

STATISTICAL GRAPHICS

D. A. Dickey, North Carolina State University, Raleigh, North Carolina

ABSTRACT

Some graphs that convey information of statistical interest are displayed and discussed. The logic of their construction is also discussed.

INTRODUCTION

This paper shows how to produce some informative statistical displays using some simple graphics available in SAS/STAT (TM) and some more complex graphics available in SAS/GRAPH. Some of the more interesting logic involved is described and the advantages of having powerful statistical analysis tools and graphics within the same package should be apparent to the reader. Some graphs display data, some are used as analysis tools, and some as vehicles for understanding the underlying statistical concepts.

DANIEL'S NORMAL PLOT

Often in industrial applications, one has several factors that may contribute to a response. It is common practice to run a screening design, such as a Plackett-Burman design (Box, Hunter and Hunter, 1978, pg. 398). In such a design, high and low levels of each experimental factor are identified and experimental trials run with each factor set at either its high or low value. For an experiment with k factors, it would take 2^k runs to observe every possible combination. To investigate the effect of wheel speed, grit size, wheel diameter, lubricant amount, feed rate, and depth of cut on the smoothness of a part being shaped by an industrial grinder would require $2^6=64$ trials. This would be called a full factorial experiment. To cut down on the necessary number of runs, various screening designs have been proposed, consisting of a small subset of the 2^k potential trials. The subset must be picked carefully in order not to lose important information. If enough care is taken, very nice statistical properties can result.

Plackett and Burman gave tables of designs, consisting of 1's and -1's for each variable and each run, indicating subsets of treatments that have such nice properties. Here is some data associated with such an experiment:

DATA	X1	X2	X3	X4	X5	X6	X7
82	1	-1	-1	1	-1	1	1
106	1	1	-1	-1	1	-1	1
122	1	1	1	-1	-1	1	-1
121	-1	1	1	1	-1	-1	1
88	1	-1	1	1	1	-1	-1
116	-1	1	-1	1	1	1	-1
87	-1	-1	1	-1	1	1	1
88	-1	-1	-1	-1	-1	-1	-1

You see that factor X1 (perhaps Ph) was high on the first trial while X2 (perhaps temperature) was low and so forth. The response on this trial was 82 which might be yield of a drug, hardness of metal, etc. Note that each column has 4 1's and 4 -1's and that the columns are orthogonal. This implies that if you regress the data on X1-X7 each coefficient will be half the difference of two means: that of the +1 data minus that of the -1 data. This difference in means is referred to as the "effect" of the variable so the effect of X1 would be $(82+106+122+88)/4 - (121+116+87+88)/4 = -3.5$ indicating that a switch from low to high level of X1 would decrease yield by 3.5.

Table 1 shows parameter estimates that result from PROC REG:

Table 1
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error
INTERCEP	1	101.250000	.
X1	1	-1.750000	.
X2	1	15.000000	.
X3	1	3.250000	.
X4	1	0.500000	.
X5	1	-2.000000	.
X6	1	0.500000	.
X7	1	-2.250000	.

So you have a nice design that can be analyzed by regression but there remains a problem - your regression has 0 error degrees of freedom. You will have no error term by which to judge the statistical significance of your results! Cathbert Daniel (1959) noted that any factor which has a theoretical 0 effect will give a regression coefficient that is normally distributed about 0. The variance of it will be just the population error variance divided by the sample size n , $n=8$ in this case, because the XX matrix is clearly a diagonal matrix with n on the diagonal. Thus assuming there are several inert factors among those being tested, there will be several coefficients which appear to be from the same normal distribution and a normal probability plot of the regression parameter estimates, such as that given by PROC UNIVARIATE or using PROC RANK, will show the inert factors in a line and the active factors off the line.

To get this normal plot, just use the OUTEST data set available in PROC REG. PROC RANK and a subsequent PROC GPLOT applied to the OUTEST data set produce a normal probability plot of the 7 coefficients as shown in Fig. 1. Daniel's idea was to estimate the standard error from the normal plot. The straight line referred to in the above paragraph goes through the origin and has a slope equal to the standard error of a regression coefficient. A suggested line having a slope 1.6 is shown in Fig. 1. Points off the line represent significant effects and clearly the largest point is well off the line. Checking the coefficients above, this is associated with factor X_2 . The other factors seem to be inert.

