall finishes correlates with all of the variables apart from substructure. This suggests that if internal wall finishes were to be included in the model without these three variables, it would implicitly include some of the effects of these three variables. This could artificially inflate its value of T in a forwards model, causing it to be added before any of the other three variables. Then, the T values of these other 3 variables will be smaller, as part of their effect on cost will already be included in the model. This will not occur with the backwards model. These three variables are already included, so the effect of removing internal wall finishes is much less. Similarly, site access is also correlated with a number of variables. Therefore it is possible that the influence of this variable is overestimated.

The independent variable frame and substructure, however, only correlate strongly with each other. It could be argued that only one of these variables could be significant within the model, therefore, it is not possible to include both. The fact that the variable frame appears more consistently suggests that it is the more cost significant variable. Nevertheless, the value of T for substructure is 4.04, much higher than 3.27, the highest value obtained by the frame type in any of the models. Thus it might be argued that, for the cost/m² models, substructure is a significant variable.
The other variable only included in 5 models is protective installations. This was excluded from the log of cost forward selection model. However, the selection process for this model did not perform particularly well. It was only able to identify 8 predictor variables, which makes the resultant model the smallest generated. Therefore, it is possible to assert that it was the failure of the selection technique to identify more significant variables which caused this omission.

Performance

The six models all performed similarly. Their relative performances are shown in Table 6.

 The log of cost backwards models outperform the other models by most of the percentage error measures. However, the differences between all the models are small. Further, it must be pointed out that one of the reasons the log of cost and log of cost/m^2 models perform so well is that it is possible to find more variables whose inclusion in these models we can be confident of. Additionally, it is not necessarily the case that this is the best dependent variable to model.

However, the R^2 value on both cost and cost/m^2 is actually higher using the cost/m^2 model, despite having fewer variables than the two log of cost backwards models. Therefore, it would appear that the best two models are the log of cost backwards selection models and the cost/m^2 model.
However, given that the predictive capabilities of all the models are very close, it might be better to look at the spread of error. Neural networks will minimise error using the least squares approach. As this can be sensitive to non-uniformity of standard deviation of error and non-normality of error, the spread and normality of error is assessed for each model.

The spread of error by cost can be assessed by considering the scatter plots of error against the value of the independent variable. The scatter plot of actual cost/m² against error for the cost/m² forward election model is shown in Figure 2.

This plot shows that the regression model is failing to predict the costs of very expensive and very cheap projects with any degree of accuracy. The tendency is to underestimate the cost of very expensive projects, and overestimate the cost of very cheap projects. The $R^2$ value of 0.3452 indicates that nearly 35% of the error can be accounted for by this phenomenon.

In addition to this, the plot also shows that the error in cost/m² does not vary significantly for different values of the cost/m². This indicates that the error, and hence the standard deviation, does not rise significantly as the cost/m² of the project rises, unlike the trend which was observed with the raw cost. This suggests that the assertion that the error is proportional to the size of the building is a reasonable one, and cost/m² is a suitable output for a cost model.

The scatter plot of the backwards selection model follows a similar pattern trend, as do the scatter plots for the log of cost/m$^2$ forward and backward models and the log of cost forward and backward models. The values of $R^2$ for these scatter plots are given in Table 7.

All of these plots display the same tendency for the models to overestimate the cost of cheaper projects, and underestimate the cost of more expensive projects. The fact that as much as 35% of the error appears to arise from this suggests that some key drivers of building cost are not being properly represented. This either arises from the non-inclusion of these key drivers, or from non-linearities in the model.

The omission of key drivers from the data set can arise in two ways. The first, and perhaps simplest is that the key variables have simply not been collected as part of the data collection process. The other possibility is that the key drivers have been omitted through poor representation of these variables. Both of these situations are difficult to verify without further research, but, if these errors arise from non-linearities in the model, then it is expected that a better model will be generated using neural networks.

While the trend for the average error to rise with cost/m$^2$ is observed with the log of cost models, the scatter plots for the log of cost and cost/m$^2$ models do not display the same even spread of standard deviation. The spread in the error appears to rise with increasing cost/m$^2$. This is interesting because it means that, concerning the
assumption that the variation in the error is proportional to the size of the model, the observed scatter plot of the resultant model appears to show this trend. However, with regard to the assumption that the error is proportional to the cost of the project, the error in the error by cost/m² plot appears to increase with increasing cost/m². Thus the scatter plots appear to follow the assumptions made in the selection of the dependent variable. This can be seen by considering the scatter plots of error against the dependent variable; an example is shown in Figure 3 for log of cost/m² forward model.

