PRESENTATION TO THE
THE SIXTEENTH ANNUAL
DEPARTMENT OF DEFENSE COST ANALYSIS SYMPOSIUM

TITLE: LEGITIMATE TECHNIQUES FOR IMPROVING THE R-SQUARE AND RELATED STATISTICS OF A MULTIPLE REGRESSION MODEL

AUTHOR: EDWIN J. CURLE
COMPTROLLER OF THE ARMY (DACA)

404132

HOSTED BY THE COMPTROLLER OF THE ARMY
SHERATON NATIONAL HOTEL
ARLINGTON, VIRGINIA
OCTOBER 4-7, 1981

ABSTRACT

Cost analysts and DOD contractors frequently use regression analysis to develop Cost Estimating Relationships, production relationships, and various forecasting equations. Invariably, those regression equations are presented in the text of the final report along with the statistical properties -- i.e. the R-Square, the Standard Error of the Estimate, the Durbin-Watson Statistic, etc. These statistics are often presented as evidence of the validity and accuracy of the resulting equation. The higher the R-square the bolder the print and the more prominently displayed.

Unfortunately, high R-square's, favorable Durbin-Watson statistics, etc. can be artificially or inadvertently inflated to appear more favorable. In reality, the equation with good statistical properties may not reflect a valid causal relationship to explain variations in the dependent variable. In many cases the regression equations prove to be of little value in forecasting or explaining the relationships with new data.

This paper discusses techniques for artifially raising the R-square and related statistical properties of regression equations. These techniques are presented for the benefit of analysts who are trying to improve the statistical properties of their equations and for the benefit of managers who must approve payment for such analysis.

LEGITIMATE TECHNIQUES FOR IMPROVING THE R-SQUARE AND RELATED STATISTICS OF A MULTIPLE REGRESSION MODEL

INTRODUCTION

There are a number of factors which contribute to the reliability and validity of a regression equation -- the underlying theoretical structure, the relevance and accuracy of the variables used, the validity of the statistical procedure used, and the achievement of clean residuals that reflect pure white noise. Each of these factors are essential to the development of good analysis.

Unfortunately, when it comes time to develop or evaluate regression equations, the statistics generated by the estimation procedure seem to get all the attention. Statistics are presented in the text of the report along with the actual equations estimated, but no mention is made of how the model was specified or how the residuals look. Instead, one of the first questions asked by analysts and reviewers is "How high is the R-square?"

Because of it apparent importance in the estimation process, this paper focuses on the R-square statistic--what it is, how it can be interpreted, and how it can be used or abused. The purpose is to discuss the sensitivity of the R-square statistic to variations in the data, different functional forms, and incorrect procedures which may inadvertently be used in the estimation process.

In the process of discussing the sensitivity of the R-square to changes in data, functional form, etc., the paper covers techniques that can be used to raise the R-square without contributing to the validity of the model.

Hopefully, this paper will point out the shortcomings of over-reliance on the R-square in developing and evaluating regression equations. It is also hoped that the discussion will stimulate an interest in looking at other statistical properties and trying to identify sensitivities of those tests to particular types of data, functional forms, or implied transformations.

What the R-Square Really Means and When it Should be Used for Comparing Different Regression Equations

It should be noted that there is some disagreement among statisticians over the appropriate formula to use in calculating the R-square. (See Belsley et al, p. 86) But there is no disagreement over the meaning or implication of the term. In the traditional sense the term R-square describes the percentage of variation about the dependent variable that is explained by the independent variables included in the model. The implication is that an equation with a higher R-square explains a greater percentage of the variability in the dependent variable and that the equation with a higher R-square is somehow better than another equation with a lower R-square. Obviously, one would like to find some mechanical procedure for developing an equation with the highest R-square value.

The step-wise regression approach does just that. In a merely mechanical way, the step-wise regression package takes a series of independent variables and adds or drops variables in successive calculations so as to obtain the one relationship among the alternatives which has the highest R-square. Many amazing discoveries have been made on the basis of results from a step-wise regression program.

