1989

NASA/ASEE SUMMER FACULTY FELLOWSHIP PROGRAM

MARSHALL SPACE FLIGHT CENTER
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

METHODS FOR TREND ANALYSIS: EXAMPLES WITH PROBLEM/FAILURE DATA

Prepared by:	Curtis K. Church, Ph.D.
Academic Rank:	Associate Professor
University and Department:	Middle Tennessee State University, Department of Mathematics and Statistics

NASA/MSFC:

Directorate:	Safety, Reliability, Maintainability & Quality Assurance
Office:	Systems Safety & Reliability
Division:	Reliability & Maintainability Engineering
Branch:	Problem Assessment
MSFC Colleagues:	Raymond Dodd, Ph.D. Frank Pizzano
Date:	July 20, 1989
Contract No:	The University of Alabama in Huntsville NGT-01-008-021

ABSTRACT

NASA Headquarters is emphasizing that statistics has an important role in quality control and reliability. Consequently, 'Trend Analysis Techniques' (NASA-STD-8070.5) recommended a variety of statistical methodologies that could be applied to time series data. The major goal of this report or 'working handbook', using data from the MSFC Problem Assessment System, is to illustrate some of the techniques in the NASA standard, some different techniques, and to notice patterns of data. Techniques for trend estimation used are: regression (exponential, power, reciprocal, straight line) and Kendall's rank correlation coefficient. The important details of a statistical strategy for estimating a trend component are covered in the examples. However, careful analysis and interpretation is necessary because of small samples and frequent zero problem reports in a given time period. Further investigations to deal with these issues are being conducted.

CONTENTS

INTRODUCTION

The purpose of this report or 'working handbook' is to discuss strategies and methods for statistical evaluation of trend for problem/failure data. Statistical analysis provides a tool to add insight and complement engineering judgement. Much of this work elaborates and clarifies the application of approaches contained in 'Trend Analysis Techniques' (NASA-STD-8070.5). Problem trend analysis tracks and categorizes problems over time. The problems may be for an entire system, subsystem, or any other appropriate level of aggregation. Techniques useful for statistically measuring a trend component will be illustrated in the next section. All examples contained in this report use data supplied by the MSFC Problem Assessment System.

There are two basic techniques for trend analysis in this report. One is regression and the second is a distribution free rank correlation method. Regression analysis is a statistical tool that utilizes the relation between two or more quantitative variables so that one variable can be predicted from the other, or others. A regression model is a formal means of expressing the two essential ingredients of a statistical relation:

1. A tendency of the dependent variable Y to vary with the independent variable or variables in a systematic way.
2. A scattering of the observations around the curve of statistical relationship.

We are interested in using a regression model to perform a test of hypothesis for trend. This requires that the experimenter postulate a probability model, commonly the normal distribution, to be used in the development of the hypothesis testing procedure. On the other hand, the distribution free approach eliminates specification of an underlying probability distribution. The rank correlation coefficient, known as Kendall's tau (τ), is used in the examples below as the distribution free basis for determining the existence of trend. For this method there are no assumptions about the form of the probability distribution involved and ther are minimal calculations.

METHODS AND EXAMPLES

The mechanics of applying the regression and rank correlation methods for problem/failure data will be covered through examples given below. The stategy for the regression approach will be as follows: examine a scatter plot of the data, fit an appropriate model to the observed data, perform a test of hypothesis for trend, and, if appropriate, generate a prediction interval for a future observation. In situations where there appears to be a positive (downward) trend the most advisable regression models are: exponential model, power (or multiplicative) model, or a reciprocal model. These models possess the desirable feature that the predicted values for the number of (normalized) problems will never be less than zero. The examples using the reciprocal model also illustrate the influence of extreme (or outlying) observations. In the case of an increasing trend a straight line model is also considered. Finally, Kendall's rank correlation coefficient will be adapted and discussed as an alternative means to perform a test of hypothesis for trend.

Fifteen sets of data on the SSME were selected from the problem assessment system. Each data set was normalized to give the rate of problem reports per 10,000 seconds of engine test firing time. Some data was at the system level, some at the subsystem level, and some at the failure mode level. These data sets are used to illustrate the issues and details of applying the above mentioned approaches. You will notice a common pattern in many of the data sets. There is an apparent adverse (increasing) trend from 1979 through 1982 plus or minus one year and then a positive trend from that point through 1988. This pattern appears in roughly 2/3 of the example data sets. In instances of an adverse trend followed by a positive trend, a model for the regression approach will be fit to the portion of the series exhibiting the positive trend, that is, the most recent six to seven years. In applying and interpreting the statistics one needs to rely on good judgement and sound engineering considerations. As you look through the examples notice similarities and dissimilarities in the data patterns. Statistical procedures cannot be applied in a vacuum.

Exponential Model

The exponential model is intrinsically linear. It is made linear by using a logarithmic transformation. Thus, applying an exponential model means that we will be regressing the natural logarithm of Y on time. The deterministic part of the model is:

$$Y = \beta_0 e^{\beta_1 t},$$

where: Y is the number of normalized problem reports,

β_0 and β_1 are parameters, and
t is a constant denoting the time period.

A positive (downward) trend is indicated if $\beta_1 < 0$. Thus, if an exponential model is fit to the data, the statistical justification for claiming a positive trend will be to perform a test of hypothesis of $H_0 : \beta_1 \geq 0$ against $H_a : \beta_1 < 0$. This presumes a flat or increasing trend in the null hypothesis in hopes that we have evidence to reject it in favor of the alternative hypothesis, which is a positive (downward) trend.

An important note is that the exponential model, through the logarithmic transformation, cannot be applied when there is a zero value in the normalized data. In instances where there are no problem reports in a time period and you wish to use the exponential model the logarithmic transformation needs to be modified. Under general conditions it is reasonable to approximate the number of normalized problem reports in the transformed data by using .5 as the number of problems before normalizing and transforming. So, if there are zero problem reports in a time period it will be replaced with .5 for the purpose of fitting the exponential model. This adjustment is relevant only for the logarithmic transformation. A justification for this modification is given in the appendix.

On the next several pages appear scatter plots and analysis summaries. The scatter plots display the number of normalized problems as asterisks (vertical axis) to the year (horizontal axis). For these five sets of data we observe the pattern of an apparent adverse trend for the first three to four years and then an apparent positive trend. These scatter plots were generated with the PC software package NWA Quality Analyst.

For each of these sets of data an exponential model provided a good fit to the data. The beginning time period for the fitted model varied from 1981 to 1983 and ended with the 1988 data. The regression analysis summaries below the scatter plots indicate the beginning time period. All of the computational work is performed with the transformed data. The PC software package Statgraphics generated the given regression analysis summaries.

Inspection of the coefficient of determination, r^2, value for each data set indicates a strong association between the normalized problems and year. Further evaluation of the appropriateness of the fitted model by inspecting a graph of the fitted curve together with the observed data or other means may be easily carried out using Statgraphics or other software package. The statistical justification for claiming a positive trend, however, lies in the hypothesis testing procedure. Fortunately, the Statgraphics regression summary contains results of all necessary calculations. To conclude acceptance of $\beta_1 < 0$ focus attention on what the regression summary labels "prob. value". For regression with two coefficients, the value of interest appears twice, once on the row identified by the slope and once in the analysis of variance table. For our purpose divide this number by two. This value is the observed significance level of the test, often called the p-value. It represents the likelihood, based on the observed data, of claiming a positive trend when there is actually not a positive trend. Thus, we want the p-value to be small, say less than .025 or possibly less than .01. Note that this value is less than .01 for each of the five data sets. The r^2 value and (twice) the p-value are circled on the following summaries.

Following the hypothesis test for trend, we may wish to forecast (or predict) a new value. We can then compare the predicted value with the new value when it becomes available to assess a continuing trend. For example, we may predict the rate of problems (i.e. the normalized data) for 1989 and compare with the first quarter rate when it becomes available. The predicted value comes from the fitted equation and then prediction limits are constructed. The 95% prediction limits will roughly be the fitted value plus or minus two standard deviations of the predicted value. The standard deviation is given by:

$$\sqrt{MSE\left\{1+\frac{1}{n}+\frac{(89-\bar{t})^2}{\sum(t-\bar{t})^2}\right\}}$$

The value n is the number of data points used in the model fit, and MSE is the mean square error that appears in the analysis of variance portion of the printout. Note that the sample size is incorporated in the prediction limits through the standard deviation.