Once you have some estimate of the standard error of each coefficient, division of the coefficient of an inert factor by this number will produce a ratio that should look like a value of Student's t . Thus you could compute these Studentized coefficients and see which of them looks to be out of reasonable range as a t statistic. This leads to another idea. PROC REG has many nice diagnostics. The RSTUDENT residuals, for example, are a set of Studentized residuals obtained by predicting each of the n points in a data set from the remaining $n-1$ and then computing a residual and standard error. If you simply regress your coefficients on a column of 1's, then, a plot of the RSTUDENT residuals should reveal the important effects. In Fig. 2 I use PROC TIMEPLOT to do this and with the ID statement, factor names can be associated with each effect. Obviously X_2 has a strong effect.

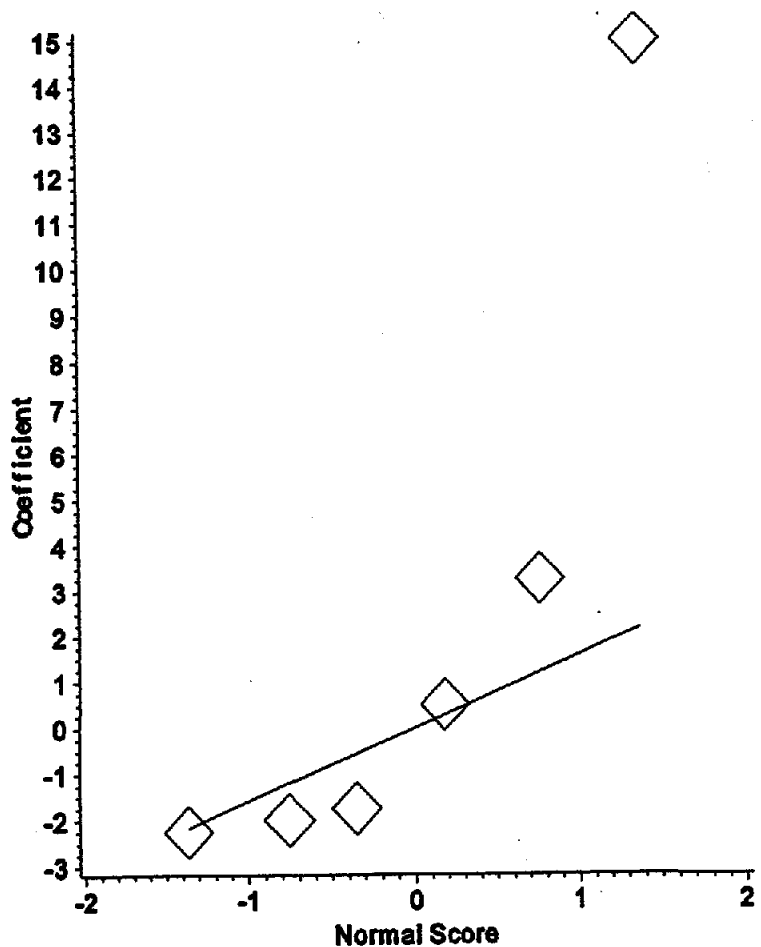


FIGURE 1: DANIEL'S NORMAL PLOT
TO GET ERROR TERM IN PLACKETT-BURMAN DESIGN

NAME OF FORMER VARIABLE	Studentized Residual without Current Obs
X1	-0.61
X2	6.65
X3	0.22
X4	-0.23
X5	-0.66
X6	-0.23
X7	-0.70