Unfortunately, many errors have been made using the Step-wise approach because of its reliance on the R-square statistic which it seeks to maximize.

Pitfalls to Avoid in Using R-square as a Criterion for Selecting the Best Regression Model

In the first place, analysts should be aware that the formula for R-square does not have the traditional meaning unless the following conditions are satisfied. Furthermore, the formula can also produce values that lie outside the 0-1 interval unless these conditions are satisfied.

(1) The OLS estimation procedure is used.

(2) The relationship being estimated is linear.

(3) The linear relationship being estimated includes a constant term. (See Aigner (1971) for a more complete discussion of the zero intercept case and an alternative formula for R-square to use in these situations.)

The above conditions are significant for the analyst who wants to compare equations where one of the terms was estimated without a constant term. Similarly, the analyst should be aware that comparing the R-square developed for the non-linear transformation does not have the same meaning as the R-square developed for the original linear relationship. As a result the R-square for a log transformation of the data is not really meaningful relative to the R-square for the equation estimated from the original data.

Other Factors that can Effect the Value of R-Square

Extreme caution should be used when using R-square to evaluate alternative regression equations where the above conditions are satisfied. The following factors create higher (or lower) R-square values without significantly enhancing the validity of the model.

1. Range of Variation on the Dependent Variable

The formula for R-square is sensitive to the range of variation in the dependent variable. Two examples are worth noting and a Monte Carlo simulation is being developed to show the significance of greater variation in the dependent variable.

The classic example used to demonstrate the sensitivity of R-square to the range of variation of the dependent variable involves the estimation of the savings and consumptions functions. Since savings is defined as the difference between income and consumption, the regression equation for savings as a function of income should be equally as good as the regression equation for consumption as a function of income. That is C=Y-S, or S=Y-C. In reality, the sum of squared residuals (or the unexplained variation) will be exactly the same for each case. But in percentage terms the unexplained variation will be greater for the savings function than for the consumption equation. As a result, the R-square for the savings function will be lower than the R-square for the consumption equation. (See Barrett (1974)).

This implies that simple substitutions of equivalent terms can significantly drive up the R-square without contributing to the validity of the relationship.

In another case, the regression equation to explain earnings of all employees has a greater R-square than does the same equation for earnings of a subset of all employees, because the total population has greater variability than the subset of the total. This implies that regressions on a subset of data could have a smaller R-square than regressions on a larger set of data.

Preliminary results from Monte Carlo simulations show that the R-square is higher for a larger population with greater variability in the data than it is for a subset of data with less variability in the dependent variables taken from the same population.

