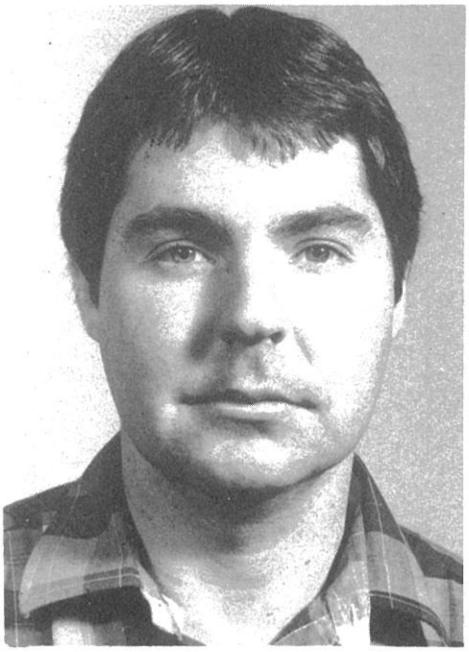
Problems in econometric cost modelling

P. A. Bowen, BSc (QS), MSc, MAQS, RQS, ACI Arb
Lecturer, Department of Quantity Surveying, University of Natal, Durban



P. A. Bowen gained his BSc (QS) with distinction in 1975 and an MSc in Construction Management at Heriot-Watt University in 1980. He has practised on his own account since 1981 and is mainly interested in mathematical cost modelling and Building Economics.

Regression cost modelling, while having considerable potential as an aid to the quantity surveyor, is a complex statistical tool and the problems associated with using it should not be underestimated. This paper highlights some of the more fundamental problem areas.

Introduction

Much has been written recently in quantity surveying journals advocating the use of multiple regression cost models. While in agreement with the researchers concerned regarding the merits of regression cost modelling, the writer believes that a somewhat oversimplified picture has been painted and that prospective users of such models need to be made aware of some of the fundamental yet important problems associated with regression analysis. Firstly, however, mention needs to be made of the use of the term "cost modelling". Cost modelling, as its name suggests, is the process of modelling (or simulating) cost as a function of certain factors. Regression analysis is a statistical tool for analysing multifactor data and is only one of the many methods of cost modelling. Thus, the terms "cost modelling" and "regression cost modelling" are not synonymous. Other cost modelling techniques include, *inter alia*, Monte Carlo simulation, simulation, optimisation using cost equations, and even unit rate estimates.

Ashworth (1981) describes regression analysis as the most popular, useful and applicable cost modelling technique. This statement is, in the writer's opinion, misleading in that it does not include the word "dangerous".

The purpose of this paper is to highlight some of the more common problems associated with regression analysis and is not intended to be a definitive list.

The Effect of Time

Any cost information (or any relationship that is likely to change over time) will be based on different points in time and therefore possibly different economic conditions. This factor needs to be taken into account when model building. There are two basic methods of catering for the influence of time.

(i) All the time dependent data are updated to a common base date (known as the model's base date) using an appropriate index (e.g., a building cost index). This then means that the model is "static" in time and, when estimating at some future point in time, the derived cost (using the model) must be updated to the time of estimate using the same index.

The choice of index is important in that it should correctly reflect the "movement" of the dependent variable over time.

A formula to update such a model is as follows:

$$\text{Estimated Total Cost} = \text{Derived Cost} \times \frac{\text{Index at time of estimate}}{\text{Index at model's base date}}$$

(ii) The alternative is to include, as one of the independent variables, some measure of time e.g., the building cost index applicable at the time of tender of the project concerned. Thus, when using the model, the index applicable at the time of estimating is inserted as one of the measured independent variables. Here the choice of the appropriate independent variable would be critical. A measure of "time" itself could also be used.

Experience within the Department of Quantity Surveying at the University of Natal has shown that the method outlined in (i) above is preferable in that the latter does not seem to allow for the effect of time adequately. Further, problems are likely to arise if the "time" variable is discarded by the computer as being statistically insignificant.

Ashworth, Neale and Trimble (1980) tried to overcome this problem by using data relating to man-hours as a surrogate for money cost. In the writer's opinion this creates another problem in that the derived model does not take into account changing productivity. A further factor requiring attention is the capital intensiveness of the firm concerned.

In today's climate of double-digit inflation a model that fails to allow adequately for the influence of time on money is obviously suspect and model builders should be aware of this pitfall.

