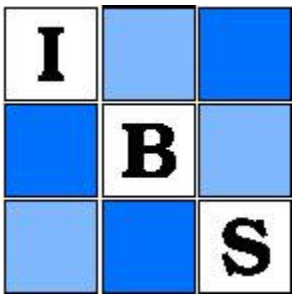

BIOMETRICS 43, 439–456
June 1987

THE CONSULTANT'S FORUM

Logarithmic Transformations in ANOVA

Donald A. Berry*

School of Statistics, 270 Vincent Hall, University of Minnesota,
Minneapolis, Minnesota 55455, U.S.A.

SUMMARY

A method is presented for choosing an additive constant c when transforming data x to $y = \log(x + c)$. The method preserves Type I error probability and power in ANOVA under the assumption that the $x + c^*$ for some c^* are log-normally distributed. The method has advantages similar to those of rank transformations—namely, it is easy to use and is resistant to extreme observations. Since the special case $c \to \infty$ corresponds in ANOVA to $y = x$, the method is a useful generalization of least squares.

1. Introduction

In many settings, statistical analyses are carried out with little opportunity for seeking out the "correct" method from the many available. For example, a pharmaceutical statistician may have to announce the method of analysis to a regulatory agency *before* the experiment is conducted.

The standard method in many circumstances is analysis of variance. It is well known that the conclusions of ANOVA can be greatly affected by outliers. Rejecting outliers will be closely scrutinized by regulatory officials: were they deleted (and others not deleted) to obtain conclusions that serve the company's best interests? And trying various models and tests to minimize the influence of unusual observations, while good statistical practice in many settings, can give reason to suspect that one's motives are less than scientific. I am not suggesting that any statistician would purposely tailor an analysis to a conclusion. But even a perfectly honest statistician may wonder whether his or her subconscious motives are purely scientific.

The approach I propose dictates standard ANOVA should the data conform to the ANOVA assumptions, but it automatically transforms the data to minimize the effect of unusual observations when such are present. The method preserves significance levels and is quite powerful. It is also robust in that it is minimally affected by extreme observations. While I do not claim that it is most robust or best in any absolute sense, I do claim that it is better than standard ANOVA. It is not as easy to use but the modifications are straightforward and, I think, intuitive.

When analyzing counts or other nonnegative data that have frequent or occasional large observations, researchers ofttimes take logarithms. When some of the observations are zero, and at other times as well, a nonnegative constant c is first added to each datum:

$$y = \log(x + c), \tag{1}$$

or perhaps only to the zeros. More generally, if all the data are positive then c is allowed to

* Research supported in part by the National Science Foundation under grant DMS 8505023.

Key words: ANOVA; Kurtosis; Least squares; Likelihood; Logarithmic transform; Log-normal distribution; Rank transformations; Robust methods; Skewness.

be negative. For example, Hill (1963) discusses a setting in which $-c$ is the incubation period of an infectious agent, so it can be natural to assume that c is negative.

I propose using the transform in (1) whenever a parametric analysis is planned. The question is how to choose c. Researchers frequently assign a value in an ad hoc fashion "to avoid taking the logarithm of 0." The choice of c greatly affects the conclusion in virtually every data set. It should depend, for example, on the scale of measurement: if c is correct when x is the count per minute then $60c$ is correct when x is the count per hour.

Model (1) is a one-parameter subfamily of model (2) of Box and Cox (1964):

$$y = \begin{cases} \dfrac{(x + c)^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x + c) & \text{if } \lambda = 0 \end{cases}.$$

Box and Cox recommend selecting a value of an unknown constant by viewing it as a parameter and using likelihood or Bayesian methods. It is clear from Hill (1963) that the Box–Cox recommendations cannot be followed in general, and not when using (1) in particular (see §4). I present a method for choosing c based on an analysis of residuals. While my discussion is limited to (1), the principles involved apply to the power transformations of Box and Cox and to other families of transformations as well. However, I have not compared the Box–Cox approach with mine in the one-parameter power transformation case that has been the focus of much research (e.g., Draper and Hunter, 1969; Andrews, 1971; Atkinson, 1973, 1985; Hinkley, 1975; Carroll, 1980; Bickel and Doksum, 1981; Hinkley and Runger, 1984; Carroll and Ruppert, 1984; Taylor, 1985).

When the data dictate the transformation used then care is required to ensure that the null distribution of the test statistic is not affected, or at least is minimally affected. This issue lies at the center of the Bickel–Doksum (1981) and Box–Cox (1964, 1982)–Hinkley–Runger (1984) controversy. In Section 5 I will show that the effect of not knowing c when using my approach is minimal.

For reasons of parsimony I consider two-way analysis of variance with one observation per cell as an example; the method obviously applies more generally.

2. Robustness

When researchers analyze data assuming a linear model with normal errors, a small number of observations can have undue influence on the analysis (Cook, 1977, 1979). The researcher would obviously delete any known "contaminant" (Beckman and Cook, 1983) from the analysis. But such an action is unwarranted when the observation is merely "discordant" (Beckman and Cook, 1983). While a discordant observation can distort an analysis, especially a parametric analysis, it contains information that should not be ignored. Observations believed to be genuine should not be rejected just because the method of analysis gives them undue influence; rather, a method should be used that can accommodate such observations. Using a robust method alleviates the need to decide whether an outlier should be rejected; see Andrews (1971), for example.

If c in (1) is suitably chosen, the subsequent analysis is robust against long-tailed alternatives. But robustness alone is not enough to attract users. To be attractive a procedure should have the following characteristics:

(i) Robust—not greatly affected by extreme observations;
(ii) Powerful, yet preserve Type I error;
(iii) Available in statistical packages with little or no additional programming; and
(iv) Easy to understand.