2. Use of Dummy Variables or Time Trends

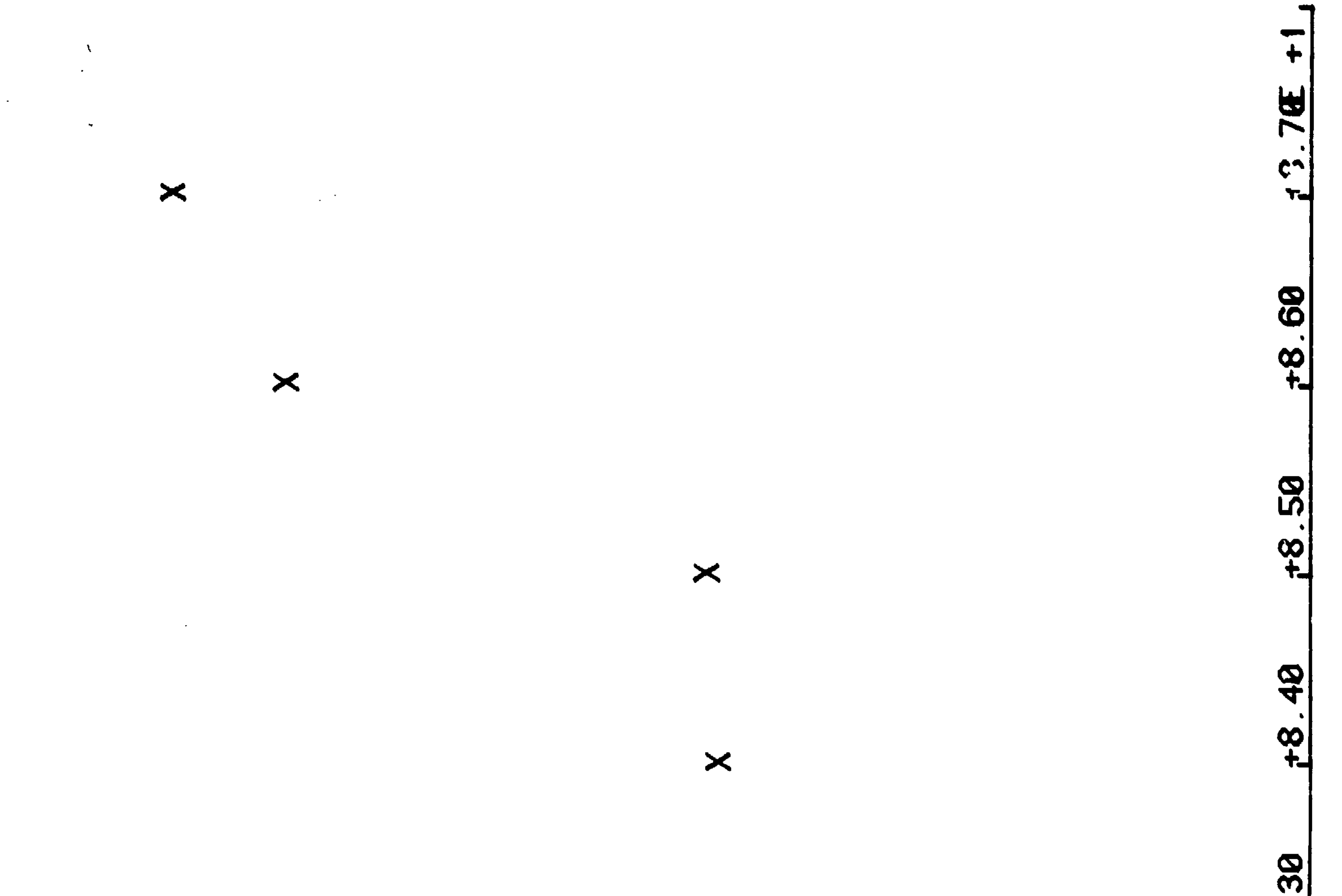
As a means of improving the explanatory value (and also the R-square) of a regression equation, statisticians will frequently introduce dummy variables or time trend data. The dummy variables are designed to capture the influence of events (strikes, wars, etc.) that cannot be quantified. The time variable captures trends in the data that may be time related and not explained by other variables in the model.

In both cases, the analyst must be sure that the dummy variable and the time trend are causal relationships and not just highly correlated.