<<<< Insert Figure 3 about here >>>>

This plot also shows same prediction tendency for very low cost or high cost projects as is observed in the cost/m² plots. The R² value of the trend line is also similar to that found in the cost/m² scatter plots. This indicates that the proportion of error associated with this phenomenon appears to be the same for both log of cost and the cost/m² models. It also shows that there little evidence for the standard deviation varying as projects become more expensive. These two observations may also be made for the backwards model.

The error of scatter of the total cost is actually testing a slightly different assertion to the log of cost/m² models. This assertion is that the standard deviation of costs remains constant as project cost increases, rather than as the cost/m² increases. This can be evaluated from the scatter plots from the remaining two models, which suggests that there might be some variation in the standard deviation of error as project cost increases. There appears to be much spread in the some of the least
expensive buildings, and a lower amount of spread. However, the effect is not very pronounced so it would be difficult to draw any conclusions from this that might apply to the data set as a whole. The plot for the backwards model appears to show similar tendencies to the forwards model. Note also that both of these models show a slight tendency to underestimate the cost of larger projects. However, this can probably be explained by the correlation between cost and cost/m$^2$. This would cause the tendency observed by the cost/m$^2$ to appear in the total cost model.

The values of $R^2$ for these scatter plots are given in Table 8.

Overall, from the scatter plots, it would appear that all three variables are suitable for the modelling of the project cost. The standard deviations of error show the even spread in terms of the dependent variable. In reality, the error cannot be completely proportional to one or the other of the floor area or the cost. Nevertheless, the error is sufficiently proportional to both to allow the three predictors to be used to construct reasonable models by minimising least squares.

In addition to the scatter plots, it is also useful to assess how normal the error distributions associated with each model are by considering their skewness and kurtosis. The values of skewness and kurtosis are shown in Table 9.

The results show that all the distributions tend to be more peaked than the normal distribution. However, the log of cost measures appear to be closest to being normally distributed, both in terms of skew and kurtosis. This could allow the log of cost measures to generally produce better models than the raw cost/m² models. However, the skewness and kurtosis of the errors for all the models are relatively small. Therefore the influence of the non-normality of the distributions of error is likely to be small. In the light of this it is recommended that both log of cost/m² and cost/m² models be tested. The reason why the log of cost/m² model is preferred is that the strong linear relationship between GIFA and cost is removed. This relationship is well documented, recognised, and used. Therefore, if it is removed then the neural network does not have to learn this relationship, but is only required to learn those relationships which are not yet understood.

**Testing ordinality**

As well as generating the general performance measures and the significance of variables, it is also possible to test the values of the predictors represented by ordinal variables. The values of the ordinal variables were determined initially by establishing how much the solutions they represented usually cost in relation to each other. For example, the cost of different staircases was evaluated in order to try to determine the relative costs of different stair solutions. This was then compared to the analyses of variance to see whether the differences in cost between projects of different stair solutions suggested the same order. However, this value was subject to the possible influences of other variables with which the stair solution correlated, so the results could not be depended on.
It is also possible to make judgements about the order of the categories with these variables using the regression models. If it is assumed that the model contains all the significant influences of cost, then the most appropriate order of the variables (i.e. the one which best expresses the relative influence of each category on the cost) can be determined by finding the order which yields the highest value of $T$ for inclusion in the model.

In addition to this method, it is also possible to perform an analysis of variances between the different groups on the residual cost error in the model. In order to do this the variable under examination must be excluded from the regression model. While the fact that the influences of the significant cost variables have been taken into account is useful, it must be understood that there are shortcomings of this test:

1. The order found is the correct order given the influences of the variable on cost which are already taken into account by correlations with other cost significant variables. The real order of influence is not necessarily the same.

2. The order found may include effects that arise from correlations with other variables which are not included in the model. While the cost influences of these may be small, if the variable correlates with a number of these then they could affect the validity of the order found.