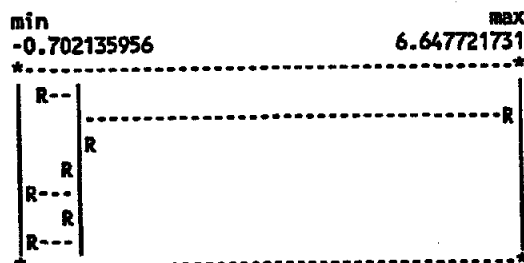


Figure 2: PROC TIMEPLOT

This idea is most effective when looking for a single active factor with the other $k-1$ factors inert as seems to be the case here. Of course PROC UNIVARIATE can be invoked on the coefficients, or their Studentized values as a check on which do not seem to be from a common normal distribution. Use the ID statement to identify factors and glance at the boxplot to see which effects are extreme.

```

Lowest ID      Highest ID
-1.41596(X7)   ) -1.1013(X1   )
-1.25863(X5)   ) 0.314658(X4   )
-1.1013(X1)    ) 0.314658(X6   )
0.314658(X6)   ) 2.04528(X3    )
0.314658(X4)   ) 9.439752(X2    )

```

UNIVARIATE PROCEDURE

Variable=SS

```

Stem Leaf      #      Boxplot
8 4            1      0
6
4
2 0            1      +-----+
0 33           2      *--*--*
-0 431         3      +-----+
-----+-----+-----+

```

MULTIPLE COMPARISONS

A lot of work has been done on so called multiple comparisons. The idea here is that you have several treatments, each replicated and thus you have several treatment means. Now any two means can be compared using a t-test. The significance level of the test, usually taken as .05, is the probability of declaring two population means to be different when in fact they are the same. With 4 means, call them A, B, C, and D, you could compare A to B, A to C, B to C, etc. for 6 comparisons in all, each having a .05 significance level or "comparisonwise error rate." What is the probability that, among the six comparisons done, at least one will falsely reject equality of means? It must certainly exceed .05 just as a person who has a .05 chance of missing a freethrow will have a much greater chance of missing at least once in six freethrows. The probability of falsely declaring two means among many to be significantly different is called the multiple comparison problem and the probability of at least one false declaration of difference in an experiment with many means is referred to as the experimentwise error rate.

While formal statistical comparison methods are available, Box, Hunter and Hunter, page 192, suggest a nice visual method - simply graph the means below a normal distribution centered on the overall average and having standard deviation equal to the standard error of a mean. Then just ask yourself if the means appear likely to have all come from this same distribution.

Figure 3 illustrates this. The data are treatment means for a rhizobium experiment reported in Steel and Torrie (1980). To get the plot, you create a data set for use in PROC GPLOT where joined line segments are used to graph the normal distribution, its baseline, and vertical shading lines for the 0.05 tail areas. Diamonds with X coordinates at the observed means and common Y coordinate locate the data just below the normal distribution. Now add a bit of rigor by taking any of the multiple comparison numbers, e.g. the Tukey comparison number as can be found on the MEANS statement in PROC GLM, and starting below each diamond, extend a line whose length is the comparison number. You see not only which means are connected, but how close they come to being significantly far apart. Suppression of axes, labels, etc. makes the graph more attractive. The normal curve is created in a data loop using the formula for the normal density.

MULTIPLE REGRESSION RESIDUALS

The data here are from a textiles experiment at North Carolina State University in which the air permeability of various fabric samples was measured as a function of the number of weft direction plies and the fabric thickness. Fig. 4 is a straightforward PROC G3D on this data. It appears that a simple linear regression of permeability on thickness and number of weft plies would suffice and indeed a nice regression is obtained from PROC GLM in Table 2.

Table 2
General Linear Models Procedure

Dependent Variable: PERM

Source	DF	Sum of Squares	Mean Squares	F Value	PR > F
Model	2	358794.2912	179397.1456	64.29	0.0001
Error	18	50224.8412	2790.2690		
Corrected Total	20	409019.1324			

R-Square	C.V.	Root MSE	PERM Mean
0.877207	19.16323	52.82300	275.647619

Parameter	Estimate	T for H0: Parameter=0	P > T	Std Error of Estimate
INTERCEPT	-157.71544	-8.24	0.0001	92.179108
WEFTPLY	124.78818	8.07	0.0001	15.464520
THICK	16184.30800	10.57	0.0001	1531.605883