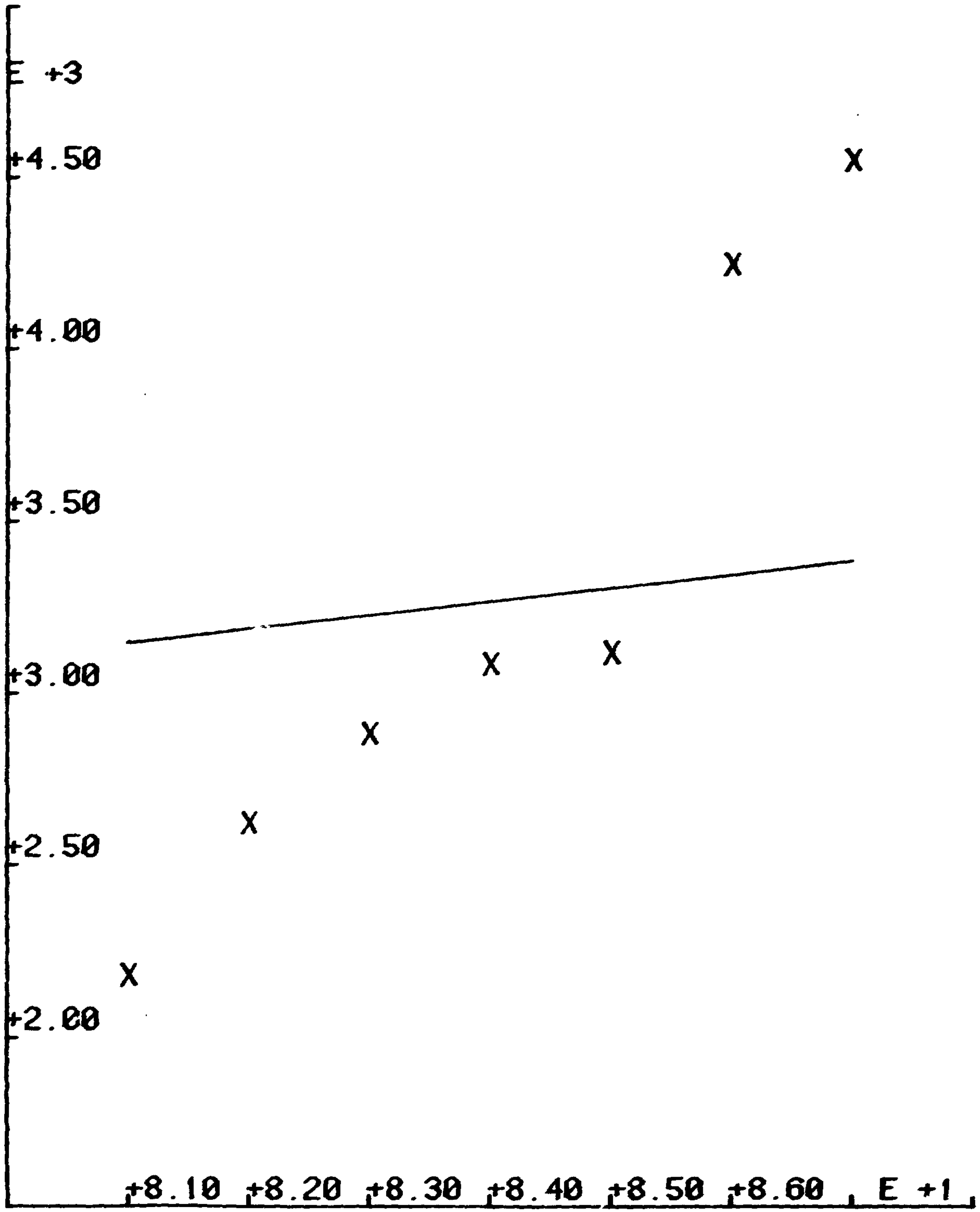
3. <u>Transformations in both data and functional form that increase the R-square Statistics</u>

Analysts frequently have an option of expressing data in constant dollar terms or in current dollars. Frequently, the choice will be motivated by which form of data will produce the greatest R-square.