For example, the data for the SSME main combustion system is:

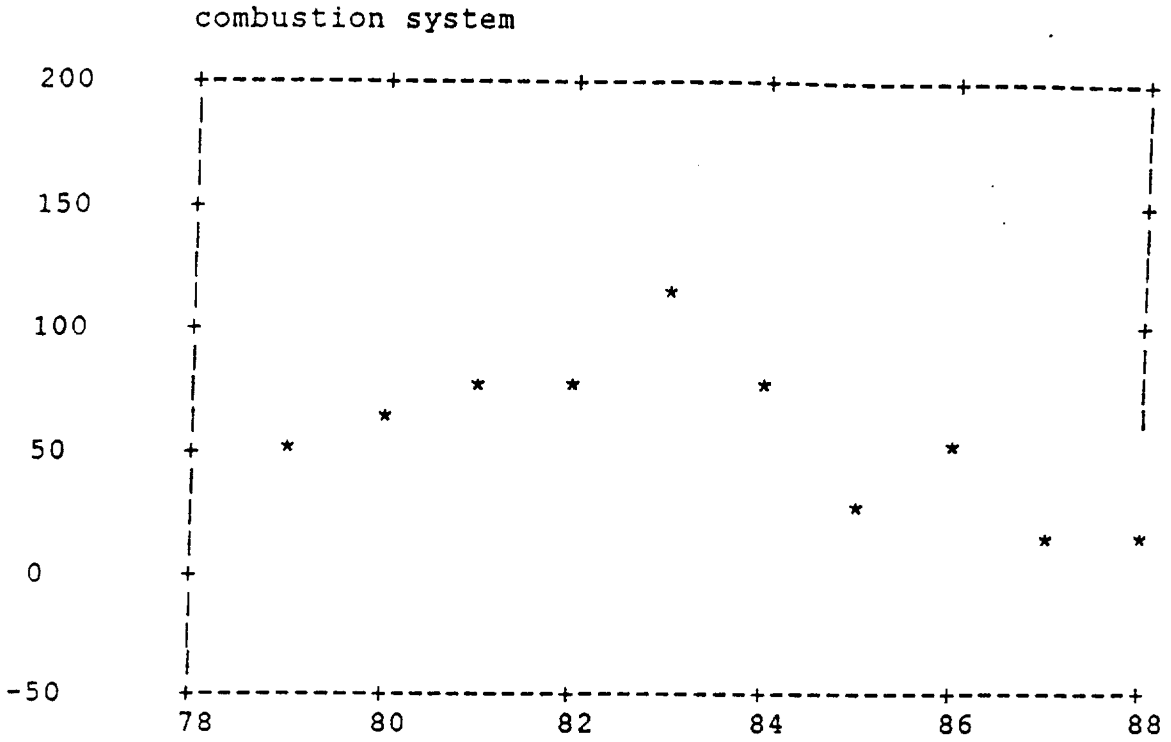
year :	83	84	85	86	87	88
normalized problems:	112.21	78.64	29.98	47.83	16.44	14.48

The predicted value for t=89 is 8.95. Through March 1989 the normalized value of problem reports is 11.88. The upper 95%

prediction limit is 25.82 and the upper 68% prediction limit (roughly one standard deviation) is 15.96. The calculations were done with the transformed data but expressed in terms of the original units.

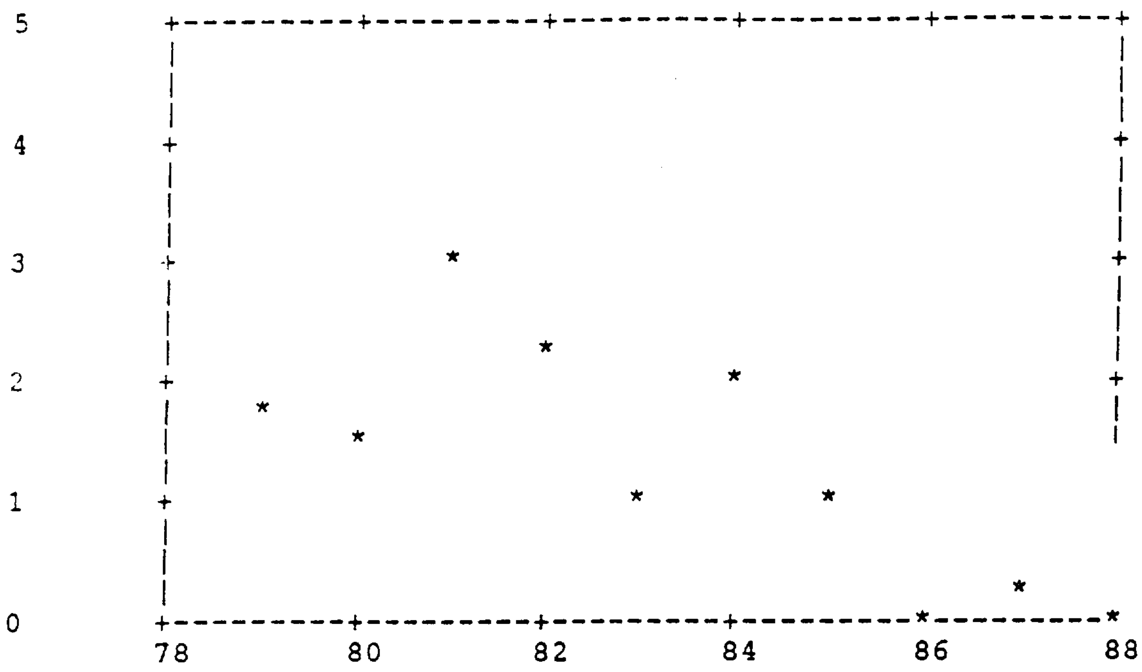
Regardless of the regression model chosen, the strategy is the same. Consequently, the other regression models that have been useful, the power model, the reciprocal model, and the straight line model, will only be briefly discussed and examples presented. The key statistical elememt is the test of hypothesis for the 'slope' parameter. This is the justification for the claim of a measurable association between problems and year.

The five sets of data that were used as examples of fitting an exponential model follow on the next several pages.



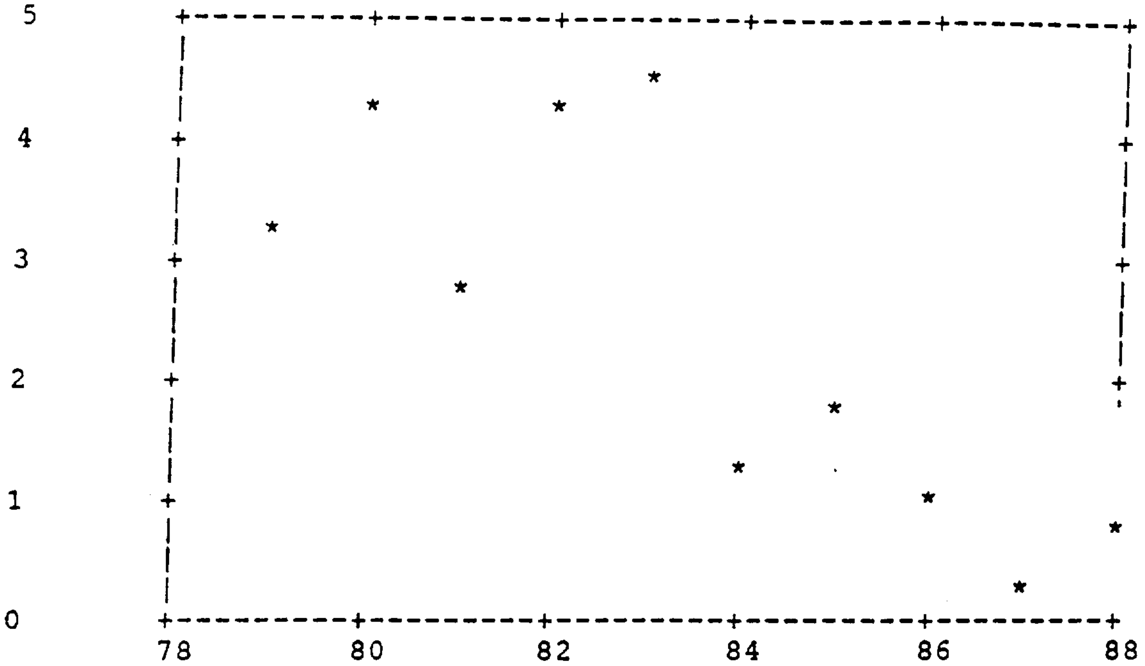
```
Regression Analysis - Exponential model: Y = exp(a+bX)
------------------------------------------------------------------------------
Dependent variable: 112.2068 78.64413 29 Independent variable: 83 84 85 86 87 88
------------------------------------------------------------------------------
                                Standard            T              Prob.
Parameter         Estimate       Error            Value            Level
------------------------------------------------------------------------------
Intercept          38.9737      6.88526          5.66046          .00480
Slope            -0.413287    0.0805133         -5.13315          .00682
------------------------------------------------------------------------------

                              Analysis of Variance
------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio  Prob. Level
Model                     2.989101      1     2.989101   26.34922       .00682
Error                     .4537670      4     .1134418
------------------------------------------------------------------------------
Total (Corr.)            3.4428682      5

Correlation Coefficient = -0.931773          R-squared = 86.82 percent
Stnd. Error of Est. = 0.336811
```