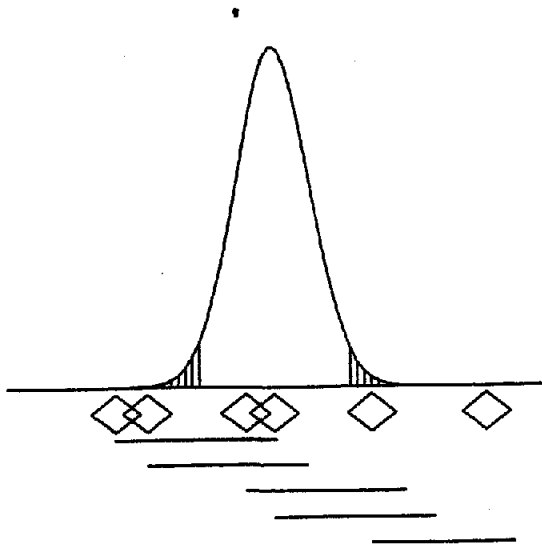


FIGURE 3: STEEL'S RHIZOBIUM DATA REFERENCED TO DISTRIBUTION OF 1 MEAN

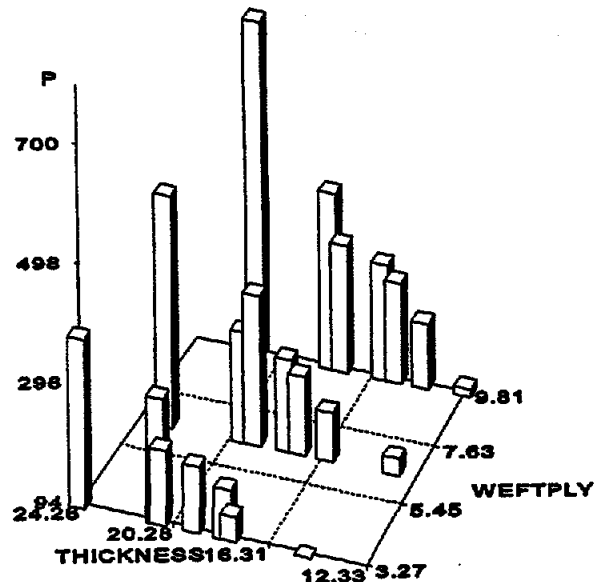


FIGURE 4: AIR PERMEABILITY FOR FABRICS

If, however, you output the residuals and plot them against the two independent variables jointly in PROC G3D a suggestion for model modification appears. The SHAPE=variable option is used in Fig. 5 to show negative residuals as diamonds and positive ones as balloons. Notice that the pattern of diamonds runs across a diagonal of the plot's floor with predominately balloons in the opposing corners. This indicates the need for a quadratic term, likely a crossproduct, in the model. The effect would be less apparent had you only used a 2 dimensional plot. Indeed a full quadratic model produces additional statistically significant terms. This data will be discussed further.

SELECTING A TRANSFORMATION

Box, Hunter and Hunter (1978) discuss transformation and specifically, a technique developed by Box and Cox (1964). The idea here is to transform Y by various power transformations such as the square root, $Y^{0.5}$ and reciprocal Y^{-1} . You model these transformed Y's, searching over a grid of powers to find the power that minimizes the error mean square. To make the comparison fair, you must scale the transformed Y's. The appropriately scaled transformed value of Y is

$$Y^*(L) = (Y^{**L} - 1)/(L * G^{**L-1}) \quad \text{for } L \text{ not } 0$$

$$Y^*(L) = G * \ln(Y) \quad \text{for } L = 0$$

where G is the geometric mean of the Y's and ln is the natural logarithm.

SAS makes it easy to perform this computation. As you read in the data, simply add this loop (after computing the geometric mean G):

```
do L = -1.5 to 1.5 by .1 ; * (choose any grid);
  if L=0 then go to one;
  Ytrans = (Y**L-1)/(L*G**(L-1));
  go to two;
  one: Ytrans = G*ln(Y);
  two: output;
end;
```

Next, run

```
proc sort; by L;
proc reg noprint outest=out1; by L;
```

to get a data set OUT1 containing RMSE, the square root of the error mean square for each L. Now just find the L giving the minimum RMSE using a plot or printing the data, for example. You can get a little fancier than this. Box, Hunter, and Hunter, page 239, consider the error sums of squares associated with each L. They suggest an approximate 95% confidence interval for L consisting of all L's for which the regression root mean square error, RMSE, is less than