**Procurement route**

Perhaps the most relevant variable to test the ordinality of is the procurement route. The route was ranked initially in terms of the average cost/m$^2$ of each of the three
groups. However, despite being significant by the analyses of variance on the various cost measures, the procurement route has not featured in any of the regression models developed. Therefore, the values of T for exclusion from the final model are compared in Table 10, along with some other key statistics.

<<< Insert Table 10 about here >>>>

The excluded after column indicates the last model procurement was included in. Model 1 contained all variables, model 2 all except one and so on until the final model is arrived at. For the first original combination this was 38 models, but for the next two it was 31. This, it seemed, was caused by the fact that the earlier exclusion of the procurement route changed the relative values of significance of other variables, causing different variables to be rejected by the selection process.

The resultant model contained 19 variables, 9 of which were among the 12 variables in the original model.

However, the performance of this new, larger model was inferior to the original model. The $R^2$ value was 0.693, while the value for the original model was 0.666. However, the values of MAPE, as well as the upper and lower confidence bounds, were 25.5%, 27.8% and –66.8% respectively. The values in the original model were 21.7%, 33.6% and –61.7%. Thus, the MAPE and range of 90% confidence are better for the original model.

On the basis that the models are approximately comparable, the values of T can be compared. This would lead to the conclusion that the order is correct when management is between traditional and design & build. However, although the T values suggest that this variable is more significant, it should also be noted that the procurement route was excluded after model 30 for this order, while for the original order, it was excluded after model 36. If the original order is truly less significant, then it would be excluded earlier, not later. As it is, there appear to be more variables which are less significant than the procurement route with the original order. So why do the values of T not correspond with this conclusion?

The reason is that as more and more variables are excluded from the model, the value of T for procurement actually tends to decrease. Thus, the value of T at model 31 is actually higher than the value at model 37. Therefore, if it is assumed that the fairest comparison of the two values of T will be obtained with the same number of variables have been excluded from the model, then the value of T obtained from the original order is the best, suggesting that this is actually the best measure.

The process was repeated for backwards regression on the log of cost/m². The results are more conclusive, and are shown in Table 11.

The final model arrived at was the same for all three orders, unlike the cost/m² models. Thus the comparisons of the values of T for the final models are valid.
These results show clearly that the original order is correct, which corresponds with the result of the cost/m² models.

Based on this, it must therefore be concluded that the order design & build, traditional, management is the most appropriate order for representation of this variable. However, it must also be recognised that this relationship is not a significant one for the cost models. Therefore it is not possible to draw any conclusions about the cost of procurement route from this result. The significant difference between the procurement routes observed when using an analysis of variances can be explained in terms of the influence (on the cost) of other cost significant variables with which this variable correlates. In other words, the apparent difference in cost between the procurement routes might only be due to typical levels of specification associated with each procurement route.

**Comparison with neural network models**

The results of the regression analysis were used to inform the development of neural network models. Models were developed using the five variables which appeared in all six regression models, the nine variables which appeared in five of the models and all the variables. In order to improve performance and eliminate bias, the ‘voting system’ technique was also adopted, which involves the creation of a number of models each using different training, verification and test sets with the output being taken as the average of these models; in this case 100 networks were used.
The best regression model is the log of cost backward model which gives an $R^2$ of 0.661 and a MAPE of 19.3%. The best neural network model is one which uses all 41 variables and a voting system using 100 networks; this gives an $R^2$ value of 0.789 and a MAPE of 16.6%. Although this is better than the regression model, other neural network models showed a comparable or poorer performance when compared to the regression model. For example, the nine variable model (without using the voting system) had an $R^2$ value of 0.688 and a MAPE of 20.1%, which is very similar to the best regression model.

However, the models developed using both techniques compare favourably with past research which has shown that traditional methods of cost estimation are less accurate as evidenced by reported values of MAPE between 20.8% (Skitmore et al., 1990) and 27.9% (Lowe, 1996).

**Conclusions**

Using raw cost as the dependent model variable will produce models which are less accurate than those which use other more appropriate representative variables and, in any event, the assumptions implicit in using raw cost as the dependent variable are open to question. Other more reliable and useful dependent variables are cost/m², log of cost and log of cost/m² and these were used to produce a total of 6 models, using forwards and backwards stepwise regression for each dependent variable.