fuel preburner injector:contamination



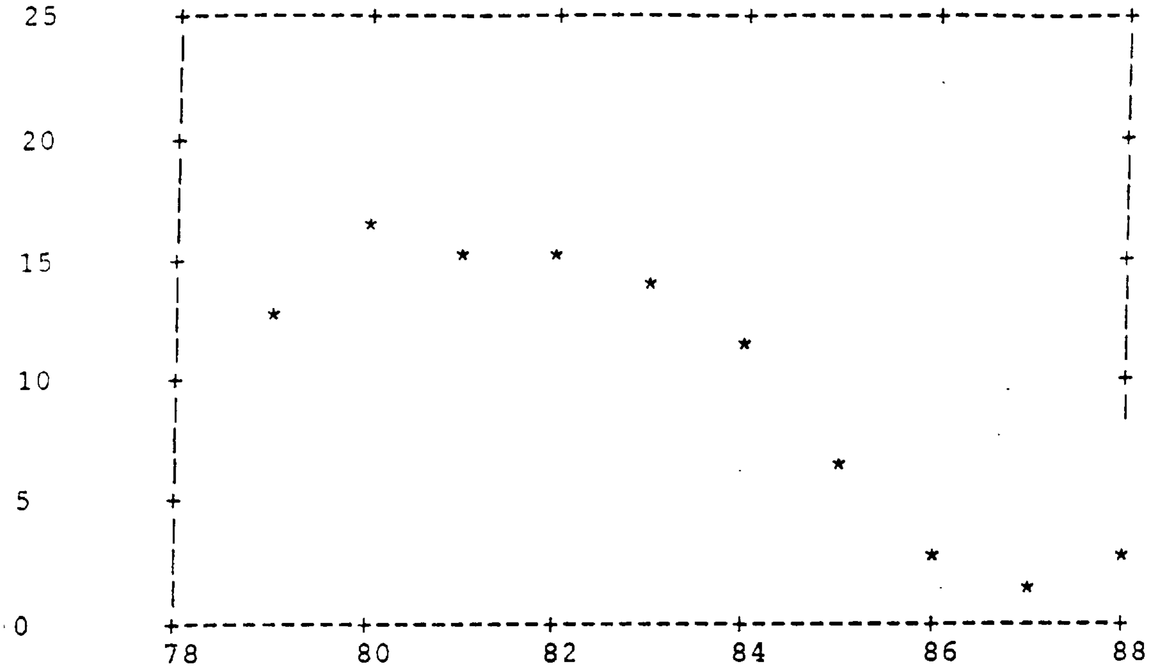
```
Regression Analysis - Exponential model: Y = exp(a+bX)
-----------------------------------------------------------------------------------
Dependent variable: 3.114 2.2049 1.0454  Independent variable: 81 82 83 84 85 86
-----------------------------------------------------------------------------------
                                 Standard              T                Prob.
Parameter          Estimate        Error             Value              Level
-----------------------------------------------------------------------------------
Intercept           37.4537        5.74436          6.52008             .00062
Slope              -0.44566      0.0679556          -6.5581             .00060
-----------------------------------------------------------------------------------

                               Analysis of Variance
-----------------------------------------------------------------------------------
Source                     Sum of Squares     Df  Mean Square    F-Ratio  Prob. Level
Model                            8.341733      1     8.341733   43.00874       .00060
Error                           1.1637263      6     .1939544
-----------------------------------------------------------------------------------
Total (Corr.)                   9.5054592      7

Correlation Coefficient = -0.936789           R-squared = 87.76 percent
Stnd. Error of Est. = 0.440403
```

fuel preburner injector:dent/crack

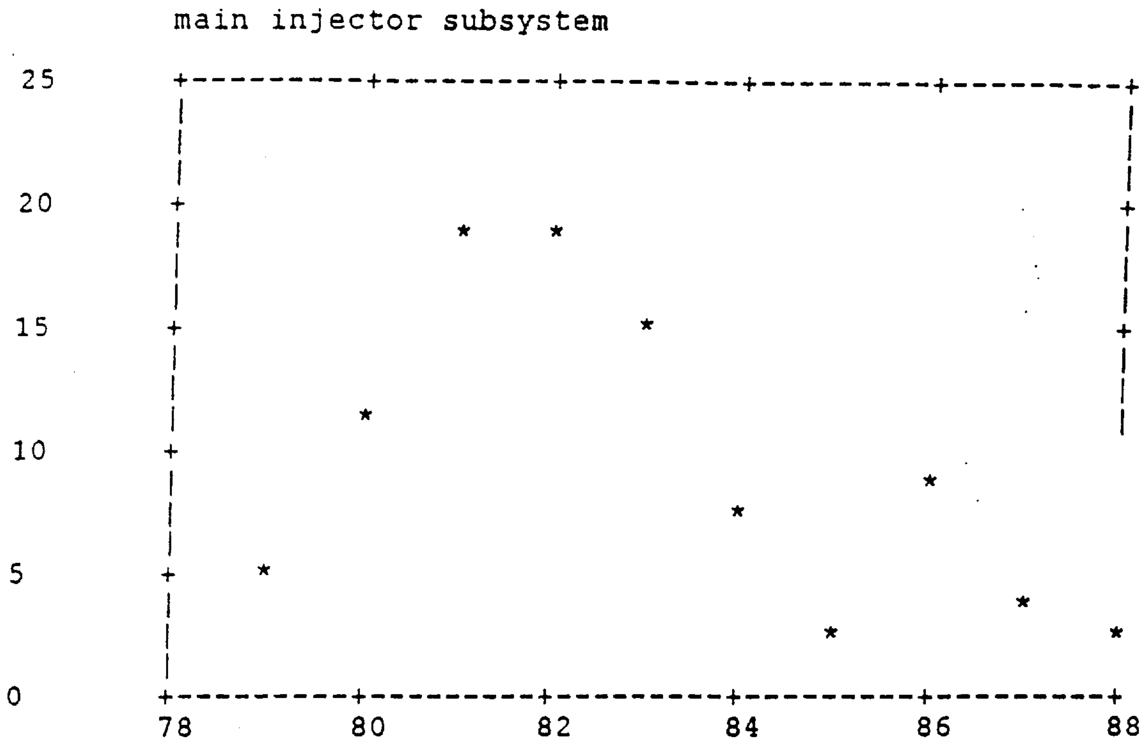


```
Regression Analysis - Exponential model: Y = exp(a+bX)
----------------------------------------------------------------------------------
Dependent variable: 4.134282 4.53009 1.2 Independent variable: 82 83 84 85 86 87
----------------------------------------------------------------------------------
                                Standard            T               Prob.
Parameter         Estimate        Error           Value             Level
----------------------------------------------------------------------------------
Intercept          35.4645       9.73999         3.64112            .01489
Slope             -0.41411       0.114556       -3.6149             .01530
----------------------------------------------------------------------------------

                           Analysis of Variance
----------------------------------------------------------------------------------
Source               Sum of Squares     Df  Mean Square    F-Ratio  Prob. Level
Model                      4.801630      1     4.801630    13.06749        .01530
Error                      1.8372425     5     .3674485
----------------------------------------------------------------------------------
Total (Corr.)              6.6388723     6

Correlation Coefficient = -0.850447            R-squared = 72.33 percent
Stnd. Error of Est. = 0.606175
```

fuel preburner subsystem



```
Regression Analysis - Exponential model: Y = exp(a+bX)
--------------------------------------------------------------------------------
Dependent variable: 15.43465 13.2418 11. Independent variable: 82 83 84 85 86 87
--------------------------------------------------------------------------------
                                 Standard            T               Prob.
Parameter          Estimate        Error           Value             Level
--------------------------------------------------------------------------------
Intercept           42.9973        9.20094          4.67314          .00547
Slope             -0.487541       0.108216         -4.50524          .00637
--------------------------------------------------------------------------------

                              Analysis of Variance
--------------------------------------------------------------------------------
Source                   Sum of Squares      Df  Mean Square    F-Ratio  Prob. Level
Model                          6.655493       1     6.655493   20.29720       .00637
Error                         1.6395101       5     .3279020
--------------------------------------------------------------------------------
Total (Corr.)                 8.2950031       6

Correlation Coefficient = -0.89574          R-squared = 80.23 percent
Stnd. Error of Est. = 0.572627
```



```
Regression Analysis - Exponential model: Y = exp(a+bX)
------------------------------------------------------------------------------
Dependent variable: 18.74208 15.33261 7. Independent variable: 82 83 84 85 86 87
------------------------------------------------------------------------------
                                Standard             T              Prob.
Parameter         Estimate       Error             Value            Level
------------------------------------------------------------------------------
Intercept          30.6808       8.45004          3.63085           .01505
Slope             -0.33945      0.0993847        -3.41551           .01893
------------------------------------------------------------------------------

                              Analysis of Variance
------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio  Prob. Level
Model                     3.226328      1     3.226328   11.66572       .01893
Error                    1.3828243      5     .2765649
------------------------------------------------------------------------------
Total (Corr.)            4.6091525      6

Correlation Coefficient = -0.83665          R-squared = 70.00 percent
Stnd. Error of Est. = 0.525894
```