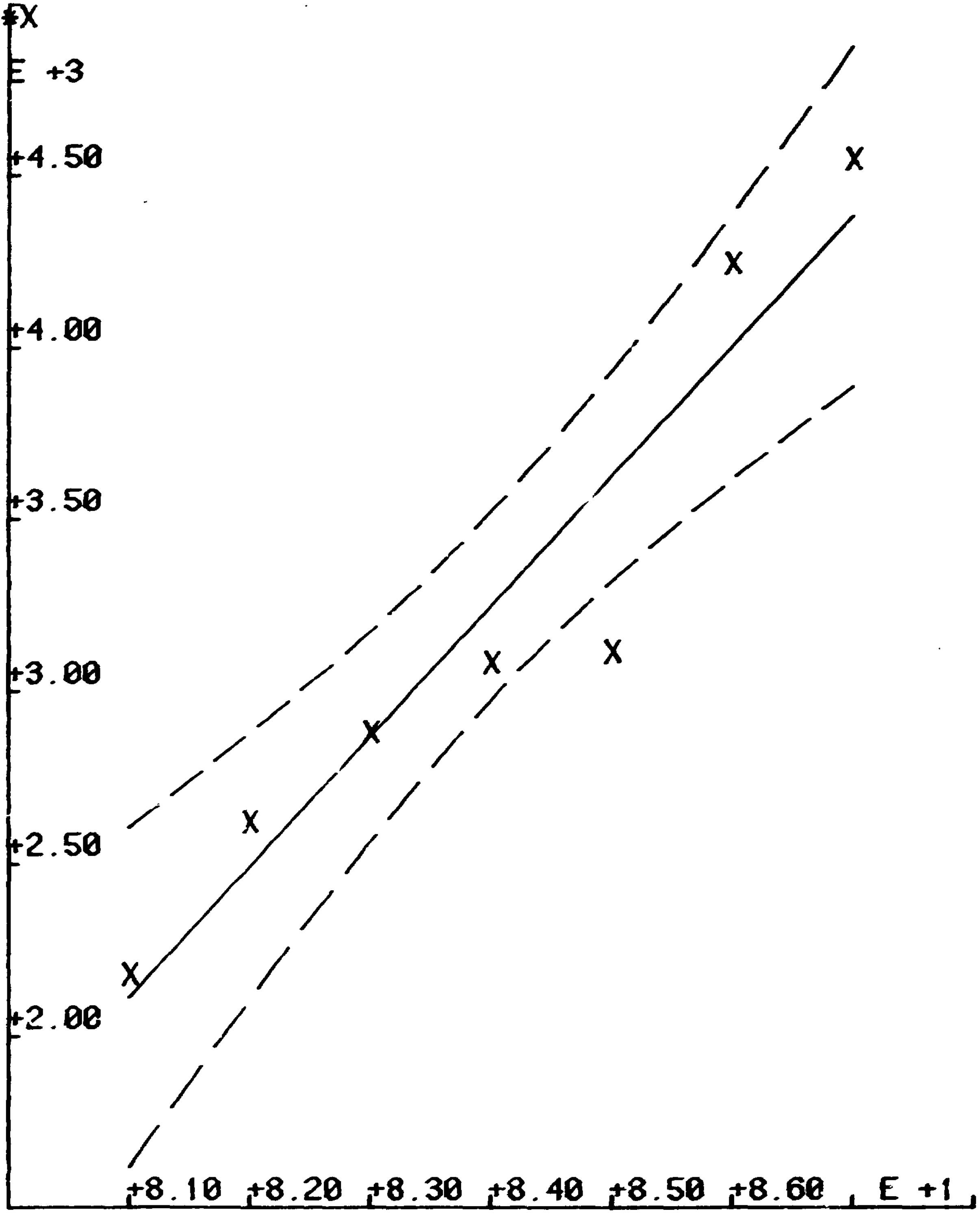
A common mistake is to express data in current dollars and then use [redacted] transformations to get a higher R-square. With the high rates of inflation that occurred in the late 70's this tends to produce [redacted] functional relationships like the following. Of course, the absurdity of this expression is demonstrated when projections are extended into the future and the forecast of the dependent variable increases at an astronomical rate.