FITTED SURFACE

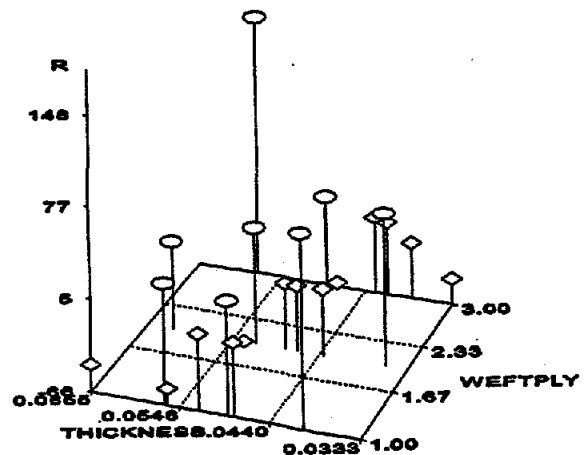


FIGURE 5: RESIDUALS FROM FIT

$$C = \min(\text{RMSE}_*) \cdot \sqrt{1 + t^*/df}$$

where $\min(\text{RMSE}_*)$ is the minimum root mean square error, t is from the Student t tables and df is the error degrees of freedom. If you plot RMSE_* versus L and put a VREF line at C , it will intersect the plotted curve in two points and thus give the upper and lower confidence limits for L . Looking at the plot, you can replot the curve with HREF lines at these two points and at the minimizing L . This is shown in Fig. 6.

Another comment about transformations is that the significance of various terms in the model may change with transformation. I fit a full quadratic model

$$Y'(L) = b_0 + b_1*U + b_2*T + b_3*T**2 + b_4*U**2 + b_5*UT$$

for each L and output the P-values that go with each coefficient. I plot these P-values against L . At the minimizing L (about .4) only one of these P-values is below 0.05. As it happens this is the crossproduct term suggested in your previous residual plot. Of course most analysts would leave the linear terms in the model too when the cross product term is significant. The untransformed Y value ($L=1$) is just outside the 95% confidence interval. Even though the minimizing L is .4, the value $L=.5$ is about as good and corresponds to the square root transformation. Finally, RMSE_* was scaled so that it fit nicely on the plot.

In Fig. 7, the information is displayed in a different way. Here the VREF lines are at the 95% t statistic cutoff points and the plotted curves are the calculated t statistics. The minimizing L with upper and lower limits are shown as in Fig. 6.

Those familiar with PROC REG will wonder how the t -statistics were produced, after all, PROC REG does not include them in the OUTEST data set. To do this, run a PROC REG on the original data and output the inverse of the $X'X$ matrix. Now for each X variable, multiply it in the dataset by the square root of the appropriate diagonal of the inverse of $X'X$. These transformed X values will produce a new inverse $X'X$ matrix with 1's now on the diagonal, implying each t test is just the (new) coefficient divided by RMSE_* . The PROBT function can be used on these to get the P-values necessary for Fig. 6.

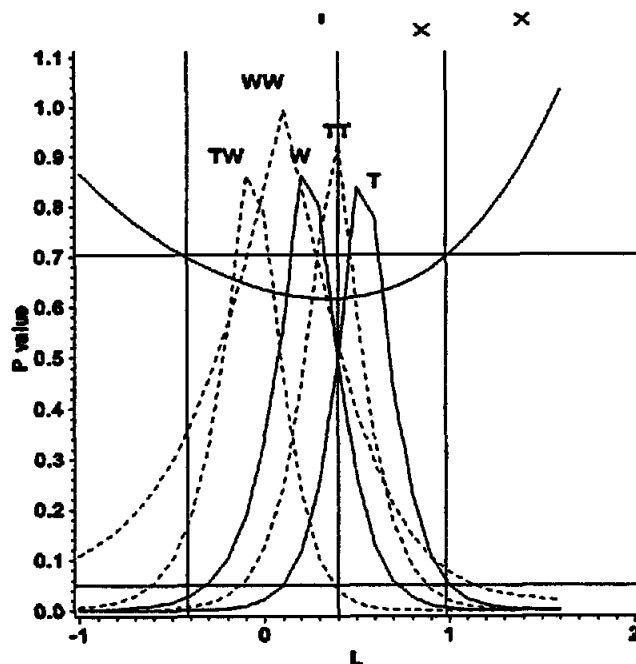


FIGURE 8: EFFECTS OF TRANSFORMATION
References at .05 and Box suggested 95% C.I. marker
Min/ $\text{RMSE}_* \sqrt{1 + t^*/df}$