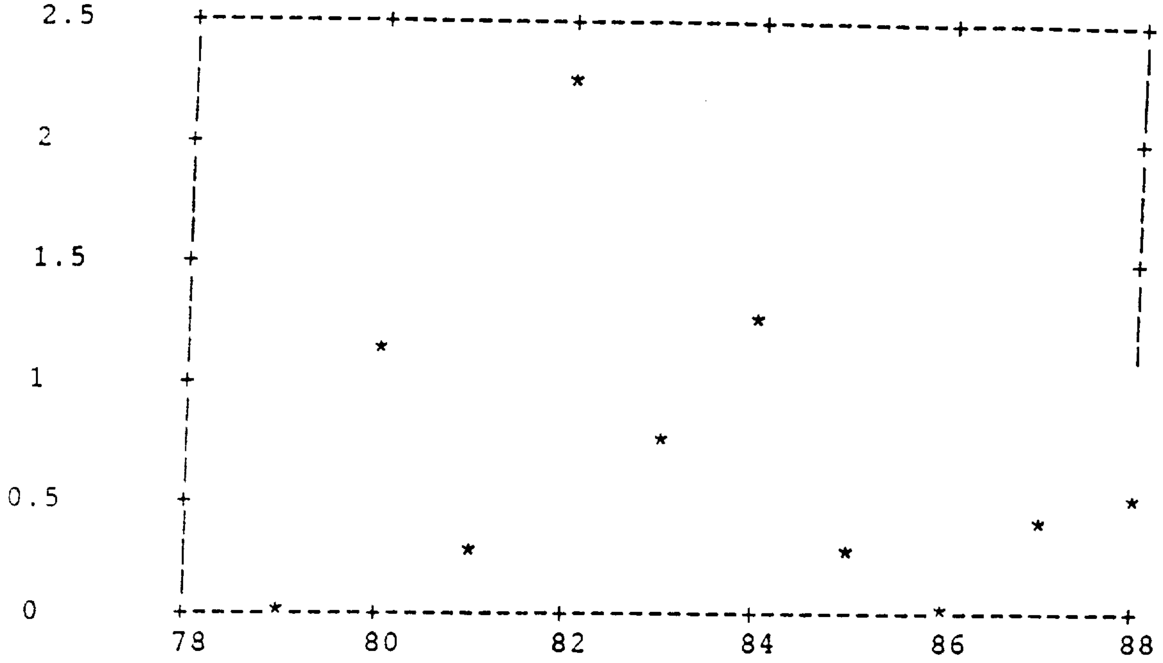
Power Model

This model is also intrinsically linear and uses the logarithmic transformation just as the exponential model. It differs in that this model uses the logarithm of both the independent and dependent variables. The regression will be for the natural logarithm of Y on the natural logarithm of time. The deterministic portion of the model is:

$$Y = \beta_0 t^{\beta_1} .$$

As with the exponential model, a positive (downward) trend is claimed if $\beta_1 < 0$. Since the logarithmic transformation is used, the same modification in the presence of no problem reports in a time period as that used with the exponential is appropriate.

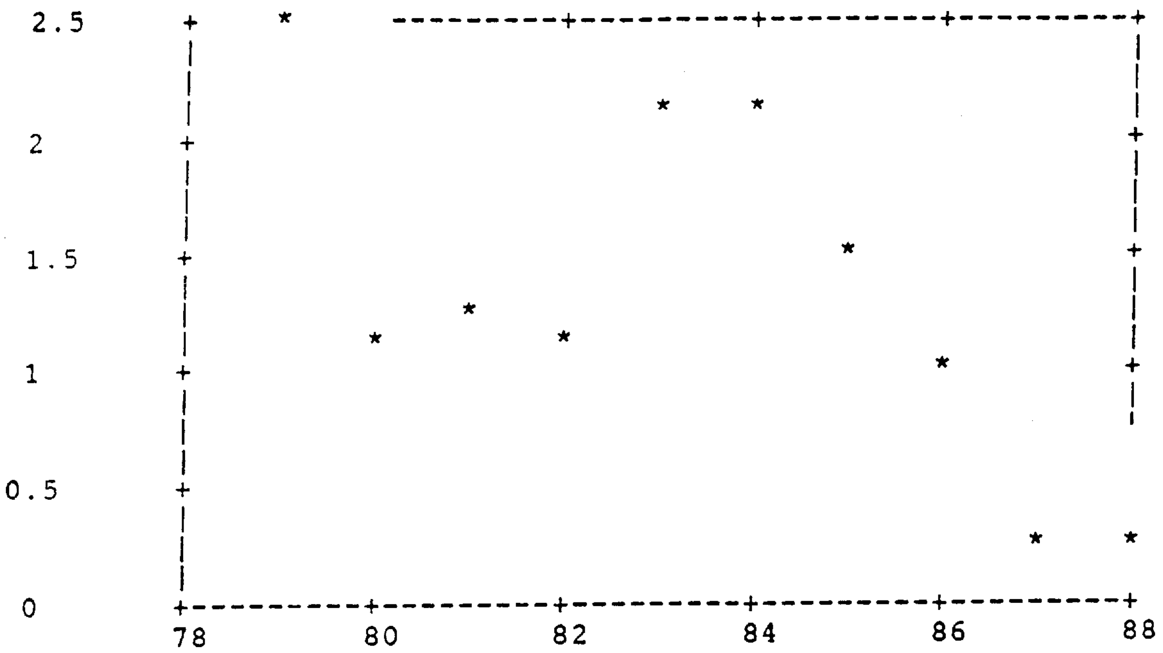
The mechanics of application are the same as with the exponential model. In the two examples below, there is some evidence of an increasing pattern followed by a decreasing trend. For the example of leakage of the HPFTP lift-off seal an adverse pattern exists from 1979 to 1983. A power (or multiplicative) model is then fit from 1982 to 1988. The coefficient of determination is .48 and, consequently, the test of hypothesis does not conclude a positive trend. Visual inspection of the scatter plot causes some concern by noticing the increase from 1986 to 1988. However, the normalized values are quite small, so there would be a good deal of (engineering) judgement in the interpretation. The second example, the main oxidizer valve subsysytem, shows an erratic pattern from 1979 to 1983 and then a decreasing pattern through 1988. There is an excellent fit for the decreasing portion of the data with strong indication of a significant downward trend (p-value less than .01).


HPFTP lift-off seal:leakage

```
Regression Analysis - Multiplicative model: Y = aX^b
------------------------------------------------------------------------------
Dependent variable: 2.2049 .6969 1.2617  Independent variable: 82 83 84 85 86 87
------------------------------------------------------------------------------
                                 Standard           T               Prob.
Parameter           Estimate        Error          Value             Level
------------------------------------------------------------------------------
Intercept*          87.6083        40.941        2.13986            .08533
Slope              -19.8174        9.2159       -2.15035            .08421
* NOTE: The Intercept is equal to Log a.
------------------------------------------------------------------------------

                           Analysis of Variance
------------------------------------------------------------------------------
Source                  Sum of Squares    Df  Mean Square    F-Ratio  Prob. Level
Model                        1.5231355     1    1.5231355   4.623994        .08421
Error                        1.6469911     5     .3293982
------------------------------------------------------------------------------
Total (Corr.)                3.1701266     6

Correlation Coefficient = -0.693156            R-squared =  48.05 percent
Stnd. Error of Est. = 0.573932
```

ORIGINAL PAGE IS
OF POOR QUALITY

main oxidizer valve subsystem



```
Regression Analysis - Multiplicative model: Y = aX^b
-------------------------------------------------------------------------------
Dependent variable: 2.0908 2.1028 1.5114 Independent variable: 83 84 85 86 87 88
-------------------------------------------------------------------------------
                                Standard             T              Prob.
Parameter          Estimate       Error            Value            Level
-------------------------------------------------------------------------------
Intercept*          196.247      40.9927          4.78737          .00873
Slope              -44.1576      9.21523         -4.79181          .00870
* NOTE: The Intercept is equal to Log a.
-------------------------------------------------------------------------------

                              Analysis of Variance
-------------------------------------------------------------------------------
Source                 Sum of Squares     Df  Mean Square    F-Ratio  Prob. Level
Model                        4.670341      1     4.670341   22.96140       .00870
Error                        .8135986      4     .2033996
-------------------------------------------------------------------------------
Total (Corr.)               5.4839396      5

Correlation Coefficient = -0.922843          R-squared =  85.16 percent
Stnd. Error of Est. = 0.450999
```

Reciprocal Model

The reciprocal model is another model with the feature that will not generate a fitted value below zero since the problem data is non-negative. It can be made linear with a simple reciprocal transformation. The deterministic portion of the model is:

