AD
TECH REPORT 68-10
DEC 1968

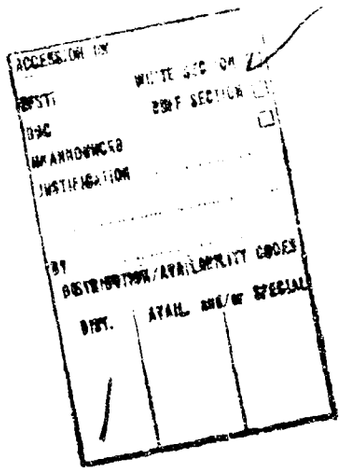
TECHNICAL REPORT

APPLICATION OF REGRESSION ANALYSIS TO COST ANALYSIS

JOHN L. HAMILTON
JAMES T. WORMLEY

SYSTEMS AND COST ANALYSIS DIVISION
COMPTROLLER AND DIRECTOR OF PROGRAMS
U.S. ARMY MATERIEL COMMAND
WASHINGTON, D.C., 20315

26



DISPOSITION

DESTROY THIS REPORT WHEN NO LONGER NEEDED.

DO NOT RETURN IT TO THE ORIGINATOR.

DISCLAIMER

THE FINDINGS IN THIS REPORT ARE NOT TO BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION

3

Technical Report 68-10

Application of Regression Analysis to Hardware Cost Estimation*

JOHN L. HAMILTON
and
CPT JAMES T. WORMLEY
US Army Materiel Command

December 1968

The views expressed in this Technical Report are those of the authors and do not purport to reflect the position of the Department of the Army or the Department of Defense. Technical Reports are reproduced by the US Army Materiel Command as a courtesy to members of its staff and to provide a means of reaching a wider professional audience.

Systems and Cost Analysis Division
Comptroller and Director of Programs
US Army Materiel Command
Washington, D.C. 20315

* The basic text of this paper was presented by the authors to the 19th national meeting of the Joint Study Group on Military Resource Allocation Methodology (JSGOMRAM), April 1968. This meeting was held at the Research Analysis Corporation at McLean, Virginia.

Acknowledgement

The authors wish to express their appreciation to Major Horace Schow II who reviewed the drafts and provided advice on methodology and techniques.

5

Abstract

This report presents an example of regression analysis which illustrates the major judgmental considerations in the development of a cost estimating relationship. The example used is the development of hardware costs of turbine aircraft engines. The methodology discussed is most useful for "quick reaction" studies and has been used by Headquarters, US Army Materiel Command for this purpose. Particular points discussed are. scatter diagrams, net scatter diagrams, causal requirements, combinations of variables, and sample selection.

TABLE OF CONTENTS

9

LIST OF FIGURES AND TABLES

APPLICATION OF REGRESSION ANALYSIS TO HARDWARE COST ESTIMATION

The purpose of this technical report is to illustrate the application of regression analysis to hardware cost estimation. The example given in this report uses hypothetical data which have been generated to best illustrate the analytical methodology being presented. These data have been selected for two reasons. First, there would be a possible breach of proprietary information if actual AMC costing equations were used. Second, it was necessary to select the data so that all important considerations could be illustrated clearly. A working knowledge of basic statistics is assumed in this discussion.

The example presented in this study is the estimation of hardware cost of turbine aircraft engines. Typically, a request would have been received for cost information which could be used for program and budget purposes and for prediction of possible cost overruns. The information was to be developed on the T3/5 family of engines with emphasis on the TX, a follow-on engine to be produced by the same contractor which produced all other family members. This study was selected because it illustrates most of the considerations to be made during a statistical cost analysis. Experience gained in the use of scatter diagrams and regression analysis will be presented.

In solving a statistical analysis problem of this type,

the first and most important step is to gather available, historical, analogous data. The Army had procured several T3/ engines for which historical costs were known. Based on data from the successive contracts for each model, slopes of the respective learning curves could also be derived. From this data the cost of the hundredth unit for each model was derived as the comparable costs for the various engines.