+8.1Ø
+8.2Ø
+8.3Ø
+8.4Ø
+8.5Ø
+8.6Ø
+8.7ØE +1
+2.ØØ
+2.5Ø
+3.ØØ
+3.5Ø
+4.ØØ
+4.5Ø
E +3

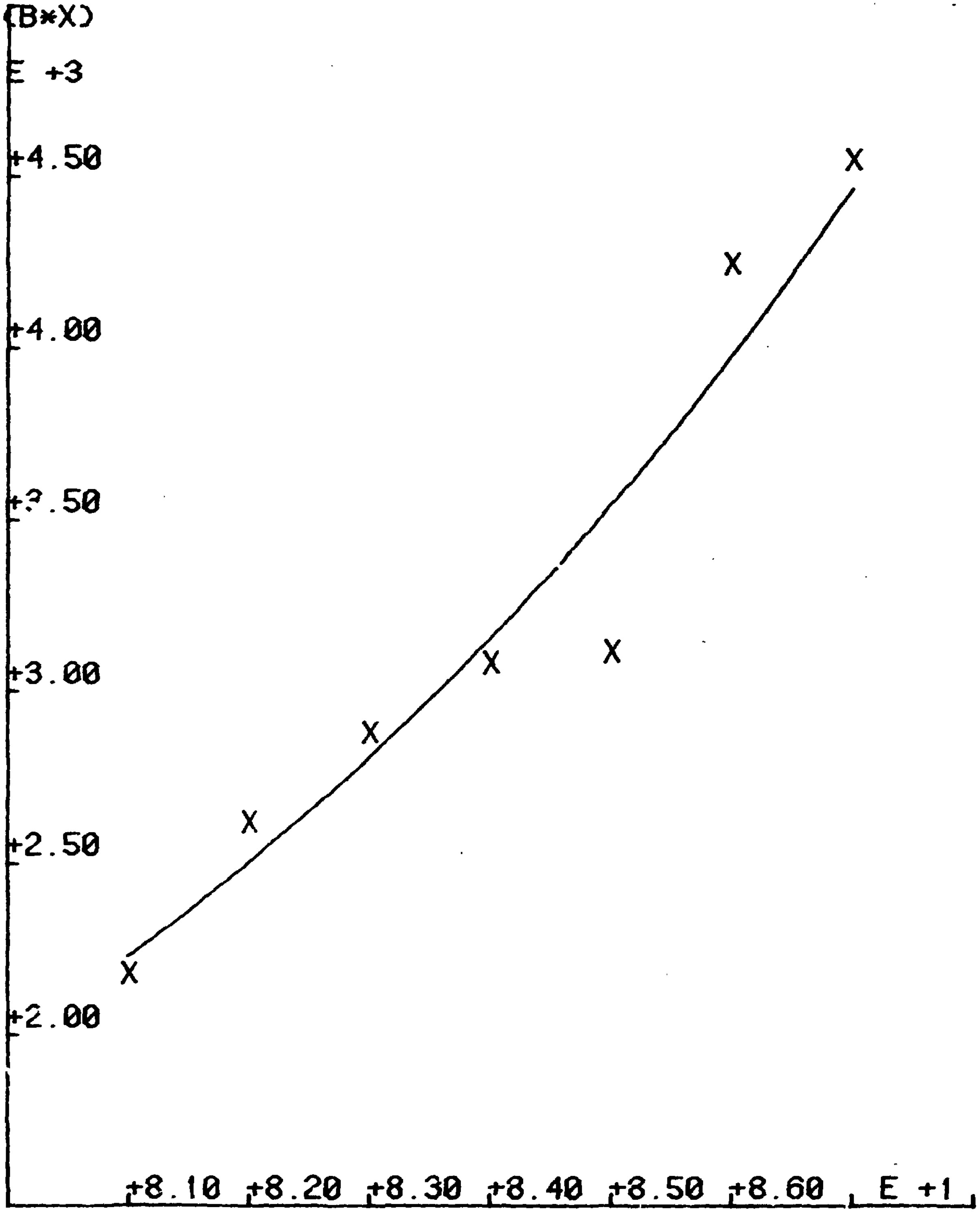
```
FOR EQUATION Y = A*X
Y = A*X

A =
 38.8386161675

R-SQUARE =
 0.17729651548

RES ERROR
 720970.517349

MAX(ABS(RESIDUAL))
 1173.82839342
```