Sample Size and Data Collection

Like most forms of estimating, regression cost models rely totally on a sufficient quantity of representative data. The first important issue here is the question of sample size, the importance of which is stressed by Beeston (1978). Researchers seem to vary in their opinion as to the number of observations required per independent variable in the equation. According to Ashworth (1981), $2\frac{1}{2}$ times the number of variables should equal the number of sets of data required. Thus, a model containing twelve variables would require thirty sets of data for analysis purposes. McCaffer (1975) appears to support this contention.

In the opinion of the writer, however, a good "rule of thumb" is 30 observations per indepent variable in the equation, i.e., 360 sets of data for a 12 variable model, especially where "normality" is being approximated. Linked to the problem of sample size is the question of the nature of the sample. Regression analysis is most applicable where the data are of a homogeneous nature, e.g., all the projects analysed would have to be of a similar type and function. This poses a problem in that the very requirements of the model building process largely preclude the building of a model—especially to private quantity surveying practices. It would be very unwise to use a cost model based on residential flats for estimating the cost of (say) warehouses.

Multicollinearity

Where some of the independent (explanatory variables) are highly correlated, a situation called multicollinearity exists. Interpretation of the multiple regression equation depends implicitly on the assumption that the explanatory variables are not strongly interrelated (Chatterjee and Price, 1977).

A simple example of multicollinearity would be where floor area and perimeter length of a typical floor were both included in the model as explanatory variables. Obviously area and perimeter length are likely to be highly correlated.

The writer, in the development of a regression cost model for framed structures, found that a number of interrelations existed (Bowen, 1980) viz.

Variables	Correlation
Building height with number of floors	0.961
Area with number of lifts	0.871
Area with number of stairs	0.846
Area with perimeter length	0.763
Type of slab with slab thickness	0.734

An examination of the residuals indicated the presence of a slight downward trend, distorting the model.

It is recommended that one should be extremely cautious about inferences based on a regression analysis in the presence of multicollinearity. One indication of multicollinearity is a high R^2 value (coefficient of determination) and all "t" values small.

Various solutions are available when multicollinearity is detected. Maddala (1977) suggests six different approaches:

1. Dropping variables
2. Using extraneous estimates
3. Ridge regression
4. Using ratios or first differences
5. Using principal components
6. Getting more data.

The whole question of multicollinearity is a complex one and detailed discussion of this problem is beyond the scope of this paper. Suffice to say that it needs to be given cognizance in any analysis.

Hetroscedasticity

Violation of the assumption that the residuals have a common variance is known as hetroscedasticity. This situation occurs where the standard deviation of the residuals tends to increase as the value of the explanatory variable increases i.e., there exist discernable trends in the residuals.

The writer, as part of the framed structure study previously mentioned, developed a cost equation based on area alone (Bowen, 1980). An examination of the residuals relating to this model indicated that hetroscedasticity existed. The coefficient of variation is 49.64%, indicating a wide residual scatter. This situation may be illustrated as follows:

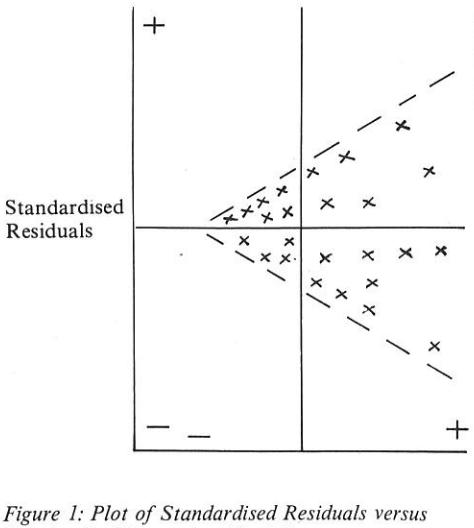


Figure 1: Plot of Standardised Residuals versus Standardised Dependent Variable

Such a situation is undesirable in that the model is inefficient and becomes unreliable. Hetroscedasticity may be removed by means of a suitable transformation.

Transformations

One of the standard assumptions in regression analysis is that the model which describes the data is linear. However, in spite of the fact that the regression coefficients may be significant and a high value of R^2 present, the linear model may not be appropriate.

A plot of the raw data may show departure from linearity for certain value ranges of the independent variables. This non linearity is likely to become more obvious when the distribution of the residuals is examined i.e., a distinct pattern is likely to be present.

Despite the presence of non-linearity, it is sometimes possible to restate the relationship in a linear form by transforming the original variables. An example of such a transformation is a logarithmic transformation on variables displaying exponential characteristics. The point to be made is that model builders should not automatically assume that a linear model is the most appropriate.

Transformations on variables can also be used to improve a model by way of accounting for more of the unexplained variation than before. Ashworth, Neale and Trimble (1980) use this technique in their build-up of the model used to predict the total number of man-hours required to complete the brickwork on a project. Buchanan (1972) uses much the same approach in his derivation of a cost model for the reinforced concrete frame of a building.