These are obviously related and have varying degrees of desirability. The procedure based on (1) suggested in this paper meets all four of these desiderata. In addition, I show in Section 4 that a parametric analysis using (1) with c sufficiently large is equivalent to such an analysis on the original, untransformed data. So using (1) represents a generalization of the "usual" parametric analysis. In this sense nothing is lost when using (1) and much can be gained.

Obvious alternatives to parametric analyses on transformed data are nonparametric analyses. In a series of papers, Conover and Iman (1981, for example) consider combining these two approaches. Namely, they apply a rank transformation and then carry out a parametric analysis. In this paper I use two of their rank transformations for comparison. The first is "rank transformation 1," or RT-1, in which all the observations in a data set are ordered and ranked 1, 2, ... The second is RT-2, in which the data are ranked in selected subsets; I have chosen to rank within blocks. These procedures also meet the four desiderata given above; as such they provide excellent alternatives to the method I propose. Moreover, they do not require choosing the value of an unknown parameter, such as an additive constant.

3. Some Examples

Four previously unpublished examples are presented in this section. Examples 1 and 3 deal with a counting process (premature heart beats) in which very large counts are common. Examples 2 and 4 concern nonnegative data in which outliers are present. These are discordant observations but it is not obvious whether they are also contaminants. I will compare the usual least squares analysis of these examples with the two Conover and Iman (1981) procedures mentioned in the previous section, and also with the method based on (1) to be developed in Section 4. I selected the examples because standard ANOVA seems inappropriate; they were not selected because they put the method in a favorable light.

The first example deals with two treatments and paired observations.

Example 1. Twelve patients who experienced frequent premature ventricular contractions (PVCs) were administered a drug with antiarrhythmic properties. One-minute EKG recordings were taken before and after drug administration. The PVCs were counted on both recordings. The results are shown in Table 1.

Table 1
Data for Example 1

Patient number	PVCs per minute		
	Predrug	Postdrug	Decrease
1	6	5	1
2	9	2	7
3	17	0	17
4	22	0	22
5	7	2	5
6	5	1	4
7	5	0	5
8	14	0	14
9	9	0	9
10	7	0	7
11	9	13	−4
12	51	0	51
Mean	13.4	1.9	11.5

The average decrease for the twelve patients is 11.50 PVCs/min with standard deviation 14.29. So $t = 2.79$ and two-sided $P = .018$. This investigator actually calculated t after every patient [see Berry (1985) for a discussion of this practice from various points of view] and found that t *decreased* (from $t = 3.56$, $P = .005$) when patient 12 was included, even though this patient's response was the most favorable in the experiment! The reason should be evident to any statistician: though the mean is somewhat less (7.91 instead of 11.50) without patient 12, the standard deviation is almost halved (7.37 instead of 14.29).

It is obvious that the assumption of normality required for the t test is not appropriate for this experiment. Various nonparametric tests are quite proper. These include the signed-rank test ($P = .002$); RT-1 of Conover and Iman (1981), in which all 24 counts are ranked and a paired-t test is carried out on the ranks ($t = 5.03$, $P = .0004$); and the sign test, which is essentially the same as RT-2 of Conover and Iman ($P = .006$). Table 2 compares the P-values for these tests with those based on (1)—with and without patient 12. The selection of c_0 is described in Section 4, but I should point out here that c_0 changes when patient 12 is deleted.

Table 2
P-values for Example 1

Transformation		df	t	P
Original data	All data	11	2.78	.018
	Pt 12 deleted	10	3.56	.005
Signed-rank	All data	—	—	.002
	Pt 12 deleted	—	—	.003
Sign test	All data	—	—	.006
	Pt 12 deleted	—	—	.012
RT-1	All data	11	5.03	4E−4
	Pt 12 deleted	10	4.63	9E−4
$\log(x + c_0)$ (§4)	All data	11	4.23	.001
	Pt 12 deleted	10	3.79	.004

I want to make it clear that this table, and similar tables in this section, are for comparison only; a small P-value speaks neither for nor against the transformation. Compare the first two and last two lines of Table 2. Without patient 12 the transform has little effect (and for more nearly normally distributed data would have no effect at all). When patient 12 is added, the value of t decreases markedly in the untransformed case for reasons discussed above. On the other hand, the effect of patient 12 is to increase t in the transformed case; this is quite appropriate since the effect of patient 12 is in the same direction as the average of the rest of the data. The three nonparametric tests behave similarly.

Example 2. Five types of electrodes were applied to the arms of 16 subjects and the resistance measured. The experiment was designed to see whether all five electrode types performed similarly. The results are shown in Table 3. After obtaining the results the experimenters decided that the reason for the two large readings on subject 15 was the excessive amount of hair on those parts of the subject's arm. They concluded that this subject's data should be deleted. Whether these readings are contaminants is not clear; the amount of hair present for the other 78 readings was not assessed relative to these two and no such assessment was made independent of the results.

Table 4 gives the F tests with and without subject 15 for these data, for all the data ranked from 1 to 80 or 1 to 76 (RT-1), for the data ranked within subjects (RT-2), and for the log transform with c_0 defined in Section 4. The two large readings on subject 15 occur

Table 3
Resistance (in k.ohms)

Subject number	Electrode type				
	1	2	3	4	5
1	500	400	98	200	250
2	660	600	600	75	310
3	250	370	220	250	220
4	72	140	240	33	54
5	135	300	450	430	70
6	27	84	135	190	180
7	100	50	82	73	78
8	105	180	32	58	32
9	90	180	220	34	64
10	200	290	320	280	135
11	15	45	75	88	80
12	160	200	300	300	220
13	250	400	50	50	92
14	170	310	230	20	150
15	66	1000[a]	1050[a]	280	220
16	107	48	26	45	51
Mean	181.7	287.3	258.0	150.4	137.9

[a] Hairy part of arm?