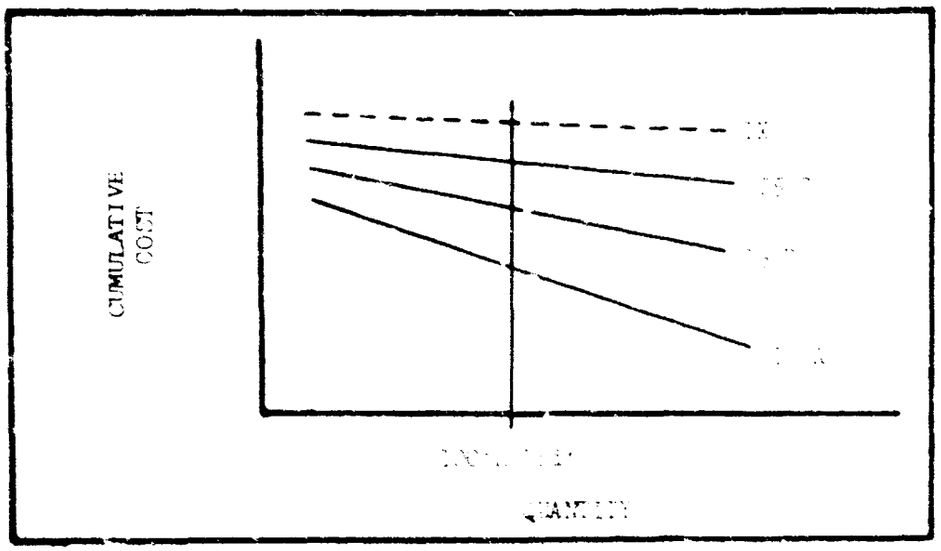


Figure 1
Graphic Representation of 100th Unit Cumulative Average Costs and Learning Curves

Figure 1 illustrates that the slopes have become shallower in this case for the newer engines. Therefore, a tentative slope for the new TX could be presumed to be even shallower. When the 100th unit cost of the TX is calculated the associated learning curve may be similar to the dashed line shown for the TX.

13

Table 1
Hypothetical Turbine Aircraft Engine Data Base

Engine Type	Revolutions per Minute	Shaft Horsepower	Specific Fuel Consumption	Installed Weight	No. of Stages	Cost of 100th Engine
T3 A	6300	900	.52	510	4	50600
T3 B	9000	900	.56	530	4	62300
T3 C	7500	960	.59	545	5	57600
T3 D	7200	1600	.63	560	5	67300
T5 A	8400	2300	.65	615	6	73100
T5 B	10500	2500	.66	640	4	88000
T5 C	10800	3700	.71	695	4	106200
T8-GE-5	7000	450	.70	136	2	48300
T4-GE-6	7500	1000	.61	723	6	61000
JET-12A-1	18000	4800	.81	882	7	210000
JET-12A-2	20000	4800	.87	882	7	235000
TX	12000	4000	.75	730	5	------

Table 1 shows the collection of all available data for turbine aircraft engines including performance data, technical data and 100th unit cost data. The list of engines for which data is available includes all those which show similar characteristics to those of the TX for which cost is to be predicted. The regression analysis developed here will relate one or more of these characteristics of the engines to their cost in a cost estimating relationship (CER).

The first step in CER development is the judgmental selection of all systems which are similar to the specific system being studied. The JET-12A 1 and -2 are not turbine engines but jet engines, so that data may be discarded. The T4 and T8 engines

were not produced by the same contractor who produced all other T3/5 engines. This data can therefore be tentatively rejected even though the cost and performance data are within the range of the T3/5 data. The four engines just discussed are not considered to have homogeneous or analogous characteristics since the costs requested are those of one particular contractor. After the rejection of four engines, there are data on seven engines remaining and this is a sufficiently large sample to form the basis of a CER. If there were data on only two or three T3/5 engines, consideration would have been given to using the similar T4 and T8 engines to provide a larger sample size. A sample must be large enough to provide statistical confidence in the resulting CER.

The next step is to examine available performance and technical characteristics to assure that the variables may actually be used to predict costs. In this case, installed weight must be rejected even though there appears to be a relationship between cost and this characteristic. Rejection is necessary because all new engines are being installed in lighter, more expensive mounts so that costs would still continue to rise as installed weight drops. We therefore reject installed weight as a future predictor. The type of mount is independent of the engine used. An old engine could also use a new mount.