However, in the writer's opinion, this can be a dangerous practice in that it tends to "force-fit" a model to the data, remembering that it is possible to improve most models by the inclusion of product and other derived variables. Surely, where causal relationships are being investigated such procedures are of little benefit and should be avoided? It would, perhaps, be better to hypothesize a theoretical model initially, test the model and draw conclusions from the analysis. Where derived variables and transformations are to be used, these manipulations need to be justified on theoretical grounds before inclusion in the model. Where this procedure is not followed, spurious relationships are likely to result. Furthermore, users of the model are likely to be more inclined to be confident of the model if they understand the relationships therein.

Model Maintenance

As previously mentioned, the regression model is generally static in time and is derived from a certain specific set of data. Such data are a function of construction relationships and do not remain static over time and, consequently, the regression cost model needs to be updated or modified from time to time.

It is thus necessary to update cost models periodically by including the most recent data available and discarding the oldest. The further into the future one predicts, the less reliable the model is likely to be. Ideally the updating should be done annually. Naturally model maintenance will place a burden on individual firms and this, together with the sample size requirement, is a strong argument in favour of a centralised data bank.

Extrapolation

In establishing a regression equation, a set of observations is used that covers a limited range of values for the variables. Caution must be exercised when making predictions about the dependent variable when the independent variables fall outside this range. Such predictions are called extrapolations. Regression analysis is best limited to the range of actual observations.

If extrapolations are made outside the intervals of the data, the model builder runs the risk of applying the derived model to a situation for which it is ill-defined. This point is best illustrated as follows:

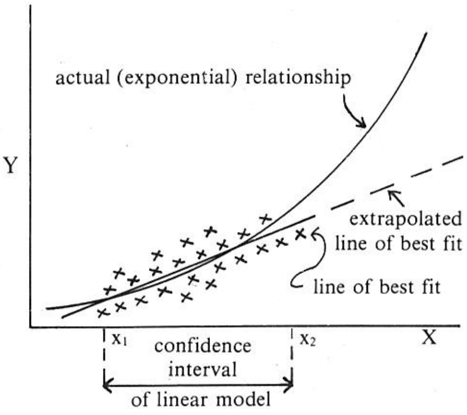


Figure 2: Extrapolation outside the Data Range

A linear model is fitted to raw data pertaining to a certain data range. The coefficient of determination (R^2) is acceptable and the residuals display no discernible trends.

If one applies the model to data beyond the x_2 limitation, the model becomes ill-defined and accuracy will suffer correspondingly. Thus it is wise to develop models based on as wide a data range as possible.

Accuracy

Naturally a regression cost model is only as useful as it is accurate and reliable. It is all very well to test the final model on new "un-seen" data but it is important to remember that this data can provide a misleading judgement of the model in that the data selected for testing the model need not be representative. Assuming that accuracy to within 5% of "actual" cost is achieved on four test cases, what does this really tell this researcher?

Surely such results are totally dependent on the test sample chosen and could just as easily have "gone the other way". In the writer's opinion, the best method of evaluating a model is to examine the residuals and the relevant statistics, operate the model in conjunction (parallel) with more conventional and traditional estimating techniques over a period of time and then decide whether or not the model is accurate enough for practical usage.

Acceptability by the Profession

Last, but by no means least, is the question of acceptability of regression cost models by the profession. As part of a research project dealing with regression cost models, the writer conducted a survey amongst thirty quantity surveyors in the Edinburgh area regarding, *inter alia,* their willingness to use regression cost models for practical purposes (Bowen, 1980). Although the sample size precludes definitive conclusions being arrived at, the writer considers that certain inferences can be drawn from the survey.

Only one of the surveyors questioned indicated that he had previously used regression cost models. Over two-thirds stated that they would be prepared to use such cost models should they be made available to them, albeit in many instances this willingness was qualified by the proviso that the model be of proven accuracy and reliability.

The writer is, however, sceptical of his results in the light of the research report published by the Department of Construction Management, Reading University (1980) entitled "Construction Cost Data Base". Here it was found that the profession has a clear preference for using in-house cost data for estimating purposes. Even within a particular office, individual surveyors have a very strong preference for using data generated on projects with which they have been personally involved. Only when this source fails to produce the necessary cost data does the surveyor begin to consider alternatives.

In the light of these findings it is unlikely that surveyors would be willing to use regression cost models, a relatively new and untried estimating technique.