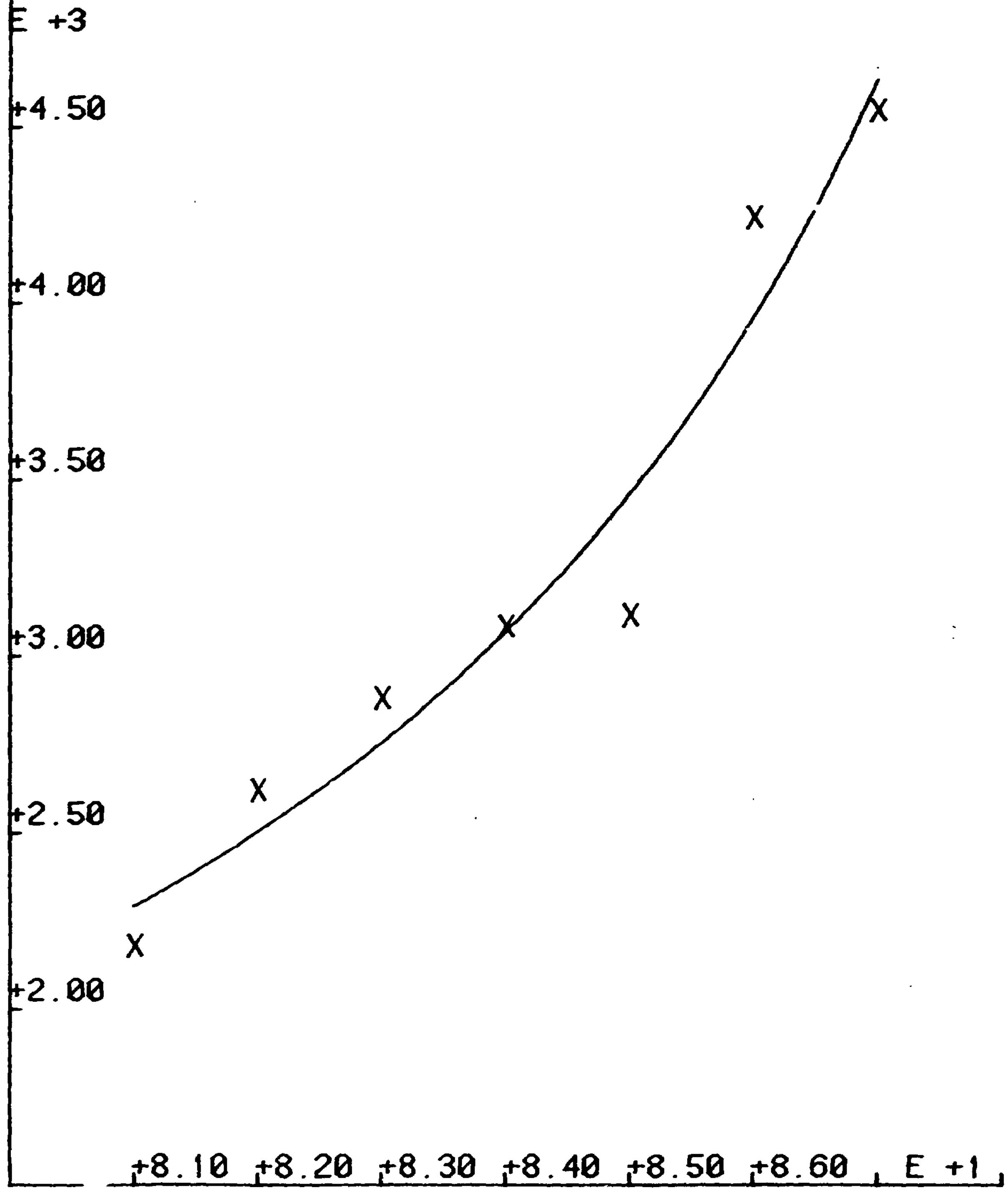
```
FOR EQUATION Y = A + B*X

Y = A + B*X

A =
 -28458.6851428

B =
 377.440535714

R-SQUARE =
 0.910355327672

RES ERROR
 78559.489539

MAX(ABS(RESIDUAL))
 500.972392857
```



```
FOR EQUATION Y = A*EXP(B*X)
Y = A*EXP(B*X)


A =
 0.191812559693

B =
 0.115567852792


R-SQUARE =
 0.934407471748

RES ERROR
 57481.5591744

MAX(ABS(RESIDUAL))
 417.738953229
```



```
FOR EQUATION Y = 1/(A + B*X)
Y = 1/(A + B*X)

A =
 0.00340976519052

B =
 -3.671143192E-5

R-SQUARE =
 0.943444186412

RES ERROR
 49562.2966831

MAX(ABS(RESIDUAL))
 333.909366583
```



4. Dropping the Outliers

There may well be a valid reason for dropping a lone data point that is far removed from the rest of the points. If the ommission can be justified by a good, logical reason, the analyst can legitimately increase the R-square of the regression equation by selectively ommitting data.