Table 4
F tests for Example 2

Transformation		Source	df	*F*	*P*
Original data	All data	Electrodes	4	3.15	.020
		Subjects	15	4.17	4E−5
	Subject 15 deleted	Electrodes	4	2.63	.044
		Subjects	14	5.34	3E−6
RT-1	All data	Electrodes	4	2.47	.054
		Subjects	15	4.62	9E−6
	Subject 15 deleted	Electrodes	4	2.19	.082
		Subjects	14	4.78	1E−5
RT-2	All data	Electrodes	4	1.40	.24
	Subject 15 deleted	Electrodes	4	1.16	.34
$\log(x + c_0)$ (§4)	All data	Electrodes	4	2.87	.030
		Subjects	15	4.74	7E−6
	Subject 15 deleted	Electrodes	4	2.04	.10
		Subjects	14	4.42	3E−5

with electrode types 2 and 3. Types 2 and 3 have the largest mean resistance both with and without subject 15. So deleting subject 15 has the effect of decreasing the electrode F statistic for all tests. It is interesting that F for the log transform decreases more than does F for the original data. The reason is that the large readings have less effect on the error variance in the transformed case. If one uses $P < .05$ to indicate "statistical significance" (a deplorable but widespread practice!) then some tests show significant differences among electrodes while others do not. Which to use? Some statisticians recommend combining these P-values, accounting for correlation among the tests, but in my view this is taking hypothesis testing much too seriously (Berry, 1985).

Example 3. Twelve patients with cardiac arrhythmias similar to those described in Example 1 were treated with three active drugs, A, B, and C, in a double-blind, three-period crossover trial. Each period consisted of 1 week of treatment followed by a 24-hour ambulatory EKG recording. (These three weeks were separated by two long periods with no drug.) Table 5 gives the mean number of PVCs per hour for each patient–drug combination; the crossover aspect of the design has been suppressed.

Table 5
Data for Example 3

Patient number	PVCs per hour		
	Drug A	Drug B	Drug C
1	170	7	0
2	19	1.4	6
3	187	205	18
4	10	.3	1
5	216	.2	22
6	49	33	30
7	7	37	3
8	474	9	5
9	.4	.6	0
10	1.4	63	36
11	27	145	26
12	29	0	0
Mean	99.2	41.8	12.3

The data suggest that drug C is better than the other two. However, there is enormous variability in PVC data within patients as well as between patients. For example, any of these patients could easily have the indicated numbers of PVCs on three successive drug-free days. On the other hand, drug C is first- or second-best among the three for all twelve patients—compare RT-2.

Table 6 gives the F tests for the data shown in Table 5, and also for these data ranked from 1 to 36 (RT-1), ranked within patients (RT-2), and the log transform. The effect of the log transform is to increase the F statistics, but despite the great variability in the data, the increases are not enormous. The conclusions for RT-1 are comparable with those for the log transform. For reasons suggested above, the P-value for RT-2 is smallest of all.

Table 6
F tests for Example 3

Transformation	Source	df	F	P
Original data	Drugs	2	2.87	.078
	Patients	11	1.03	.45
RT-1	Drugs	2	3.63	.043
	Patients	11	2.03	.076
RT-2	Drugs	2	5.54	.011
$\log(x + c_0)$ (§4)	Drugs	2	3.50	.048
	Patients	11	1.66	.15

Example 4. Fifteen subjects were administered digoxin on the same schedule each day for several weeks. Another drug was given concurrently at various times throughout the study. Plasma digoxin levels were measured at the same time on each of six days. The data are shown in Table 7. The question addressed is: Does the second drug affect the blood

Table 7
Data for Example 4

Subject number	Plasma digoxin (ng/ml/100)					
	Day A	Day B	Day C	Day D	Day E	Day F
1	34	40	56	45	121	34
2	74	15	44	49	38	37
3	31	49	52	58	41	23
4	40	22	38	25	29	32
5	143	33	57	51	49	46
6	54	58	59	43	41	35
7	31	48	65	53	44	30
8	52	61	70	62	66	38
9	51	63	76	50	32	16
10	21	36	56	40	39	38
11	42	38	68	52	55	51
12	20	30	36	40	31	40
13	51	47	47	38	46	31
14	27	28	61	67	92	75
15	52	43	73	56	56	44
Mean	48.2	40.7	57.2	48.6	52.0	38.0

levels of digoxin? While the dosing schedule of the second drug is very relevant for answering this question, I will suppress this schedule and restrict consideration to whether there is a difference among the days.

There are two obvious outliers. Subsequent laboratory analyses virtually duplicated the data as shown and the bioanalysts had no explanation for these observations. It is tempting to delete them before analyzing the data, substituting estimates (the 143 becomes 41 and the 121 becomes 43) as described by Snedecor and Cochran (1967, §11.9), and this in fact was done. But regulatory agencies and other consumers of statistical analyses look askance at such procedures: Were various combinations of "outliers" deleted until a favorable conclusion was obtained? Why, for example, were the 15 (subject 2, day B) and 92 (subject 14, day E) not deleted?