15

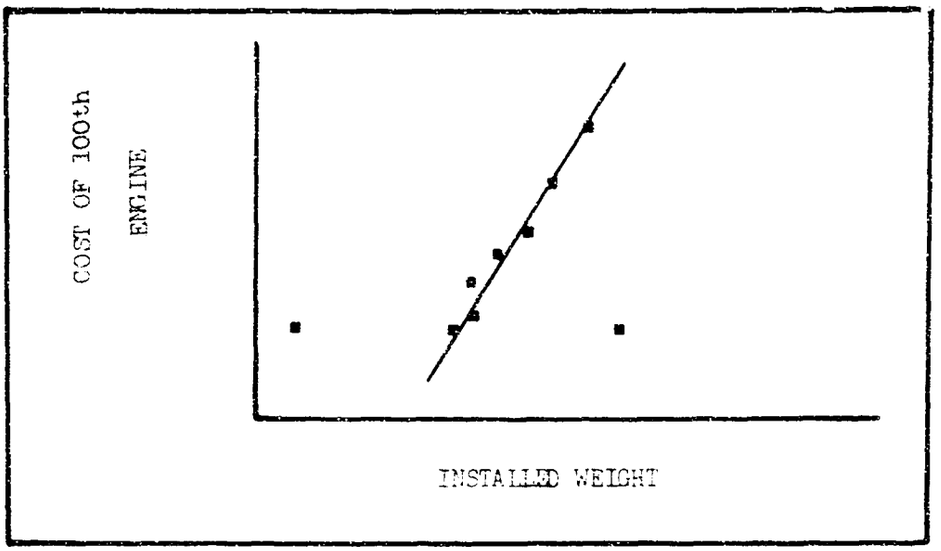


Figure 2
Turbine Aircraft Engine Installed Weight vs. Cost

Figure 2 which shows installed weight versus cost, is shown only to illustrate the point that data which appears excellent may have to be rejected on closer evaluation. Also, notice that the T4 and T8 engines which are produced by another contractor have no apparent relationship to the T3/5 group. After the deletion of installed weight, the remaining independent variables are revolutions per minute, military shaft horsepower, specific fuel consumption and number of compressor stages.

One must be very careful to insure that there is a cause and effect relationship between the independent variables and cost. As a recent Cost-Effectiveness Newsletter* stated, "There is probably a good correlation between men's

*The CE Newsletter, Volume 3, Number 1, February 1968, page 3.

shoe sizes and their heights. Therefore, one likely way to reduce a man's height is to chop off his toes." The moral is; make sure the long feet actually cause height before you cut off the man's toes.

Figure 3 shows shaft horsepower plotted against cost. An analyst familiar with mathematical functions may observe the possibility that a square root function may "fit" this data.* The data for this variable can easily be transformed into a square root function and plotted again to check this assumption.