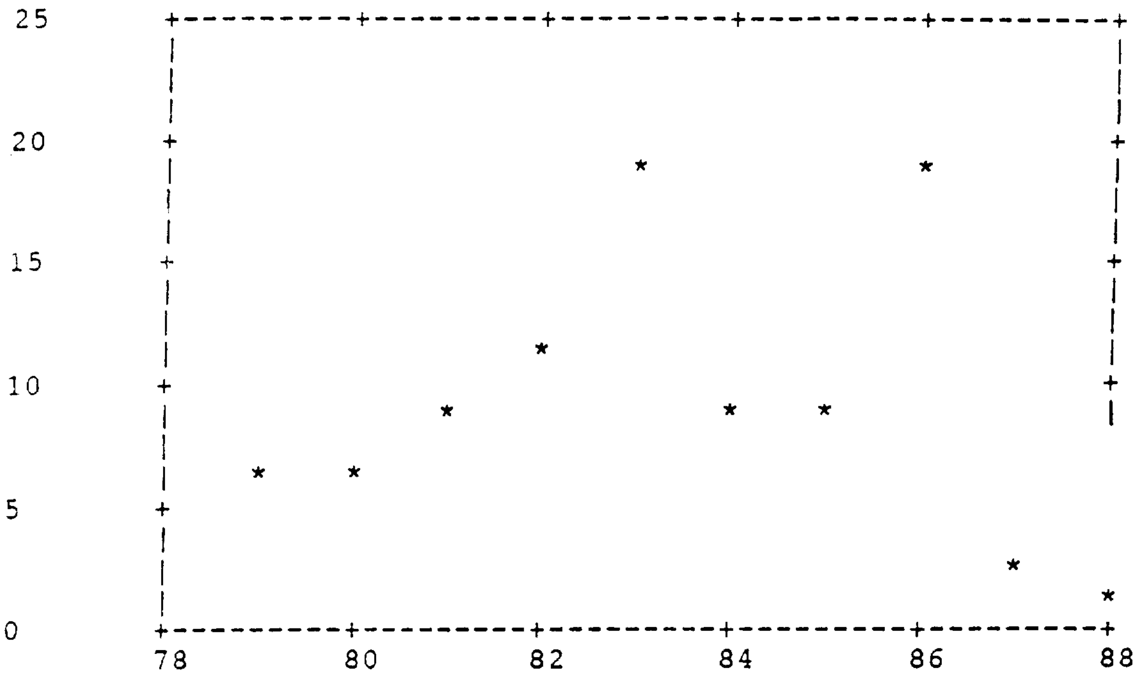
$$Y = \frac{1}{\beta_0 + \beta_1 t}$$

The application if this model regresses 1/Y on t. In contrast to the above two models, a positive (downward) trend is indicated if $\beta_1 > 0$. This model is not applicable if there are zero problem reports in a given time period.

The two examples below employ a reciprocal model. The data for the SSME combustion chamber subsystem shows again an adverse pattern followed by a positive one. There is an increasing trend from 1979 to 1983. A reciprocal model fit to the data beginning in 1983 gives marginal statistical evidence of a positive trend. The extreme value in 1986 has substantial influence on the goodness of the fit. Removing this observation and fitting again a reciprocal model beginning in 1983 the coefficient of determination, r^2, increases from .68 to .87. An extreme observation (or outlier) can dramatically effect the quality of the fitted model.

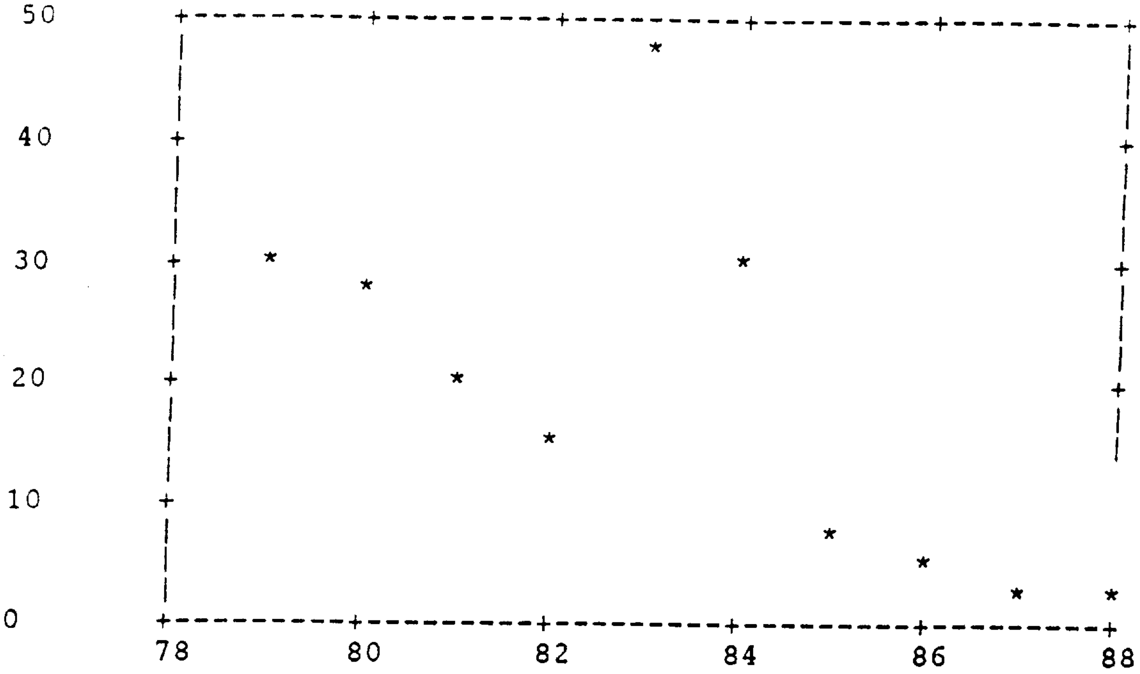
The second example, the nozzle assembly subsystem, has a pattern that appears decreasing from 1979 to 1988 with two outlying observations. There is no adverse trend in the early years. Fitting a reciprocal model beginning in 1979 yields statistical evidence of a positive (downward) trend. The p-value is less than .01. Deleting the 1983 and 1984 observations, the coefficient of determination, r^2, goes from .71 to .84 for the reciprocal model. Also given below is a regression summary of an exponential model fit to the data beginning in 1983.

main combustion chamber subsystem



```
Regression Analysis - Reciprocal model: 1/Y = a+bX
-----------------------------------------------------------------------------
Dependent variable: 18.46883 8.831693 8. Independent variable: 83 84 85 86 87 88
-----------------------------------------------------------------------------
                                 Standard            T             Prob.
Parameter        Estimate          Error           Value           Level
-----------------------------------------------------------------------------
Intercept        -9.51288         3.35553        -2.83499          .04711
Slope            0.114038       0.0392381         2.90631          .04384
-----------------------------------------------------------------------------

                          Analysis of Variance
-----------------------------------------------------------------------------
Source                 Sum of Squares     Df  Mean Square    F-Ratio  Prob. Level
Model                        .2275812      1     .2275812   8.446613       .04384
Error                        .1077739      4     .0269435
-----------------------------------------------------------------------------
Total (Corr.)                .3353551      5

Correlation Coefficient = 0.823788         R-squared =  67.86 percent
Stnd. Error of Est. = 0.164145
```

Comment: nozzle assembly subsystem



```
Regression Analysis - Reciprocal model: 1/Y = a+bX
---------------------------------------------------------------------------------
Dependent variable: 28.98 28.35 19.62 15 Independent variable: 79 80 81 82 83 84
---------------------------------------------------------------------------------
                                 Standard              T              Prob.
Parameter          Estimate        Error             Value            Level
---------------------------------------------------------------------------------
Intercept          -2.92075       0.692515          -4.2176          .00293
Slope             0.0364916     8.28869E-3          4.40258          .00228
---------------------------------------------------------------------------------

                            Analysis of Variance
---------------------------------------------------------------------------------
Source                   Sum of Squares    Df  Mean Square     F-Ratio  Prob. Level
Model                           .109860     1      .109860    19.38268       .00228
Error                          .0453435     8     .0056679
---------------------------------------------------------------------------------
Total (Corr.)                  .1552035     9

Correlation Coefficient = 0.841335            R-squared =  70.78 percent
Stnd. Error of Est. = 0.0752857


Regression Analysis - Exponential model: Y = exp(a+bX)
---------------------------------------------------------------------------------
Dependent variable: 46.69478 30.70065 7. Independent variable: 83 84 85 86 87 88
---------------------------------------------------------------------------------
                                 Standard              T              Prob.
Parameter          Estimate        Error             Value            Level
---------------------------------------------------------------------------------
Intercept           54.7792        7.47381           7.32949          .00184
Slope             -0.615199      0.0873955          -7.03925          .00215
---------------------------------------------------------------------------------

                            Analysis of Variance
---------------------------------------------------------------------------------
Source                   Sum of Squares    Df  Mean Square     F-Ratio  Prob. Level
Model                          6.623225     1     6.623225    49.55105       .00215
Error                          .5346587     4     .1336647
---------------------------------------------------------------------------------
Total (Corr.)                 7.1578840     5

Correlation Coefficient = -0.961928           R-squared =  92.53 percent
Stnd. Error of Est. = 0.365602
```