Table 8 compares the F tests for various transformations. "Outliers replaced" means that the two largest observations were replaced with the estimates indicated above. The outliers

Table 8
F tests for Example 4

Transformation		Source	df	F	P
Original data	All data	Days	5	2.19	.065
		Subjects	14	1.59	.10
	Two outliers replaced	Days	5	4.72	9E−4
		Subjects	14	2.58	.005
RT-1	All data	Days	5	5.17	4E−4
		Subjects	14	2.71	.003
	Two outliers replaced		5	6.16	9E−5
			14	3.05	.001
RT-2	All data	Days	5	6.53	5E−5
	Two outliers replaced		5	7.81	7E−6
$\log(x + c_0)$ (§4)	All data	Days	5	3.38	.009
		Subjects	14	1.89	.042
	Two outliers replaced	Days	5	4.85	7E−4
		Subjects	14	2.44	.007

have a dramatic effect on the conclusions of ANOVA applied to the original data. This is despite the fact that the means of days A and E are in the midst of the means of the other days both with and without these outliers. Again, the explanation lies in the effect of outliers on error variance. While less dramatic, this effect on ANOVA of the transformed data is in the same direction. The rank transformations are less affected by the outliers than is the log transformation. In my opinion it is not appropriate to use ANOVA on the original data, nor is it appropriate to replace the outliers; rather, the statistician should use a robust method that accommodates any such outliers.

4. Choosing an Additive Constant

Consider a two-way model of n_1 treatments and n_2 blocks assuming no treatment–block interaction. If one plans to add a constant, take logs, and then do ANOVA, one must be assuming, at least approximately, that

$$y_{ij} = \log(x_{ij} + c) = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$, where ε_{ij} are independent $N(0, \sigma^2)$.

Box and Cox (1964) suggest estimating c using maximum likelihood, or using the posterior distribution of c calculated assuming an improper prior. For reasons indicated below, such likelihood-based approaches cannot work (Hill, 1963). In terms of the original data x_{ij}, the likelihood is proportional to

$$\sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i,j}[\log(x_{ij}+c)-\mu-\alpha_i-\beta_j]^2\right\}\prod_{i,j}\frac{1}{x_{ij}+c},$$

where $n = n_1 n_2$ [cf. Box and Cox, 1964, eq. (5)]. With standard notation, for any c the maximum likelihood estimates of μ, α_i, β_j, and σ^2 are as usual:

$$\hat{\mu} = \hat{\mu}(c) = \bar{y} = \sum_{i,j} y_{ij}/n,$$

$$\hat{\alpha}_i = \hat{\alpha}_i(c) = \bar{y}_{i.} - \bar{y},$$

$$\hat{\beta}_j = \hat{\beta}_j(c) = \bar{y}_{.j} - \bar{y},$$

$$\hat{\sigma}^2 = \hat{\sigma}^2(c) = \sum_{i,j}(y_{ij} - \hat{y}_{ij})^2/n,$$

where

$$\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j.$$

For fixed c the maximized likelihood is proportional to

$$[\hat{\sigma}(c)]^{-n}\prod_{i,j}\frac{1}{x_{ij}+c}. \qquad (2)$$

This is a badly behaved function of c. It tends to ∞ as $c \to -\min x_{ij}$. And it tends to a positive constant as $c \to \infty$. [Depending on the data, there may be a local maximum at an intermediate value of c; this happens in Examples 2 and 4 but (2) is monotone in Examples 1 and 3. Griffiths (1980) shows some graphs of (2) and discusses the literature involving the three-parameter log-normal distribution.] So the maximum likelihood estimate of c cannot be used in any analysis; and there can be no proper posterior distribution for the prior assumed by Box and Cox. Hill (1963) gives a proper Bayesian analysis, interpreting $-c$ as an incubation period. In the current setting c is quite artificial: it measures lack of normality but has no physical meaning. Unless the researcher has previous

experience with similar data sets and can quantify this into a proper prior distribution whose support is bounded away from $-\min x_{ij}$, Bayes' theorem cannot be used.

To see that $c \to \infty$ corresponds to the untransformed case, write the sum of squares in the numerator of the F statistic (for treatment, say) as

$$\sum_j \left[\log(x_{ij} + c) - \sum_{i,j} \log(x_{ij} + c)/n\right]^2 = \sum_j \left[\log(x_{ij}/c + 1) - \sum_{i,j} \log(x_{ij}/c + 1)/n\right]^2.$$

Using the first term in the Taylor expansion, appropriate for large c, one obtains the approximation

$$\frac{1}{c^2} \sum_j \left[x_{ij} - \sum_{i,j} x_{ij}/n\right]^2.$$

Writing the denominator sum of squares similarly also gives the multiple c^{-2}, which therefore cancels. So, in the limit, the value of F in the transformed case is the same as for the untransformed case. The convergence rate depends on the data set (see the upcoming Figures 3, 4, 5, and 9).

A number of alternative ways for choosing c have been proposed (Draper and Hunter, 1969; Harris and DeMets, 1972; Hoyle, 1973; Carroll, 1980). If the assumption of normality is appropriate then the residuals will tend to be symmetric and moderate in size. So one might choose a value of c to make the residuals as symmetric as possible and the kurtosis small. Skewness and kurtosis are usually defined as

$$g_1 = g_1(c) = \sum_{i,j} (y_{ij} - \hat{y}_{ij})^3/(n\hat{\sigma}^3), \tag{3}$$

$$g_2' = g_2'(c) = \sum_{i,j} (y_{ij} - \hat{y}_{ij})^4/(n\hat{\sigma}^4) - 3.$$

When the y_{ij} are normal, g_1 has mean 0 and, following Cramer (1946, p. 386), g_2' has mean $-6/(d + 2)$ where $d = (n_1 - 1)(n_2 - 1)$, the number of error degrees of freedom. So to measure kurtosis I will use

$$g_2(c) = g_2'(c) + 6/(d + 2). \tag{4}$$

Define

$$g_0(c) = |g_1(c)| + |g_2(c)|. \tag{5}$$

Let c_k denote the value of c that minimizes $|g_k|$, $k = 0, 1, 2$. The values of c_k tend to be rather close to each other for data sets I have analyzed. (They can all have rather small likelihoods, but not as small as the untransformed data, $c = \infty$.) The choice of the additive constant c_0 has at least two advantages over c_1 and c_2: (i) c_0 generally has greater power (cf. §5); and (ii) c_0 is unique when both n_1 and n_2 are at least 3, while for some data sets this is not true of either c_1 or c_2 (c_2 is not unique in Examples 2 and 3). [Another advantage (shared by c_0 and c_2 over c_1) is that the residuals are always symmetric when either n_1 or n_2 is 2, and so in that case $g_1(c) = 0$.]