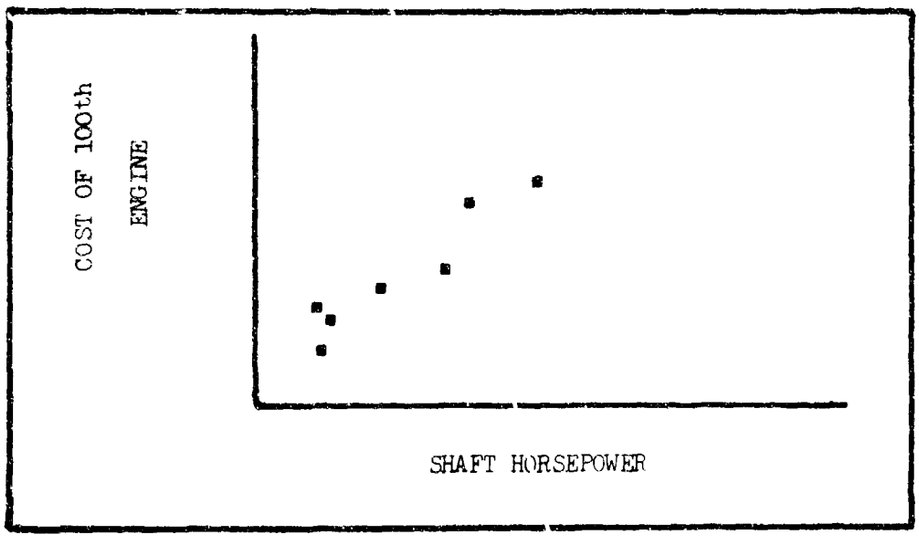


Figure 3
Shaft Horsepower vs. Cost

The data is transformed into a square root function in Figure 4. A general square root function seems to "fit" the transformed data fairly well. It will be useful to consider

*See Appendix for explanation of use of nonlinear terms.

17

the square root of shaft horsepower as an independent variable, a variable which logically may cause cost and therefore be a good predictor of cost.

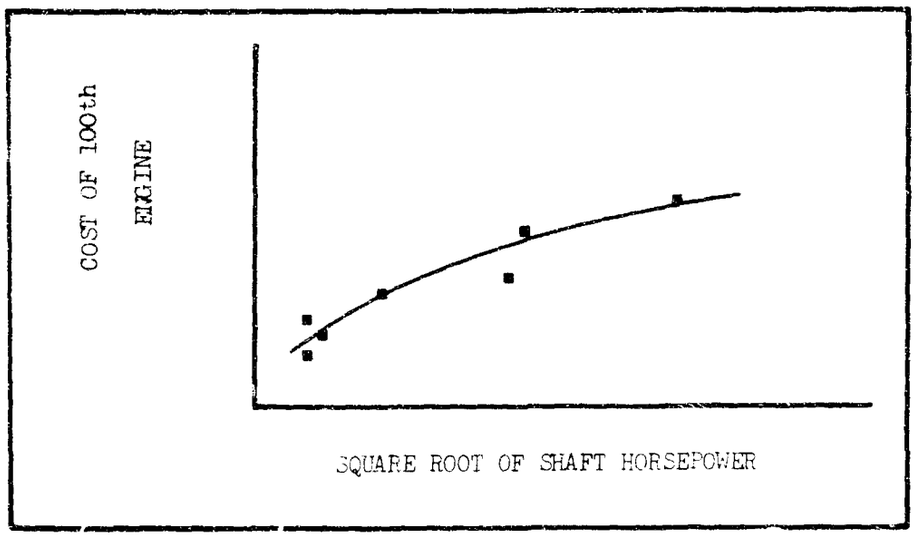


Figure 4
Square Root of Shaft Horsepower vs. Cost

Figure 5 shows revolutions per minute times number of stages plotted against cost. This combination was chosen because it represents an engineering relationship. Combinations based on engineering relationships often make good CERs because they are often the relationships which actually cause cost. The observations on this graph are very scattered and are not considered a good possibility for mathematical expression, so this relationship was rejected.

After all likely variables, transformations of variables and combinations of variables have been chosen, a least

squares line or multiple linear regression is computed using standard statistical regression techniques. These techniques are well known, are covered in standard textbooks such as those referenced at the end of this report, and will not be given here.