The number of independent output variables included in each model varied between 8 and 14, with more variables consistently being included in the backwards models.
number of variables include in each model is also a reflection of the correlation which exists between variables, such that when a particular variable is included another variable which correlates well with that variable will not be included. The five variables included in all six models were: GIFA, function, duration, mechanical installations and piling, suggesting that they are the key linear cost drivers in the data.

All six models performed similarly and while the best regression model can claim to be the log of cost backward model, which gives an $R^2$ of 0.661 and a MAPE of 19.3%, the difference between the performance of all models is small. One of the reasons that the log of cost backwards model performed well is that it contains more variables whose inclusion in the model we can be confident of and it does not necessarily mean that this is the best variable to use as the dependent variable.

Scatter plots of cost/m$^2$ against error show that all models tend to underestimate the costs of very expensive projects and overestimate the cost of very cheap projects. While a test of ordinality for the procurement route variable showed that the correct order, in terms of cost, for the three options was: design & build, traditional, management. However, this does not mean that, for example, design & build will always be cheaper than traditional procurement methods, but gives an order for these variables to be used when data are modelled in the future.

Performance of the regression model is slightly inferior to the neural network models, but the differences are small. Although, there is potential for better models to be developed, the models produced by both techniques compare favourably with traditional methods of cost estimation.

Acknowledgements

The authors gratefully acknowledge the support of the Engineering and Physical Sciences Research Council (EPSRC), who funded the research through two grants; our industrial collaborators: Paul Moore - EC Harris, the late Chris Powell – formerly with Tweeds (then Faithful and Gould), Alun Williams – Symonds, and Joe Martin – Building Cost Information Service (BCIS); and the contribution made by the research assistants: Mick Gregory and Adam Hickson and by Dr Roy Duff, former senior lecturer, Manchester Centre for Civil and Construction Engineering, UMIST.

References


Birnie, J.W. 1993: A behavioural study using decision analysis of building cost prediction by chartered quantity surveyors PhD thesis, University of Ulster


Ellis, C. and Turner, A. 1986: *Procurement Problems* Chartered Quantity Surveyor, April, p 11


McCaffer, R. 1975: Some examples of the use of regression analysis as an estimating tool. Quantity Surveyor, December 81-86


Ogunlana, S.O. and Thorpe, A. 1987: Design phase cost estimating: the state of the art
International Journal of Construction Management and Technology, 2(4), 34-47


Smith, N.J. 1995: Purposes and stages of estimating in "Project cost estimating"
Edited by N.J. Smith, Thomas Telford Limited, London


Table 1  Classification of input variables

Project strategic variables

<table>
<thead>
<tr>
<th>Contract form</th>
<th>Procurement strategy</th>
<th>Tendering strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>Purpose</td>
<td></td>
</tr>
</tbody>
</table>

Site related variables

<table>
<thead>
<tr>
<th>Site access</th>
<th>Type of location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topography</td>
<td>Type of site</td>
</tr>
</tbody>
</table>

Design related variables

<table>
<thead>
<tr>
<th>Air conditioning</th>
<th>Internal doors</th>
<th>Roof finishes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceiling finishes</td>
<td>Internal walls</td>
<td>Roof profile</td>
</tr>
<tr>
<td>Electrical installations</td>
<td>Internal wall finishes</td>
<td>Shape complexity</td>
</tr>
<tr>
<td>Envelope</td>
<td>No. lifts</td>
<td>Special installations</td>
</tr>
<tr>
<td>External doors</td>
<td>Storeys above ground</td>
<td>Stairs</td>
</tr>
<tr>
<td>External walls</td>
<td>Storeys below ground</td>
<td>Substructure</td>
</tr>
<tr>
<td>Floor finishes</td>
<td>Mechanical installations</td>
<td>Structural units</td>
</tr>
<tr>
<td>Frame</td>
<td>Piling</td>
<td>Upper floors</td>
</tr>
<tr>
<td>Function</td>
<td>Protective installations</td>
<td>Wall-to-floor ratio</td>
</tr>
<tr>
<td>GIFA</td>
<td>Quality*</td>
<td>Windows</td>
</tr>
<tr>
<td>Height</td>
<td>Roof construction</td>
<td></td>
</tr>
</tbody>
</table>

* Note that in previous publications *Quality* has been classified as a project strategic variable. Here it has been re-classified as a design related variable.
Table 2  Measures of the performance of various cost predictors in a regression model