However, this technique can also be used for not-so-legitimately raising the R-square. Unfortunately, when the analyst drops data, he does not include the explanation for doing so. When an explanation is included, it is frequently buried in a footnote and not highlighted as well as the R-square and related statistics.

Alternatives to the R-square for Evaluating Regression Models

This paper is not intended to be an indictment of R-square or those who use the R-square for evaluating regression models. Rather, the purpose of this paper is to point out some of the pitfalls that may result from over-reliance on the R-square in developing regression equations.

Analysts should also be encouraged to look at the following alternatives which can supplement the more traditional statistics used.

1) The Mean Square Error

In the above example of regressing earnings data on the total population and the subset of engineers, it was noted that the R-square for the equation on the total population exceeded the R-square for the earnings of engineers. This would imply that the equation for the total population should be used instead of the equation for engineers.

However, the mean square error for the data on engineers was greater that the means square error for the regression using data for the total population.

In this case, the regression equation for engineers with the lower mean square error would have greater predictive value than the equation with the higher R-square.

2) Splitting the Data Base to Validate Estimates of Coefficients

Another technique for validating estimates is to randomly split the sample into two groups and run the regression for both groups. If the estimated coefficients are not significantly different, one can assume that the equation accurately identifies the relationship among variables.

3) Back-casting and Forecasting

It is frequently helpful in evaluating the merits of a regression model to estimate the dependent variable using data from a previous period of time and/or for future periods to see if the results of the equation are reasonable.

4) Tests for Specification Error

There are numerous tests available for detecting specification errors. The Durbin-Watson test for autocorrelation can be a good indication that a significant explanatory variable has been ommitted. Ramsey (1974) has developed a rather interesting test for detecting specification errors using estimates of the dependent variable in subsequent regressions.

5) T-Statistics

Of course, regression equations which have low t-statistics for the explanatory variables should be re-estimated or dropped in favor of equations where all the explanataory variables have statistically significant variables.

6) Does the Estimate Make Sense?

There must be some plausible causality between the dependent variable and each of the independent variables. This criterion eliminates the possibility of inducing variables with spurious correlation (i.e. sunspots, weather, etc.) This appeal to common sense also eliminates models where the coefficients take on the opposite sign from that which one would expect.

CONCLUSION

Some caution must be taken to insure that the statistics generated by the estimation procedure are meaningful and valid. In this case the R-square has been shown to be misleading unless reasonable care is taken in selecting the varibles to include in the model, the type of data to use, and the functional form to use.

Because of these shortcomings behind the R-square statistic, it becomes even more important to develop a strong theoretical structure behind the model and to correctly specify the equation before any attempt is made to select an estimation procedure.

Finally, it is extremely important to look beyond the R-square for other statistics and techniques that can support the model estimated.

REFERENCES

Aigner, D. (1971) Basic Econometrics Englewood Cliffs, N.J.: Printice Hall.

Barrett, J. (1974) 'The Coefficient of Determination--Some Limitations' The American Statistician vol. 28, February, pp. 19-20.

Belsley, David A., Edwin Kuh, and Roy E. Welsch. (1980) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, New York: Joh Wiley & Sons.

Chatfield, C. (1979) The Analysis of Time Series: Theory and Practice, London: Chapman and Hall.

Draper, N.R. and H. Smith. (1966) Applied Regression Analysis, New York, N.Y.: John Wiley & Sons, Inc.

Kennedy, Peter. (1979) A Guide to Econometrics, Cambridge, MA: The MIT Press.

Ramsey, J.B. (1974), "Classical Model Selection Through Specification Error Tests," in P. Zarembka, Ed. Frontiers in Econometrics, New York: Academia Press.

Theil, Henri. (1971), Principles of Econometrics, New York: John Wiley & Sons.

Wynn, R.F. and K. Holden. An Introduction to Applied Econometric Analysis, New York: John Wiley & Sons.