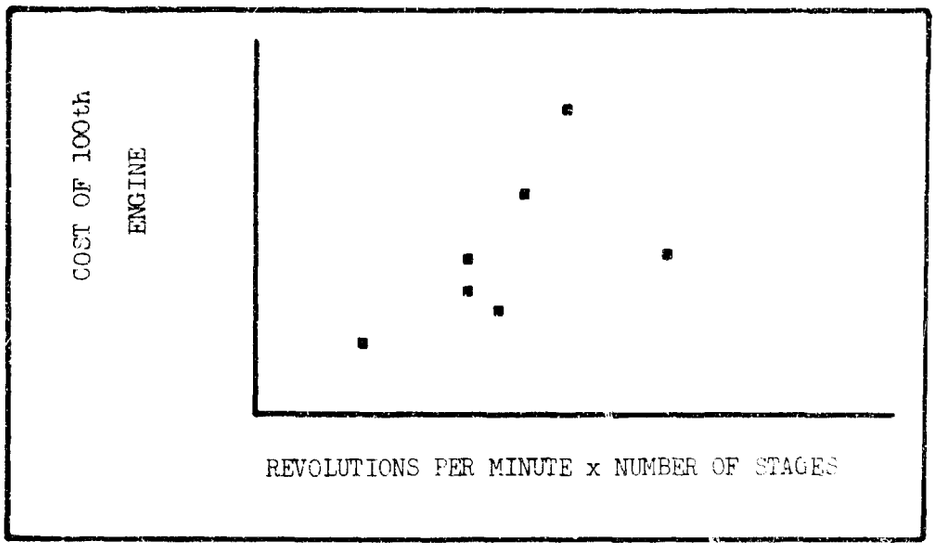


Figure 5
Revolutions per Minute vs. Cost

What has been done so far is not difficult or time consuming. However, to adequately consider all possible relationships and derive the best available CER requires many scatter diagrams and calculations of multiple variable regression lines. If multiple variables are used, simple one variable scatter diagrams of the type illustrated contain the effects of more than one variable and become useless to the analyst using visual inspection methods.

19

With the assistance of a small computer these problems can be solved quickly. The rush requirement is especially typical of the Army environment where all cost analysis is due yesterday. Since a computer of any size is usually a limited resource and access time is slow, the cost analyst must make the best use of all available techniques to expedite finding the one or more variables that best explain cost. The procedure to be described requires a minimum of two multiple regressions. If more computation effort can be afforded so much the better. Planning the procedure to be used will afford good, timely results.

The first step in this procedure is to compute a linear multiple regression using as many variables or combinations of variables as good judgment and the computer will allow. Hopefully, the program will discriminate and select only significant variables. Significant variables are those which in some way relate to or explain changes in cost.

After the linear regression line has been computed it is used to recompute the costs of all of the engines in the sample. These new costs are called "computed costs" as opposed to the observed or actual costs. The difference between computed and observed costs for each observation is called the residual. These residual costs with the proposed regression line will be used to help identify as yet undiscovered but meaningful relationships.

The scatter diagrams shown in earlier figures showed the variation or residual about a least squares line computed with only one independent variable. This variation contained no effects from other variables but did include an error term. The multiple regression line now contains the influence of one or more significant variables relating to cost. These residuals are determined by all the independent variables included in the line plus an error term.

It is possible to isolate and look at one variable at a time on a "net" scatter diagram by plotting the regression line at the mean of all but one of the independent variables and diagramming the one isolated variable versus cost.