For many data sets the values of the c_k are very different from values dictated by likelihood considerations. Suppose we set aside data sets that give rise to decreasing likelihoods for all $c > -\min x_{ij}$. Then there is at least one relative maximum $\hat{c}$ that is greater than $-\min x_{ij}$. The ANOVA model suggests that $\hat{c}$ should result in residuals that are more nearly normal than are those from nearby c's. But $g_1(\hat{c})$ and $g_2(\hat{c})$ can both be far from 0 and the residuals for $\hat{c}$ can fail the usual tests for normality. On the other hand, typically $g_1(c_0)$ and $g_2(c_0)$ will be close to 0 (c_0 minimizes $|g_1(c)| + |g_2(c)|$) and the residuals at $c = c_0$ appear ideally normal!

Minimizing skewness plus kurtosis has an important implication in data analysis: when the third and fourth moments are small, the resulting data set tends to be free of outliers. If one takes logarithms in the presence of large observations to minimize their influence on the analysis, one can create a data set in which *small* observations become influential. This does not happen when using $\log(x + c_0)$, but it can easily happen using likelihood-based methods to estimate c. For example, Atkinson (1985, Chap. 9) considers a data set in which his analysis leads to using a small value of c and forces him to reject the smallest observation as an outlier. Minimizing skewness plus kurtosis for these data gives a much larger value of c and the residual for the smallest observation is at the 20th percentile among the residuals.

A question of some importance is the influence that a small number of observations can have on the choice of c and, especially, on the conclusions of the statistical analysis. If there are several large (or small) observations, then deleting any observation affects c_0 and the resulting conclusions very little. However, suppose there is a single large observation (as in Example 1). Then deleting this observation can have a dramatic effect on c_0. Including a large observation will tend to make c_0 moderate in size and deleting it can make $c_0 = \infty$ (if the remaining residuals are close to normal in the untransformed case). But the conclusion is always quite resistant to deleting any observation. As a case in point, patient 12 in Example 1 has a greater effect on the t statistic when using $\log(x + c_0)$ than does any other patient. Without patient 12 the P-value is .004, and with patient 12 it is .001. This higher level of statistical significance is consistent with the dramatic decrease in PVCs for patient 12.

In the next section, analyses of variance using $\log(x_{ij} + c_k)$ for $k = 0, 1, 2$ are compared via simulation with using $\log(x_{ij} + c^*)$, when c^* is the actual value of c, and also with ANOVA using RT-1 and RT-2.

5. Power Comparisons

For the purposes of this section, fix $n_1 = 3$ and $n_2 = 12$ (as in Example 3); other sample sizes provide qualitatively similar results. Also, throughout this section the actual value of c is $c^* = .01$. Assume that the $(y_{1j}, y_{2j}, y_{3j})'$ are independent trivariate normals:

$$\begin{pmatrix} y_{1j} \\ y_{2j} \\ y_{3j} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} \right). \tag{6}$$

Then the $(x_{1j} + c^*, x_{2j} + c^*, x_{3j} + c^*)'$ are trivariate log-normal with

$$\mathrm{E}(x_{ij} + c^*) = \exp(m_i + \sigma^2/2), \quad \mathrm{var}(x_{ij}) = [\exp(\sigma^2) - 1]\exp(2m_i + \sigma^2). \tag{7}$$

The parameter ρ measures block effect; if $\rho = 0$ then there is none.

To calculate power, $n_2 = 12$ normal triples as in (6) were simulated using the Cray 2 computer and the GGNPM normal variate generator in IMSL. Then the "original data"

$$x_{ij} = \exp(y_{ij}) - c^*$$

were calculated. Values of c over a fine grid were added to each x_{ij} and logarithms taken; only values of c for which $x_{ij} + c > 0$ for all i and j were used. An analysis of variance was carried out for each value of c. For each c it was noted whether the null hypotheses

$$H_0: \quad \alpha_1 = \alpha_2 = \alpha_3 = 0,$$

$$H_0': \quad \beta_1 = \cdots = \beta_{12} = 0,$$

were rejected at the .05 significance level ($F_{2,22} > 2.26$ and $F_{11,22} > 3.44$). The same was

noted for $c = c_0$, c_1, and c_2, the values of c that minimize g_0, g_1, and g_2, respectively. (In case of nonuniqueness the smallest value of c_k was used.) In addition, ANOVAS using rank transformations RT-1 and RT-2 were carried out for each data set. Repeating this procedure gives estimates of power for the five methods.

Of special concern is that the significance levels for testing H_0 and H_0' be preserved. Over a wide range of parameter values, the actual significance levels using $\log(x_{ij} + c_k)$ for $k = 0, 1, 2$ are always between .05 and .055. So in this setting little is lost by using a transformation dictated by the data (cf. Carroll and Ruppert, 1984). (On the other hand, choosing c corresponding to the largest value of the F statistic, for example, can give a significance level greater than .08.) Similarly, the significance levels for RT-1 are generally between .05 and .055, with that for RT-2 occasionally as large as .06.