ORIGINAL PAGE IS OF POOR QUALITY

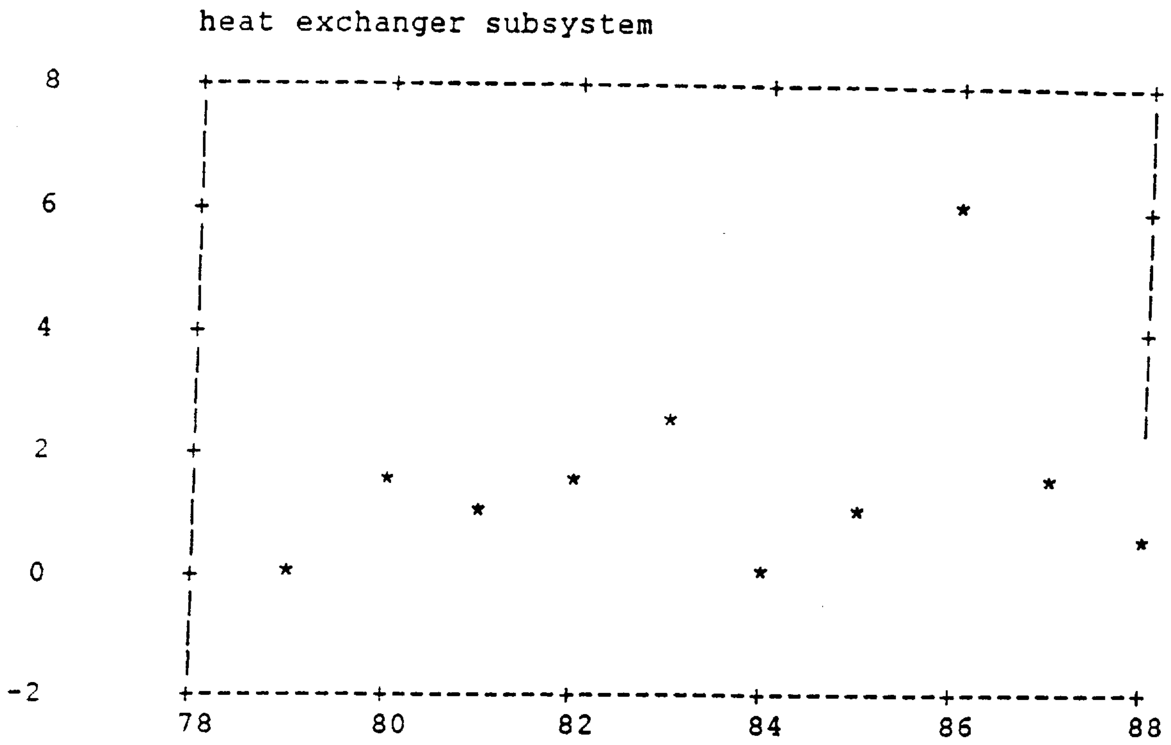
Straight Line Model

The straight line model, even though simple, can be useful and informative in a couple of situations. In instances of an increasing trend the straight line model is effective in providing statistical support. For example, a straight line model applied to the early years data (1979-1983) of the heat exchanger subsystem indicates a significant increasing trend. The scatter plot for intermittent sparking in the igniters subsysytem also shows a hint of an adverse pattern for 1979 to 1982. Refer to the scatter plots and regression summaries below to observe this. Again we see that curious pattern of increasing measurements through the first few years.

A second situation arises when the data appear to be random, but, the level of the data is close to zero. A straight line fit with both slope and intercept zero could then be viewed as desirable. The regression summary contains calculations for the test of hypothesis that the intercept is zero, as well as doing so for the slope. So, when the data are close to zero, a random scattering might indicate a positive situation. A great deal of engineering assessment and judgement is required.

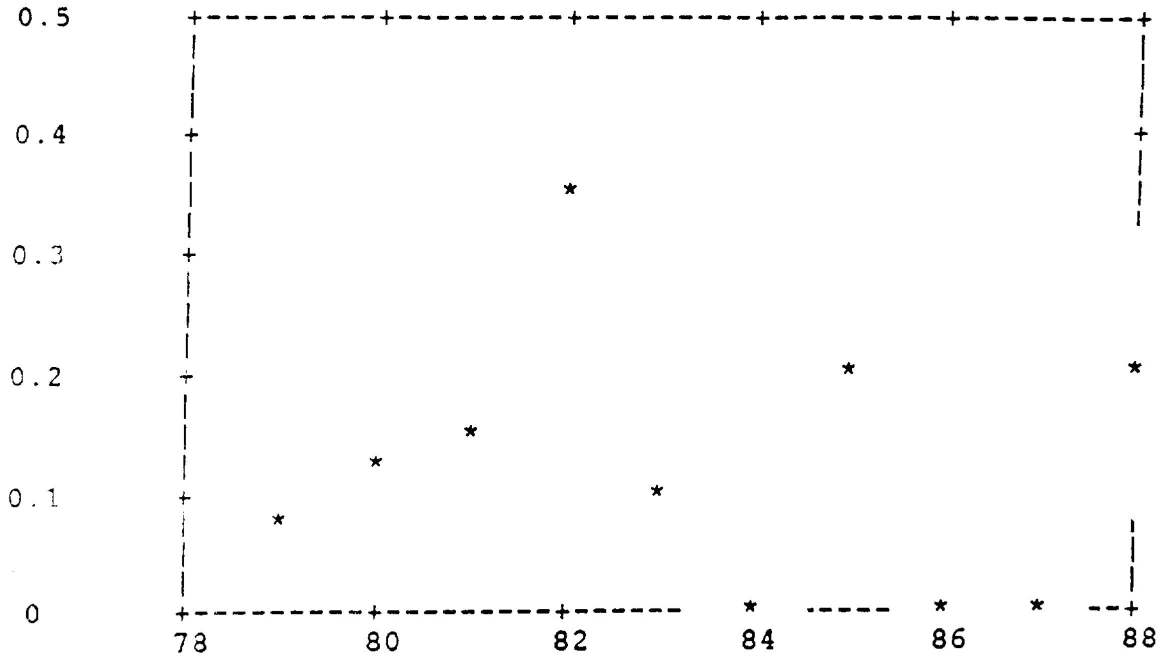
Referring back to the intermittent sparking in the igniters subsystem, the regression summary is for a straight line fit from 1983 to 1988. Note that the r^2 value is very low, .035, indicating a random pattern. The test of hypotheses that both the slope and intercept are zero would be accepted. This is seen with a large p-value for both tests. The two scatter plots following the igniters subsystem data summary have the appearance of a random scattering of the data. The regression summaries for a straight line fit beginning in 1979 for both the oxidizer preburner erosion data and the combustion subsystem leakage data had a slope and intercept that were not measurably different from zero.

A data summary for a broken main fuel valve is also given. This data set indicates a possible decreasing trend or at least, if the first one or two observations are deleted, a flat line through the remaining data. Note again that the straight line model for the data from 1981 to 1988 would conclude no measurable deviation from zero for both the slope and intercept. An exponential model fit to the full range of the data points to marginal acceptance of a positive trend, as seen in the corresponding regression summary.



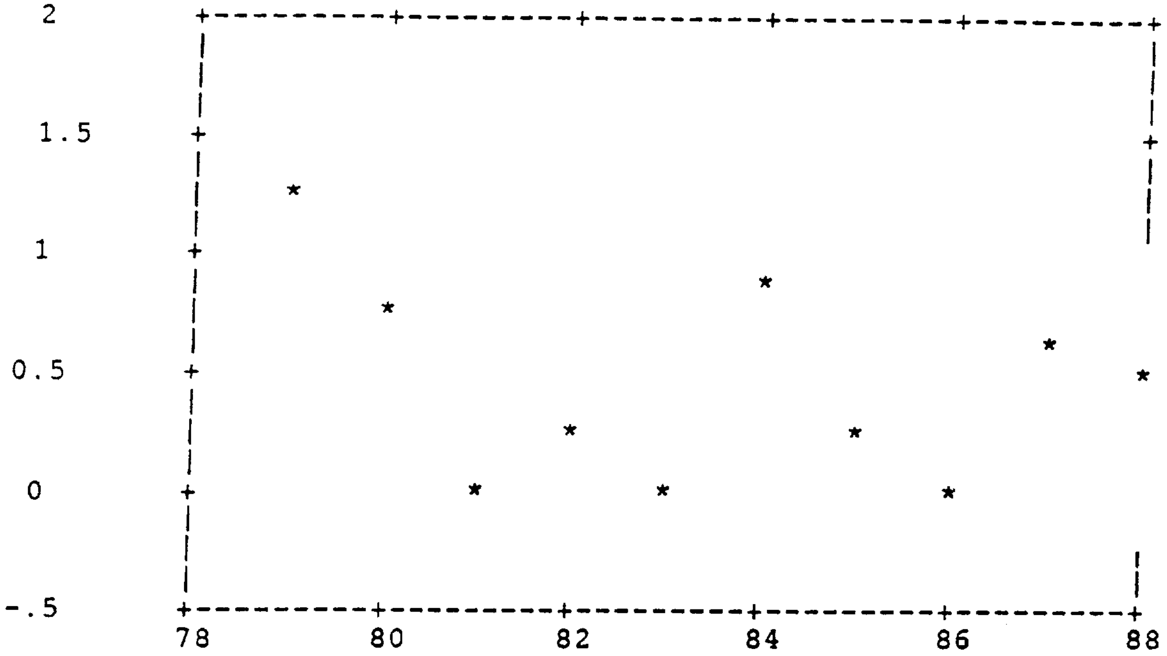
```
Regression Analysis - Linear model: Y = a+bX
-----------------------------------------------------------------------------
Dependent variable: .21 1.405811 .934201    Independent variable: 79 80 81 82 83
-----------------------------------------------------------------------------
                               Standard            T               Prob.
Parameter        Estimate        Error            Value            Level
-----------------------------------------------------------------------------
Intercept        -36.7937       10.8388         -3.39463          .04263
Slope            0.470646       0.133792         3.51775          .03898
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square     F-Ratio  Prob. Level
Model                     2.215077      1     2.215077    12.37455       .03898
Error                     .5370077      3     .1790026
-----------------------------------------------------------------------------
Total (Corr.)            2.7520843      4

Correlation Coefficient = 0.897147          R-squared =  80.49 percent
Stnd. Error of Est. = 0.423087
```