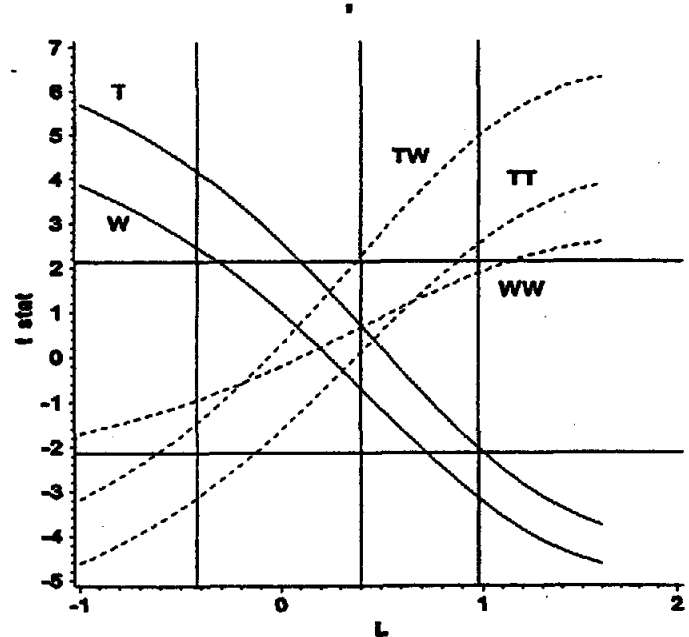


FIGURE 7: EFFECTS OF TRANSFORMATION
HREF at minimum L and 95% C.I. endpoints
VREF at critical value of t
T, W solid=linear, dot=quadratic, X=crossproduct

UNDERSTANDING A TRANSFORMATION

From Fig. 6, you see that without transformation all the terms in a full quadratic model are at least close to significance. With a square root transformation, only the crossproduct term is significant and just barely so. Evidently, a transformation can have an effect on the shape of the response function. Further, it is fairly common advice to suggest that if the variance of your residuals goes up as the predicted value does, then you might consider a transformation such as the logarithm. This is supposed to homogenize the variance, and it can also remove certain types of curvature from the plot as well. To understand the effect of a transformation, you can view Figs. 8 and 9. Here PROC G3D has been used to plot the transforming function, the logarithm in this case, and the ANNOTATE facility adds the original data, the transformed data, and fitted curves. The two figures show different ROTATE and TILT perspectives.