Rank transformation RT-2 is effectively dominated by RT-1 in this log-normal setting, the latter being up to 15% more powerful. Also, using $\log(x_{ij} + c_0)$ generally results in better power than using either c_1 or c_2, though the three are really quite comparable (c_2, kurtosis, is typically second-best). Therefore, I shall concentrate on RT-1 and $\log(x_{ij} + c_0)$.

Consider the case $m_1 = m_2 = 3$, $\sigma^2 = 3$, and $\rho = .25$; the results are similar for other parameter values. From (7), the mean of the first two treatments is about 90 and the standard deviation is about 390. Obviously, there is substantial probability near 0 and the right tail is long; for any m_3, the skewness and kurtosis of the residuals average about 1.5 and 7.5, respectively. Figure 1 shows power as a function of c for $m_3 = -1, 1, 3, 5, 7$. Twenty thousand simulations were performed for each curve, so the standard errors are less than .004. Note the differences between the power curves for $m_3 = 3 - \delta$ and $m_3 = 3 + \delta$. Also note the low power that results from not transforming. (The smallest value of c that can be used for σ^2 as large as 3 is $c^* = .01$ since some of the x_{ij} are very

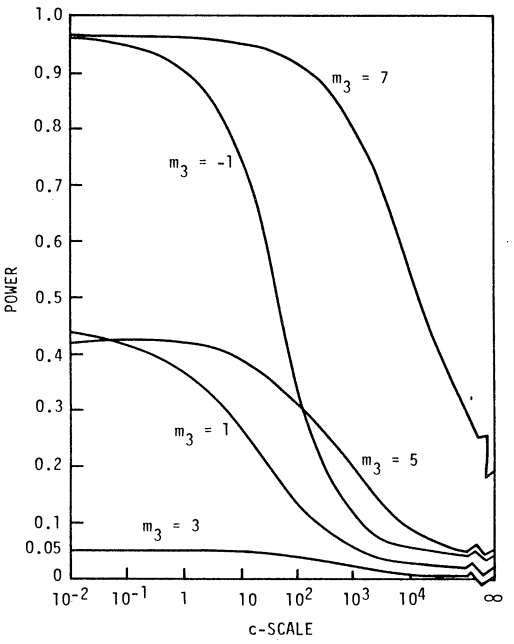

Figure 1. Power curves for $\log(x + c)$ with $n_1 = 3$, $n_2 = 12$, $m_1 = m_2 = 3$, $\sigma^2 = 3$, $\rho = .25$, and $c^* = .01$.

close to 0; when σ^2 is sufficiently small then values of c less than c^* are possible and I find that the power drops sharply as c decreases from c^*.)

Figure 2 compares the power functions (of m_3) for ANOVA using x_{ij}, $\log(x_{ij} + c^*)$, $\log(x_{ij} + c_0)$, and RT-1; $m_3 = 3$ is the null case. The figure was drawn using 20,000 simulations at each of $m_3 = -1, -\frac{1}{2}, 0, \frac{1}{2}, \ldots, 7$; the standard error is less than the width of the lines. Evidently, there is little power lost when using either RT-1 or $\log(x_{ij} + c_0)$; the former is slightly more powerful for $m_3 < 3$ and the latter is better for $m_3 > 3$. This asymmetry about $m_3 = 3$ in using $\log(x_{ij} + c_0)$ follows from the corresponding asymmetry evinced by Figure 1. On the other hand, it is difficult to reject H_0 without transforming, whether or not H_0 is true; actual Type I error probability is less than .01.

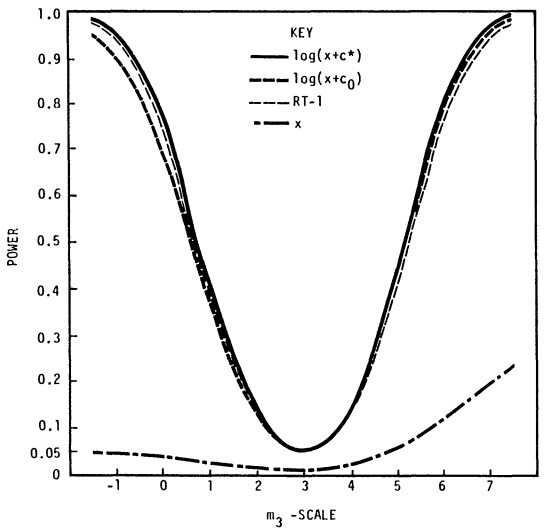

Figure 2. Power curves for $n_1 = 3$, $n_2 = 12$, $m_1 = m_2 = 3$, $\sigma^2 = 3$, $\rho = .25$, and $c^* = .01$.

There is little to choose between $\log(x_{ij} + c_0)$ and ranks on the basis of Figure 2, though RT-1 has a slight edge for the case $c^* = .01$ shown in this figure. However, in the untransformed case ($c^* = \infty$) the log transformation uniformly dominates RT-1 from the point of view of power, though the difference is less than .05. In the latter case, the power function using $\log(x_{ij} + c_0)$ is virtually identical with that of (correctly) assuming that the data are normal. Incidentally, in the normal case the method proposed here leaves the data untransformed ($c_0 = \infty$) about 50% of the time, and when $c_0 < \infty$ the conclusions are essentially the same as if c_0 were ∞.

All power comparisons described here for treatment effect apply as well for block effect. In particular, using c_0 is more powerful than using c_1 or c_2, and RT-1 is more powerful than RT-2.