igniters subsystem:intermittent sparking RTV voids

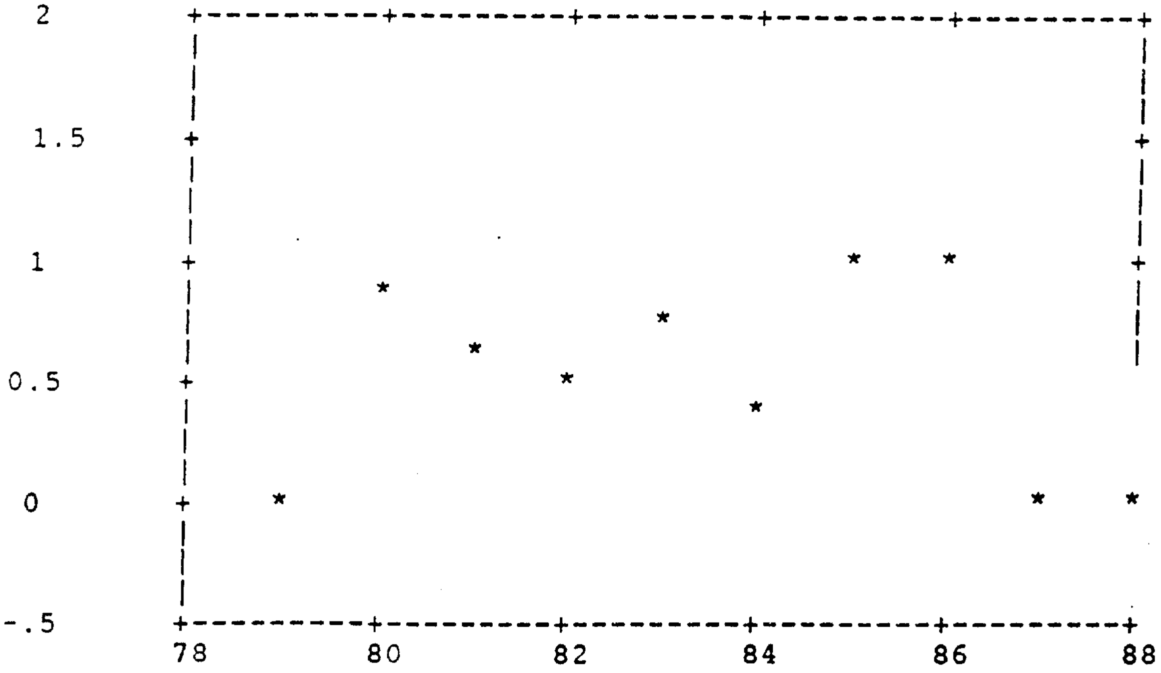


```
Regression Analysis - Linear model: Y = a+bX
-------------------------------------------------------------------------------
Dependent variable: .09 0 .189 0 0 .194  Independent variable: 83 84 85 86 87 88
-------------------------------------------------------------------------------
                                 Standard           T               Prob.
Parameter        Estimate          Error           Value            Level
-------------------------------------------------------------------------------
Intercept       -0.729752         2.10973       -0.345898          .74685
Slope          9.45714E-3       0.0246703        0.383341          .72097
-------------------------------------------------------------------------------

                            Analysis of Variance
-------------------------------------------------------------------------------
Source                Sum of Squares     Df  Mean Square    F-Ratio  Prob. Level
Model                       .0015652      1     .0015652    .146950       .72097
Error                       .0426037      4     .0106509
-------------------------------------------------------------------------------
Total (Corr.)               .0441688      5

Correlation Coefficient = 0.188244           R-squared =    3.54 percent
Stnd. Error of Est. = 0.103203
```

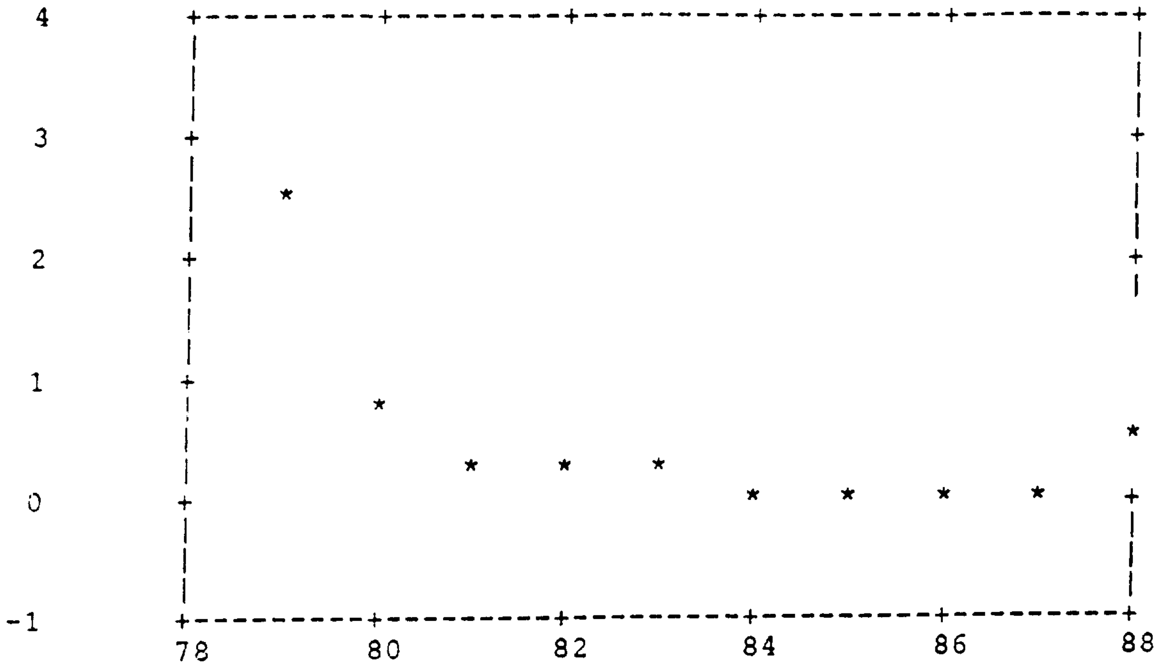
oxidizer preburner:erosion



main combustion chamber subsystem:leakage



main fuel valve subsystem:broken



```
Regression Analysis - Linear model: Y = a+bX
------------------------------------------------------------------------------
Dependent variable: .3114 .2756 .3485 0  Independent variable: 81 82 83 84 85 86
------------------------------------------------------------------------------
                                Standard            T               Prob.
Parameter        Estimate         Error           Value             Level
------------------------------------------------------------------------------
Intercept          1.4636        2.72275        0.537547           .61023
Slope          -0.0152345        0.03221       -0.472974           .65295
------------------------------------------------------------------------------

                          Analysis of Variance
------------------------------------------------------------------------------
Source                Sum of Squares     Df  Mean Square    F-Ratio  Prob. Level
Model                       .0097478      1     .0097478    .223705       .65295
Error                       .2614465      6     .0435744
------------------------------------------------------------------------------
Total (Corr.)               .2711943      7

Correlation Coefficient = -0.189589          R-squared =    3.59 percent
Stnd. Error of Est. = 0.208745

Regression Analysis - Exponential model: Y = exp(a+bX)
------------------------------------------------------------------------------
Dependent variable: 2.52 .7 .31 .28 .35  Independent variable: 79 80 81 82 83 84
------------------------------------------------------------------------------
                                Standard            T               Prob.
Parameter        Estimate         Error           Value             Level
------------------------------------------------------------------------------
Intercept         13.3621        7.11328         1.87847           .09713
Slope           -0.172212      0.0851386        -2.02272           .07773
------------------------------------------------------------------------------

                          Analysis of Variance
------------------------------------------------------------------------------
Source                Sum of Squares     Df  Mean Square    F-Ratio  Prob. Level
Model                      2.4466945      1    2.4466945   4.091409       .07773
Error                      4.7840627      8     .5980078
------------------------------------------------------------------------------
Total (Corr.)              7.2307572      9

Correlation Coefficient = -0.581699          R-squared =   33.84 percent
Stnd. Error of Est. = 0.77331
```

Kendall's τ

Kendall's rank correlation coefficient is a most efficient distribution free or nonparametric measure to test for linear trend. In applying Kendall's τ we will be testing a series of data for randomness, the null hypothesis, against a decreasing trend, the alternative hypothesis. Application of this procedure involves comparing the values of the time series in terms of larger or smaller. Whereas a parametric procedure, like regression, uses the recorded values in computations, the nonparametric approach only notes greater than or less than from the observed values and, hence, is not influenced by extreme values.

Given a series $y_1, y_2, \ldots, y_n$, let us count the number of cases in which $y_j > y_i$ for $j > i$. Call this number P. There are n(n-1)/2 pairs for comparison. The expected number in a random series is n(n-1)/4. The excess of P over this number, if significant, suggests a rising trend; a deficiency suggests a falling trend. The rank correlation coefficient, known as Kendall's τ , is then:

$$\tau = \frac{4P}{n(n-1)} - 1 \ .$$

This coefficient may vary from -1 to +1. Its expected value in a random series is zero, and its variance is given by:

$$\text{var}\ \tau = \frac{2(2n+5)}{9n(n-1)} \ .$$

In working with the normalized problem data there can be one or more values in the series that equal zero. The only possible occurence of equal values are at zero. The above results for the calculation of τ and for the variance are based on no tied values in the data. However, it is appropriate to use these computations if we regard successive zero values as a continuing positive trend. There are adjustments to the variance computation for tied values; but, with our only possible tied values at zero, we will count successive zero values in favor of a positive (downward) trend. So, the value for P will be tabulated as described in the above paragraph where n is the number of values in the series. We will not disregard multiple values of zero, but regard them as desirable.