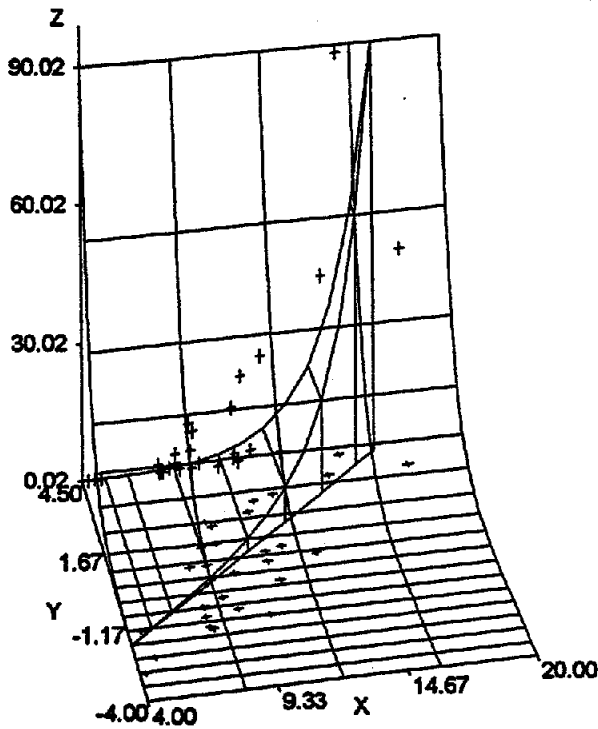


FIGURE 8: LOG TRANSFORMATION
Median estimator - Rotate=10

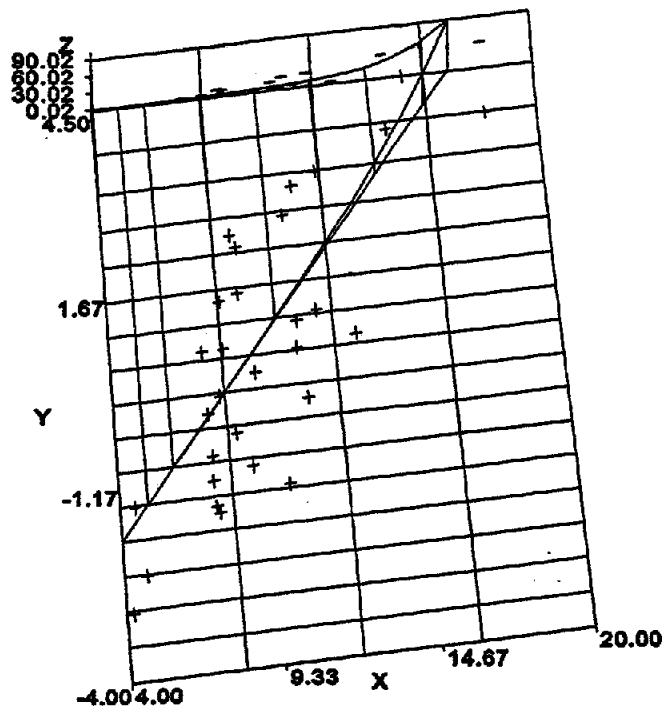


FIGURE 9: LOG TRANSFORMATION
Median estimator - Rotate=5 Tilt=5

A few comments about the code are in order. First, a small MACRO was written to form a plus sign in the floor of the plot. As each point is read into the data set, the MACRO is invoked. It simply uses MOVE and DRAW functions and small perturbations of two coordinates, X and Y, while setting the third coordinate, Z, to the minimum Z on the plot. Notice how this approach insures that the "horizontal" bar of the plus sign is really parallel to the appropriate axis, regardless of rotation and tilting. A similar MACRO plots points on the left wall.

You can see that the original data, Z versus X, plotted in the left wall, show a slight curvature and definite increase in variance as X increases. The plotted curve is $Z = \text{EXP}(Y)$ and so $Y = \ln(Z)$. Take any point on the left wall. Move horizontally, perpendicular to the wall, until you hit the curved surface. Then drop straight down to the floor. Doing this for each point yields a plot of the transformed data, $Y = \ln(X)$ versus X in the floor. The plot is more like a straight line and the variance is more uniform. Fitting a straight line to the data in the floor, you can now back transform all the points on this line to get a fitted function, a curve, on the wall. Note that some authors suggest a bias adjustment when doing this based on the fact that, if $Z = \text{EXP}(Y)$ then the mean of Z is not $\text{EXP}(\mu)$ but in fact is $\text{EXP}(\mu + .5 S^2)$ where S^2 is the error variance and μ the mean on the Y scale.

CONCLUSION

Statistical graphics are useful not only for presenting data analyses, but also for understanding statistical principles. When used in conjunction with the powerful statistical analysis tools available in SAS software, graphics can provide very informative visualization capabilities.

REFERENCES

- Box, G.E.P., W. G. Hunter and J. S. Hunter (1978). *Statistics for Experimenters*, Wiley, New York.
- Box, G.E.P., and D. R. Cox (1964). "An analysis of transformations," *J. Royal Statistical Society, Series B*, 26, 211.
- Daniel, Cuthbert (1959). "Use of half-normal plot in interpreting factorial two-level experiments," *Technometrics*, 1, 149.
- Steel, R.G.D. and J. H. Torrie (1980). *Principles and Procedures of Statistics*, McGraw Hill, San Francisco.
- SAS, SAS/GRAPH, and SAS/STAT are registered trademarks of SAS Institute Inc. in the USA and other countries. (TM) indicates USA registration.