The power comparisons in this section between $\log(x + c_k)$ and rank transformations have been somewhat unfair to the latter. The class of models assumed for the log transformation includes the actual model used in simulating. So the only problem when using logs is to find the right additive constant. The method I am proposing does that quite well. But rank transformations cannot exactly fit the true model in the class I have assumed. So the fact that they are very powerful in this class is quite surprising.

6. The Examples Revisited

To use the procedure described in the previous two sections, one carries out a separate ANOVA (or a more general linear model analysis) for transformed data $y = \log(x + c)$ for each c on a grid of values. (It is enough to consider $-\min x < c < 10 \cdot \max x$ together with $c = \infty$). Then one calculates the sum of third and fourth powers of residuals and evaluates g_1, g_2, and g_0 as in (3), (4), and (5)—plots of these functions can aid in choosing c. The analysis used is the one with $c = c_0$, corresponding to minimizing g_0. I will demonstrate this process using the examples introduced in Section 3.

Example 1 is somewhat special in that $n_2 = 2$ and so skewness is 0 for all c. Figure 3 shows g_2 and the treatment F statistic (1 and 11 degrees of freedom) for these data. The function g_0 is not shown because it is simply $|g_2|$. The value of c that minimizes g_0 is $c_0 \approx 2.8$. The resulting analysis was presented earlier in Table 2 (wherein $t = \sqrt{F}$). To show the influence of the outlying patient 12, I carried out the same analysis with this patient deleted. This is shown in Figure 3 and indicated by carets. Now, $\check{c}_0 \approx 45$ and the analysis is also shown in Table 2. [That $\check{g}_2(\infty)$ is near 0 suggests that the normal analysis is more reasonable in the absence of patient 12, but $\check{c}_0 < \infty$ indicates that the residuals for the untransformed data are still somewhat nonnormal even without patient 12.] Since the response of patient 12 is in the same direction as the other patients, only markedly so,

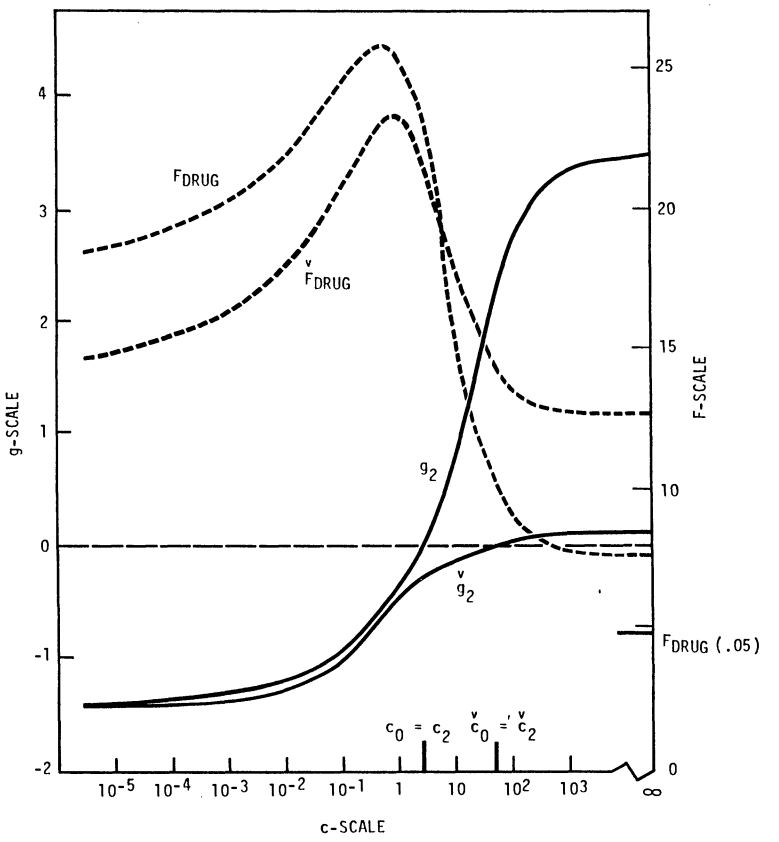

Figure 3. Graphical analysis for Example 1, with and without (v) patient 12.

F_{DRUG} is greater than $\check{F}_{DRUG}$ for small values of c. But as c increases, patient 12's response so inflates the variance that the relationship between the F's reverses. In my view it is wrong to delete patient 12 because the analysis used cannot cope with it; instead, transform to $\log(x + 2.8)$, or use ranks.

Figure 4 shows the analysis for Example 2. The smallest number in the data set is 15, so c is restricted to exceed -15. [I have chosen to show c on the horizontal scale in Figure 4—and also Figure 9—to facilitate reading off the values of c_k; generally the picture would be neater using $c + \min x_{ij}$ or $\log(c + \min x_{ij})$ instead.] In this example (and also in Example 3) the equation $g_2 = 0$ has two roots. (There are data sets for which this equation and also $g_1 = 0$ have more than two roots.) The value of c that minimizes g_0 is $c_0 \approx 32$. Table 4 gives the corresponding analysis.