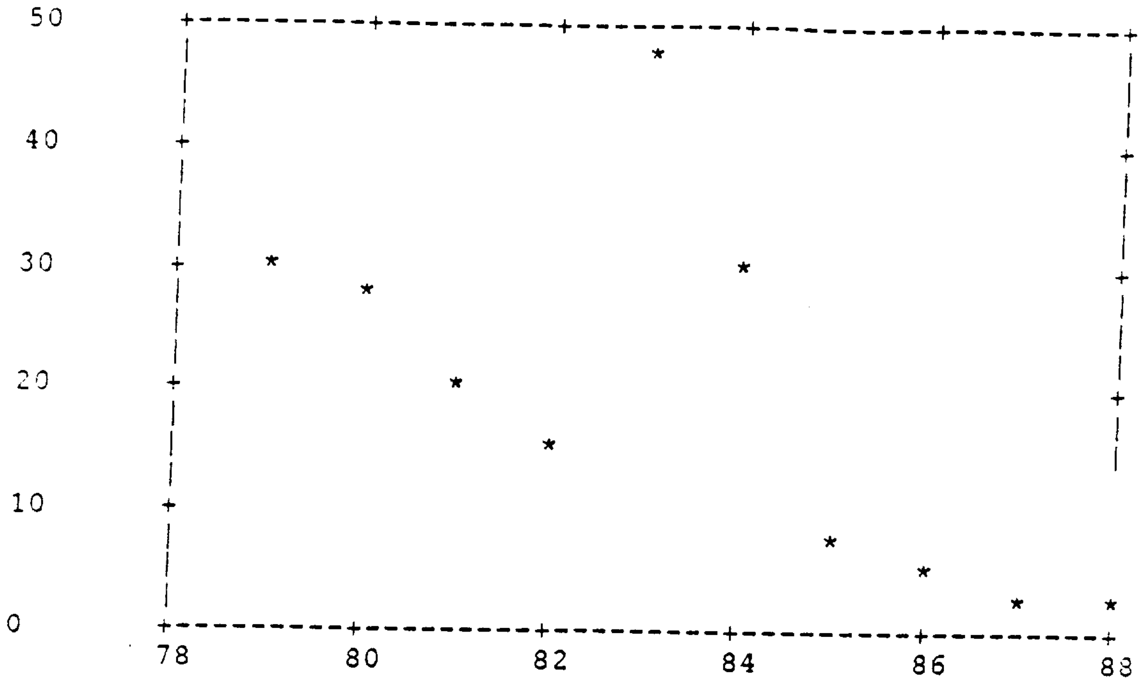
The distribution of τ tends rapidly to normality. Hence, the test statistic under the hypothesis of randomness will be a standard normal variable. If n is less than 10, the calculation of τ will use 4(P+1) instead of 4P. This is a continuity correction factor used when testing randomness versus decreasing trend. The calculated statistic will then be:

$$Z = \frac{\tau}{\sqrt{\mathrm{var}\ \tau}}$$

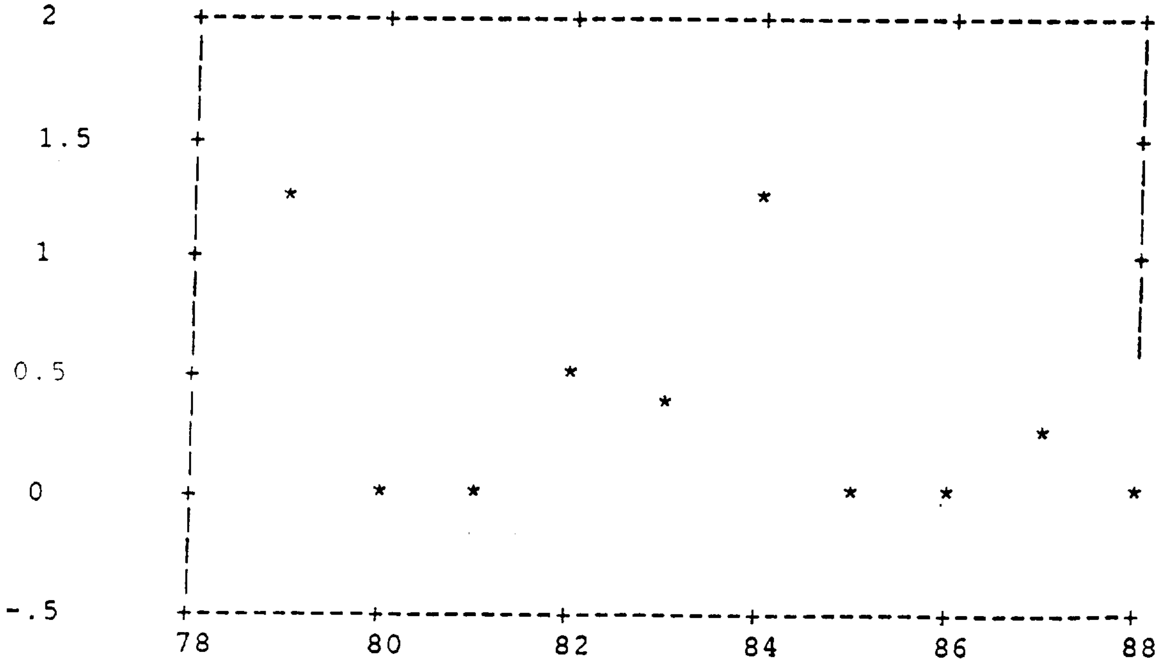
where Z represents a standard normal variable. If this value is less than -2.33 that will be evidence of a decreasing trend. This is a significance level of .01.

Let's apply this procedure to the nozzle subsystem data and to the oxidizer preburner contamination data. The scatter plots are seen below. For the nozzle subsystem, which was previously modeled with a reciprocal model, the calculated value for τ is -.644, and the calculated value for the Z statistic is -2.59. This corresponds to our previous analysis which concluded a significant downward trend. The p-value for this test is less than .01. The scatter plot for the oxidizer preburner contamination reveals more of a random scattering than decreasing pattern. Note that there are five zero values. Successive zero values are regarded as positive (downward) in our application of Kendall's τ. The calculated value of τ is -.422 and the calculated Z is -1.70, not evidence enough of a positive trend. If you cover up the 1984 observation, there is some hint of a trend. Kendall's τ is one of the, if not the, most efficient distribution free approaches for detecting trend.

nozzle assembly subsystem



oxidizer preburner:contamination



REMARKS

The foregoing examples are enough to show that trend fitting and trend estimation are very far from being a purely mechanical process which can be handed over regardless to an electronic computer. There is great scope, even a necessity, for personal judgement. To a scientist it is felt as a departure from correctness to incorporate subjective elements into his work. The student of time series cannot be a purist in that sense. What he can do, of course, is to make available the data on which he worked and explain unambiguously how he has treated them.

There are other approaches and considerations, such as increasing the sample sizes with semiannual or quarterly problem reports and time to failure patterns, that could be investigated. It is possible that some experimental design ideas could be used to identify significant factors in problem reporting. Further exploration of the data bases for problem reporting to address these and other issues seems to be the next step in developing trending approaches.

REFERENCES

1. Kendall, M. G., Rank Correlation Methods, 4th Edition, Charles Griffin & Company Ltd. (1975) London
2. Neter, J., Wasserman, W., Kutner, M. H., Applied Linear Statistical Models, Second Edition, Richard D. Irwin Inc. (1985) Homewood, Ill.
3. NSTS Problem Trending Special Study C43a-3, Calspan Corp., June 30, 1989

APPENDIX

The logarithmic transformation used with the exponential and power models needs modification because it is undefined at zero. If R denotes the number of problem reports in a given time period and m denotes the number of seconds of engine firing, one plausible way of modifying the transformation is to define a transform as

$$\ln\left(\frac{R+a}{m} * 10^4\right)$$

We would then choose the constant a so that the expected value of the above quantity is as nearly as possible $\ln(\theta * 10^4)$ where θ denotes the true fraction of problems per second.

Write $R = m\theta + Z\sqrt{m}$, where

$$E(Z) = 0 \quad , \quad E(Z^2) = \theta(1-\theta) \, ,$$

E is the expected value operator and Z is of order one in probability as $m \longrightarrow \infty$. Then

$$E\left\{\ln\left(\frac{R+a}{m} * 10^4\right)\right\} - \ln(\theta * 10^4) = \ln\left\{1 + \frac{(Z\sqrt{m} + a)^2}{2m^2\theta^2}\right\}$$

$$\doteq \frac{a - \frac{1}{2}(1-\theta)}{m\theta}$$

where we have neglected terms of smaller order than 1/m in probability. As $\theta \longrightarrow 0$ and $m \longrightarrow \infty$ in such a way that $m\theta$ remains constant the above quantity is zero as a approaches 1/2.