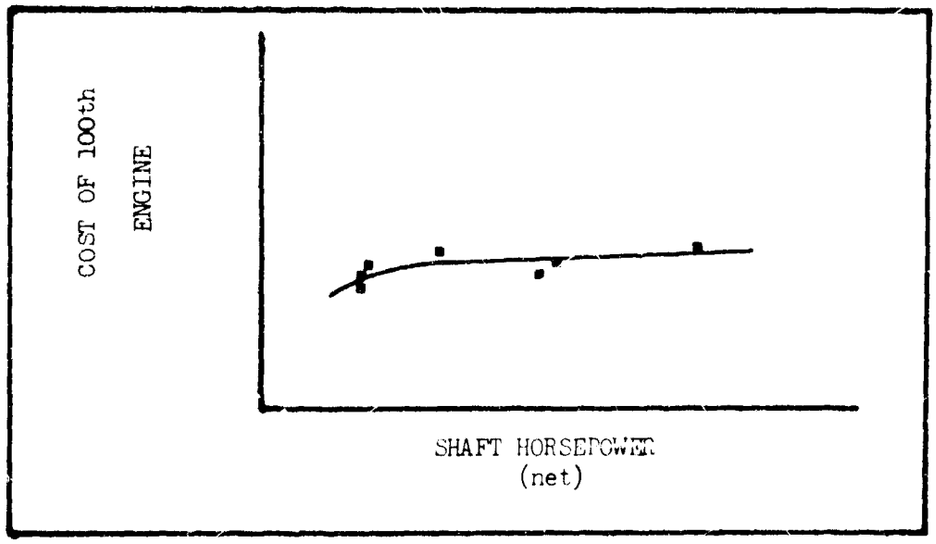


Figure 6
Net Scatter Diagram of Shaft Horsepower vs. Cost

The two-variable net scatter diagram illustrated in Figure

21

6 for shaft horsepower and cost nets out, or eliminates, variation caused by other independent variables so that the true relationship between cost and shaft horsepower can be studied in a manner similar to the previous single-variable scatter diagrams. Here the regression equation (or tentative CER) is plotted with all other independent variables valued at their mean so the scatter of residuals about the line can be studied for some clue as to the true function of only shaft horsepower and cost. In Figure 6, an overlayed general square root function (line) illustrates a good "fit" to the data. Therefore, shaft horsepower should be considered as a good predictor of cost when transformed into a square root function.

Figure 7 shows the net scatter diagram for revolutions per minute. This chart indicates that the linear (untransformed) function of revolutions per minute should also be considered as a possible variable.

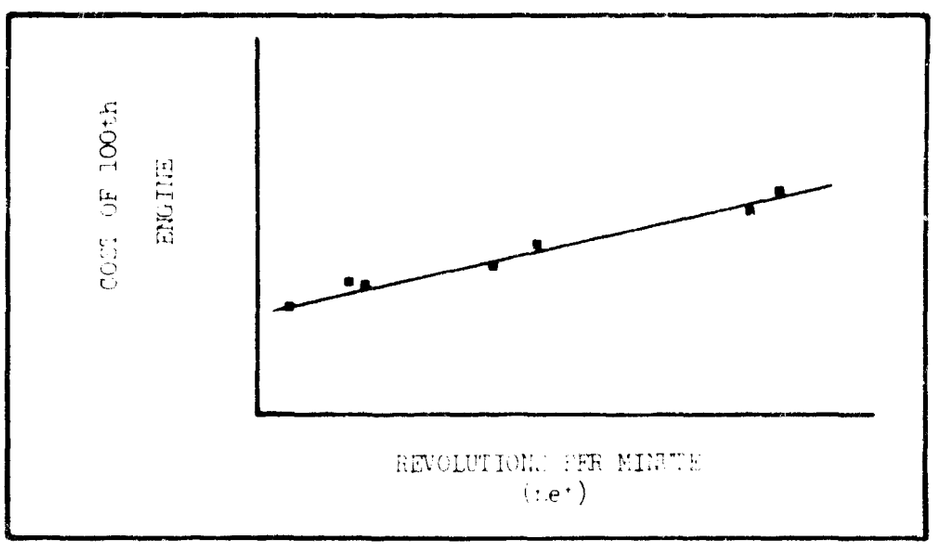


Figure 7
Net Scatter Diagram of Revolutions Per Minute vs. Cost

The net scatter diagrams can be used to determine what, if any, new functions of variables should be tried in the multiple regression. Otherwise, net effects of a variable can be buried by the interactions of other variables and a good relationship will be ignored. In this case the two best variables were RPM and the square root of shaft horsepower. Of course, more than two independent variables may be used as necessary. After assuming these two variables as "best" there arise two statistical problems associated with the use of these two particular variables. First, as the number of variables used in an equation increases, the statistical confidence in the equation decreases. This means that an equation with two variables is not likely to be as statistically significant as an equation with one variable. Second, there is a high degree of correlation between these two particular independent variables. This means, for example, that RPM may "cause" shaft horsepower as well as cost. The result is an equation with two unknowns. The only significant correlation allowable for an independent variable is with cost. These statistical problems usually would cause further search for more independent variables. In this example there is another solution. An alternative source of causitive variables is an engineering combination of the two variables. The statistical problems above do not occur if two variables are combined into

one variable. The combination is simply treated as one variable. And engineering relationships may, indeed, be true cost-causing variables.