<table>
<thead>
<tr>
<th></th>
<th>Cost</th>
<th>Cost/m$^2$</th>
<th>Log Cost/m$^2$</th>
<th>Log Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>R$^2$ on cost</td>
<td>94.1%</td>
<td>94.4%</td>
<td>92.4%</td>
<td>92.7%</td>
</tr>
<tr>
<td>R$^2$ on Cost/m$^2$</td>
<td>9.3%</td>
<td>66.8%</td>
<td>64.8%</td>
<td>64.4%</td>
</tr>
<tr>
<td>MAPE</td>
<td>65.3%</td>
<td>20.8%</td>
<td>20.0%</td>
<td>20.1%</td>
</tr>
</tbody>
</table>
Table 3  Variables included in the regression models

<table>
<thead>
<tr>
<th>Variables</th>
<th>Cost/m²</th>
<th>Log Cost/m²</th>
<th>Log Cost</th>
<th>Total</th>
<th>Mean T</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIFA (log)</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>6</td>
</tr>
<tr>
<td>Function (log)</td>
<td></td>
<td></td>
<td>•</td>
<td>•</td>
<td>2</td>
</tr>
<tr>
<td>Duration (log)</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>4</td>
</tr>
<tr>
<td>Mechanical installations</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>6</td>
</tr>
<tr>
<td>Piling</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>6</td>
</tr>
<tr>
<td>Internal wall finishes</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>5</td>
</tr>
<tr>
<td>Frame</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>5</td>
</tr>
<tr>
<td>Site access</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>5</td>
</tr>
<tr>
<td>Protective installations</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>5</td>
</tr>
<tr>
<td>Internal walls</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>3</td>
</tr>
<tr>
<td>Substructure</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>1</td>
</tr>
<tr>
<td>Wall/floor ratio</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>2</td>
</tr>
<tr>
<td>Special installations</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>1</td>
</tr>
<tr>
<td>External walls</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>1</td>
</tr>
<tr>
<td>Floor finishes</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>1</td>
</tr>
<tr>
<td>Height (log)</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>2</td>
</tr>
<tr>
<td>Units</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>2</td>
</tr>
<tr>
<td>Electrical installations</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>11</td>
<td>9</td>
<td>14</td>
<td>8</td>
</tr>
</tbody>
</table>
Table 4  Correlation between building function and other variables

<table>
<thead>
<tr>
<th>Correlation with building function</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Floor finishes</td>
<td>0.460</td>
</tr>
<tr>
<td>Ceiling finishes</td>
<td>0.458</td>
</tr>
<tr>
<td>Mechanical installations</td>
<td>0.455</td>
</tr>
<tr>
<td>Internal wall finishes</td>
<td>0.330</td>
</tr>
<tr>
<td>Air conditioning</td>
<td>0.326</td>
</tr>
<tr>
<td>Quality</td>
<td>0.325</td>
</tr>
<tr>
<td>Windows</td>
<td>0.315</td>
</tr>
<tr>
<td>Shape</td>
<td>0.279</td>
</tr>
<tr>
<td>Protective installations</td>
<td>0.252</td>
</tr>
<tr>
<td>Special installations</td>
<td>0.204</td>
</tr>
<tr>
<td>Electrical installations</td>
<td>0.154</td>
</tr>
</tbody>
</table>
Table 5  Correlations between excluded and included variables

<table>
<thead>
<tr>
<th></th>
<th>Frame</th>
<th>Internal wall finishes</th>
<th>Site access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substructure</td>
<td>0.428</td>
<td>0.060</td>
<td>0.070</td>
</tr>
<tr>
<td>External walls</td>
<td>0.100</td>
<td>0.246</td>
<td>0.198</td>
</tr>
<tr>
<td>Special installations</td>
<td>0.070</td>
<td>0.155</td>
<td>0.076</td>
</tr>
<tr>
<td>Floor finishes</td>
<td>-0.003</td>
<td>0.479</td>
<td>0.181</td>
</tr>
</tbody>
</table>
Table 6  Performance of regression models