Conclusion

Regression cost modelling must surely be considered an exciting development, a logical step forward of the quantity surveyor in the costing process, albeit its application to the construction industry is, as yet, fairly new.

Despite its obvious potential as an estimating tool, regression cost modelling needs to be undertaken with caution, the modeller being aware of the many pitfalls associated with this technique. It is all too easy, using one of the many regression packages available, to input the raw data, perform the regression analysis and accept the output without careful consideration of the many associated problems, some of which are discussed in this article. Regression analysis can, in the writer's opinion, be learnt only by experience and a surveyor would be well advised to enlist the aid of a statistician when embarking on the development of a model.

An appropriate quotation comes to mind; "There are lies, damned lies, and statistics".
Benjamin Disraeli.

References

Ashworth, A., 1981; "Cost Modelling for the Construction Industry", *The Quantity Surveyor,* July, pp. 132-134.

Beeston, D., 1978; "Cost Models", *Chartered Surveyor,* Building and Quantity Surveying Quarterly, Vol. 5 No. 5, pp. 56-59.

Bowen, P. A., 1980; An Investigation into the Feasibility of Producing an Econometric Cost Model for Framed Structures, MSc Project, Heriot-Watt University, Edinburgh.

Buchanan, J. S., 1972; "Cost Models for Estimating", Special Report, The Royal Institution of Chartered Surveyors.

Chatterjee, S., and B. Price, 1977; *Regression Analysis by Example,* John Wiley and Sons Incorporated, New York.

Department of Construction Management, Reading University, 1980; *Construction Cost Data Base,* A report prepared on behalf of the Directorate of Quantity Surveying Services, Department of the Environment (Property Services Agency).

Maddala, G. S., 1977; *Econometrics,* McGraw-Hill Incorporated, Tokyo.

McCaffer, R., 1975; "Some Examples of the use of Regression Analysis as an Estimating Tool", *The Quantity Surveyor,* December, pp. 81-86.

Trimble, E. G., 1974; "Regression Analysis—New uses for an Established Technique", Paper presented at the Conference entitled "The Training of Estimators", Department of Civil Engineering, Loughborough University of Technology, March.

INPLAN AWARD 1982

As no competitor met the full requirements of the 1982 competition, the jury reluctantly agreed unanimously that the Inplan Trophy—the premier award—could not be presented this year.

In these circumstances, and because of the two different approaches to the subject, two second prizes (joint) were awarded to:

Mr Rob Peebles, a sixth year post graduate student reading for a Diploma in Architecture at the School of Architecture, Leicester Polytechnic.

and

Mr Mark A. Preskey, a fourth year student in the Department of Building reading Construction at the Sheffield City Polytechnic.

The object of the Inplan Award competition is to encourage building students to put forward development ideas which will sustain interest in the use of rigid polyurethane-based products as effective insulants throughout the construction industry.

The Inplan jury operated entirely independently of the sponsors and was made up of selected representatives of the learned Institutes under the chairmanship of Professor D. R. Harper (Professor Emeritus of Building, UMIST), CBE. BArch, PhD (Tech), FRIBA, MRTPI, FCIOB, FCIArb. The other members of the jury were: Professor A. W. Pratt, DSc, FInstP, Head of Department of Construction & Environmental Health, University of Aston in Birmingham; Col D. Sherret, MC, MArch, RIBA, FCIOB, FBIM, Deputy Chief Executive, The Chartered Institute of Building; Mr W. Challenger, AIAS, Chairman of the Manchester and District Branch, representing the Incorporated Association of Architects and Surveyors; Mr M. R. Edwards, FIQS, AMCST, FCIArb, representing the Institute of Quantity Surveyors; and Mr P. Galloway, BEng, of the Cranfield Institute of Technology.

The jury declared that Rob Peebles' studied contribution indicated a possible (very advanced) future housing structure which deeply interested the jury, and Mark Preskey, while offering no new design as such, showed how detailed study and calculation on matters of heat flow through structures, ventilation, air spaces, etc. could be used in pursuit of the much more stringent U-values which might be applied by 1990.

Each winner received a commemorative plaque and a cash prize of £200. Their colleges received a wall plaque and a cash prize of £100.

The presentations were made by Mr Alan G. Turner, Chief Executive BPB Industries plc at a celebratory luncheon in London on 10th March, 1982.

The Inplan Award competition is sponsored by Coolag Limited, Plaschem Limited, Urethane Foam Operatives and Co Limited who all market rigid polyurethane products, together with ICI Polyurethanes group which markets polyurethane chemicals.

From left to right: Mr R. Peebles, Mr M. A. Preskey and Mr A. G. Turner, Chief Executive of BPB Industries plc.