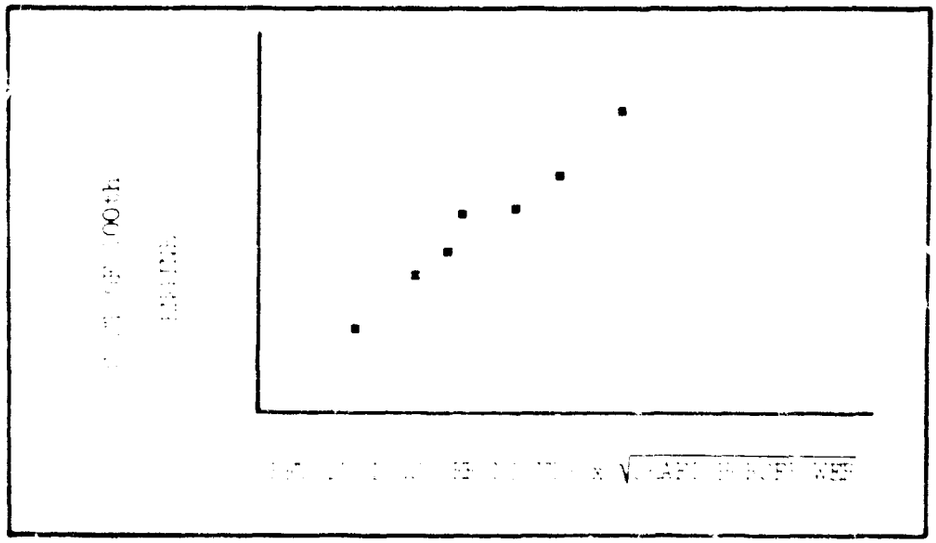


Figure 8
Scatter Diagram of RPM x Square Root of SHP vs. Cost

Figure 8 illustrates the data for the engineering combination (multiplication of RPM and square root of shaft horsepower) to form one new independent variable

The regression line in Figure 9 was developed from this combined data. It is obviously a very good fit to the data. The equation is statistically significant using statistical measures such as the F test. Also, four requirements are satisfied. That is, the CER is unbiased, consistent, efficient and sufficient. The standard deviation of the estimate is less than ten percent of the mean value of the

independent variable. The equation, which appears on the figure, predicts the actual observed data with an average deviation of 2.6 percent of the observed value. The individual deviations run from 0.7 percent to 5.8 percent -- all well within the expected deviation of a good cost estimate. The equation explains 98.5 percent of the variation of the data about the mean.

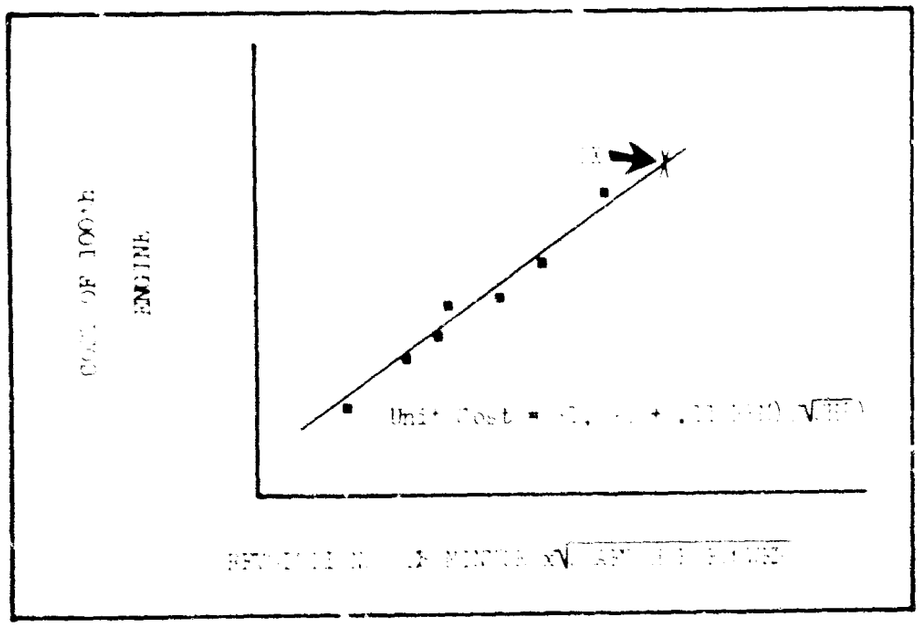


Figure 9
Turbine Aircraft Engine Cost Estimating Relationship

Anyone can solve for cost using this equation by simply substituting the product of RPM and square root of shaft horsepower of the system to be costed, for instance the IX. The 100th unit cost of the IX using the proposed specifications