<table>
<thead>
<tr>
<th></th>
<th>Forward</th>
<th>Backward</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cost/m²</td>
<td>Log Cost/m²</td>
</tr>
<tr>
<td>$R^2$ on cost/m²</td>
<td>66.8%</td>
<td>64.8%</td>
</tr>
<tr>
<td>$R^2$ on cost</td>
<td>94.4%</td>
<td>92.4%</td>
</tr>
<tr>
<td>MAPE</td>
<td>20.8%</td>
<td>20.0%</td>
</tr>
<tr>
<td>Mean % error</td>
<td>6.2%</td>
<td>3.4%</td>
</tr>
<tr>
<td>St dev. % error</td>
<td>30.5%</td>
<td>28.3%</td>
</tr>
<tr>
<td>Lowest outlier</td>
<td>-59%</td>
<td>-54%</td>
</tr>
<tr>
<td>Highest outlier</td>
<td>253%</td>
<td>208%</td>
</tr>
<tr>
<td>90% confidence</td>
<td>-36.9%</td>
<td>-35.5%</td>
</tr>
<tr>
<td></td>
<td>50.3%</td>
<td>44.9%</td>
</tr>
</tbody>
</table>
Table 7  \( R^2 \) for scatter plots of actual cost/m\(^2\) against error

<table>
<thead>
<tr>
<th></th>
<th>Cost/m(^2)</th>
<th>Log Cost/m(^2)</th>
<th>Log Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward</td>
<td>0.3452</td>
<td>0.3089</td>
<td>0.2855</td>
</tr>
<tr>
<td>Backward</td>
<td>0.3163</td>
<td>0.2367</td>
<td>0.2367</td>
</tr>
</tbody>
</table>
Table 8  R\textsuperscript{2} for scatter plots of actual cost/m\textsuperscript{2} against error

<table>
<thead>
<tr>
<th></th>
<th>Log Cost/m\textsuperscript{2}</th>
<th>Log Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward</td>
<td>0.2777</td>
<td>0.0313</td>
</tr>
<tr>
<td>Backward</td>
<td>0.2503</td>
<td>0.0354</td>
</tr>
</tbody>
</table>
**Table 9** Skewness and kurtosis of the error distributions

<table>
<thead>
<tr>
<th></th>
<th>Error values forward</th>
<th></th>
<th>Error values backward</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cost/m²</td>
<td>Log Cost/m²</td>
<td>Log Cost</td>
<td>Cost/m²</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.887</td>
<td>-0.117</td>
<td>-0.155</td>
<td>0.377</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>2.340</td>
<td>1.505</td>
<td>1.769</td>
<td>2.446</td>
</tr>
</tbody>
</table>
Table 10  Statistics for test of ordinality for procurement route for cost/m² forward

<table>
<thead>
<tr>
<th>Excluded after</th>
<th>36</th>
<th>30</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>21.6</td>
<td>11.9</td>
<td>-21.3</td>
</tr>
<tr>
<td>T</td>
<td>1.59</td>
<td>1.84</td>
<td>1.74</td>
</tr>
<tr>
<td>Sig</td>
<td>0.113</td>
<td>0.067</td>
<td>0.083</td>
</tr>
<tr>
<td>T at 31</td>
<td>1.92</td>
<td>1.84</td>
<td>1.74</td>
</tr>
<tr>
<td>Design and build</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Management</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Traditional</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 11  Statistics for test of ordinality for procurement route for log of cost/m² backward

<table>
<thead>
<tr>
<th>Excluded after</th>
<th>20</th>
<th>15</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.036</td>
<td>0.014</td>
<td>-0.009</td>
</tr>
<tr>
<td>B (approx Prop)</td>
<td>3.6%</td>
<td>1.4%</td>
<td>-0.9%</td>
</tr>
<tr>
<td>T</td>
<td>0.80</td>
<td>0.38</td>
<td>0.07</td>
</tr>
<tr>
<td>Sig</td>
<td>0.427</td>
<td>0.704</td>
<td>0.942</td>
</tr>
<tr>
<td>Design and build</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Traditional</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Management</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Figure 1  Scatter plot of error against actual cost

Scatter plot of Error against actual cost
Figure 2 Scatter plot of actual cost/m² against error for cost/m² forward selection model

\[ R^2 = 0.3368 \]
Figure 3  Scatter plot of error by log of cost/m², forward selection model

$R^2 = 0.3064$

Log of cost/m²

Error