25

computes to about $115,000, a figure slightly outside the range of the known data. In this case extrapolation beyond the range of the historical data is acceptable since the TX is only a small amount larger in all characteristics than the largest known sample and because the CER fits the data so well. That is to say, the statistical variation is very small so the prediction interval is also small, even when extrapolated a small amount.

This technique is used by the Cost Analysis Branch to prepare credible cost estimates without expending extensive resources in the process. In review, the process involves the following major activities:

1. Collect and analyze the relevant data on only those systems for which data are available.

2. Hypothesize relationships affecting cost and plot scatter diagrams.

3. Test promising variables using multiple variable regression analysis and plot new scatter diagrams.

4. Find the best relationships and compute a CER and test for statistical significance.

5. Insure that the CER is logical, reasonable and useful before publishing.

27

REFERENCES

Ezekiel, Mordecai, and Karl A. Fox, Methods of Correlation and Regression Analysis, (3rd Edition), John Wiley and Sons, Inc., New York, N.Y., 1959

Hale, Jack, Multi-Variable Nonlinear Regression Handout, Cost and Economic Analysis Department, Air Force Logistics Command, Wright-Patterson Air Force Base, Ohio, Dec 1956

Hamilton, John L. and James T. Wormley, "T53/T55 Cost Analysis Methodology", presented to 19th National Meeting of the Joint Study Group on Military Resource Allocation Methodology, Research Analysis Corporation, McLean, Virginia, April 19, 1968

Richmond, Samuel B., Statistical Analysis (2d Edition), The Ronald Press Company, New York, N.Y., 1964

Headquarters, U.S. Army Materiel Command, Engineering Design Handbook, AMCP 706-110 July 1963

APPENDIX

The title "Linear Regression" does not mean that non-linear relationships cannot be considered. Nonlinear functions such as square roots, log functions, etc., can be transformed for use in a linear equation by simple methods such as those illustrated in AMC Pamphlet 706-110 (reference 5 to this report).

In general, the procedure is this:

if $Y = a + b\,[f(x)]$

set $f(x) = x'$ and solve for the regression equation in the normal manner. The "assumption of linearity" is basic to linear regression. An "error in specification" results if this assumption is not valid. In a "quick-reaction" context, the only test for error in specification is the use of net scatter diagrams.

DISTRIBUTION LIST

Copies	
20	Defense Documentation Center, Cameron Station, Alexandria, Virginia 22314
130	Commanding General, U. S. Army Materiel Command, ATTN: AMCCP-S, Gravelly Point, Washington, D.C. 20315

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) HQ US Army Materiel Command Washington, D.C. 20315	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

APPLICATION OF REGRESSION ANALYSIS TO HARDWARE COST ESTIMATION

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

5. AUTHOR(S) (First name, middle initial, last name)

John L. Hamilton
James T. Wormley

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
December 1968	17	4

8a. CONTRACT OR GRANT NO. b. PROJECT NO. c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) TR 68-10 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT

UNLIMITED

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	HQ US Army Materiel Command Washington, D.C. 20315

13. ABSTRACT

This report presents an example of regression analysis which illustrates the major judgmental considerations in the development of a cost estimating relationship. The example used is the development of hardware costs of turbine aircraft engines. The methodology discussed is most useful for "quick reaction" studies and has been used by Headquarters, US Army Materiel Command for this purpose. Particular points discussed are: scatter diagrams, net scatter diagrams, causal requirements, combinations of variables, and sample selection.

DD FORM 1 NOV 68 1473 REPLACES DD FORM 1473, 1 JAN 64, WHICH IS OBSOLETE FOR ARMY USE.

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Cost Analysis						
Cost Estimation						
Regression Analysis						
Linear Programming						
Cost Estimating Relationship						