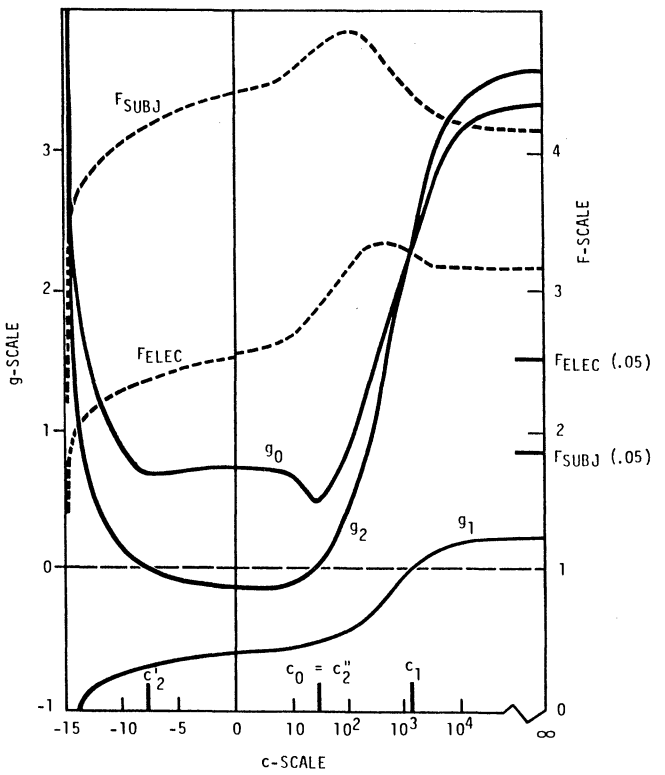

Figure 4. Graphical analysis for Example 2.

Figure 5 shows the curves relevant for Example 3. The value of c that minimizes g_0 is $c_0 \approx 7.9$. Table 6 gives the corresponding analysis. Since the dimensions of this example correspond to the simulations of the previous section, I have chosen it to show the pattern of residuals vs fitted values for the transformed data $y = \log(x + c_0)$, the untransformed data, and the ranked data using RT-1; these are shown respectively in Figures 6, 7, and 8, along with histograms of residuals. (The reader is invited to guess the cell corresponding to the most extreme residual in Figure 7. *Answer:* patient 12, drug B.)

Figure 9 shows the curves for Example 4. As in Example 2, the smallest value in the data set is 15; so $c > -15$. The value of c that minimizes g_0 is $c_0 \approx -6.2$. Table 8 gives the corresponding analysis.

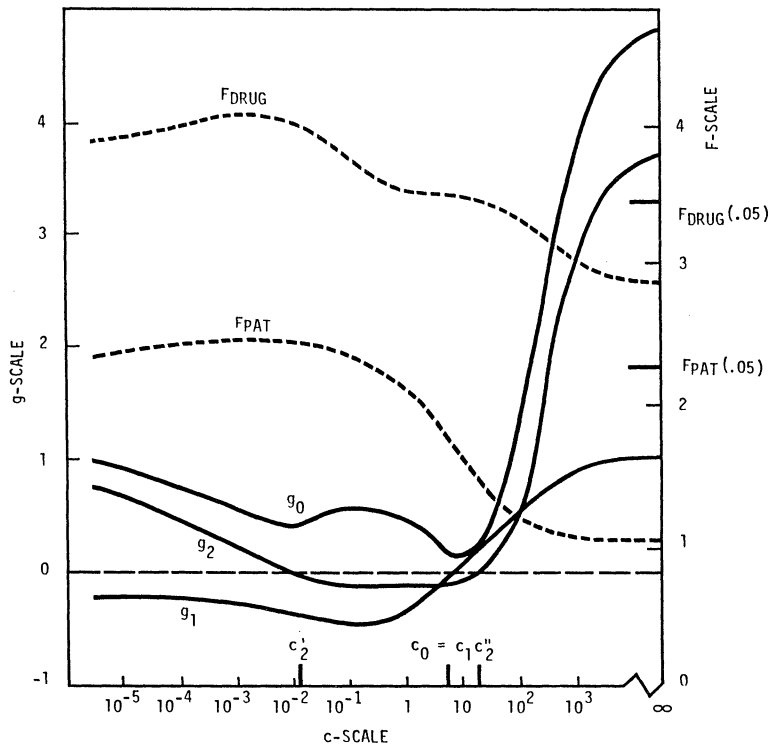

Figure 5. Graphical analysis for Example 3.

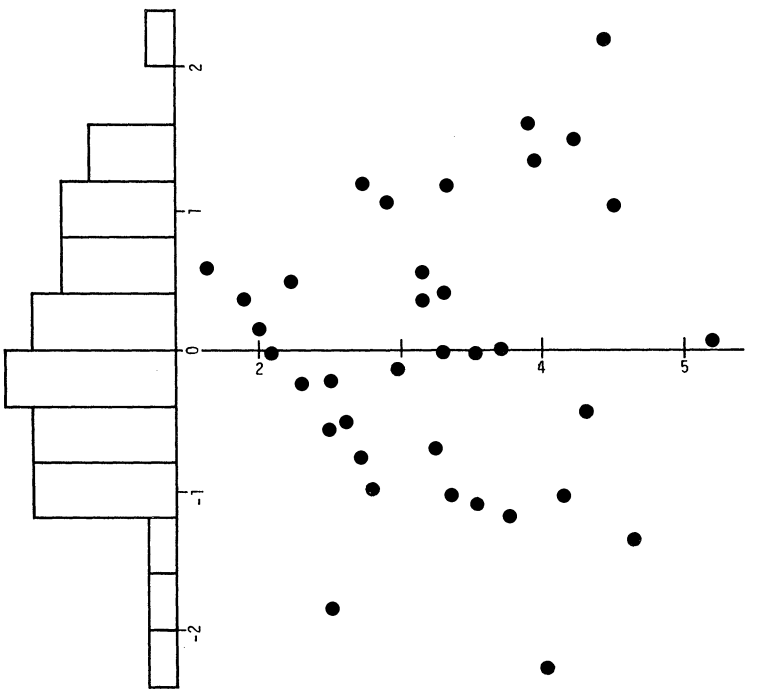

Figure 6. Standardized residuals vs fitted values and histogram of residuals for Example 3; $c = c_0 = 7.9$.

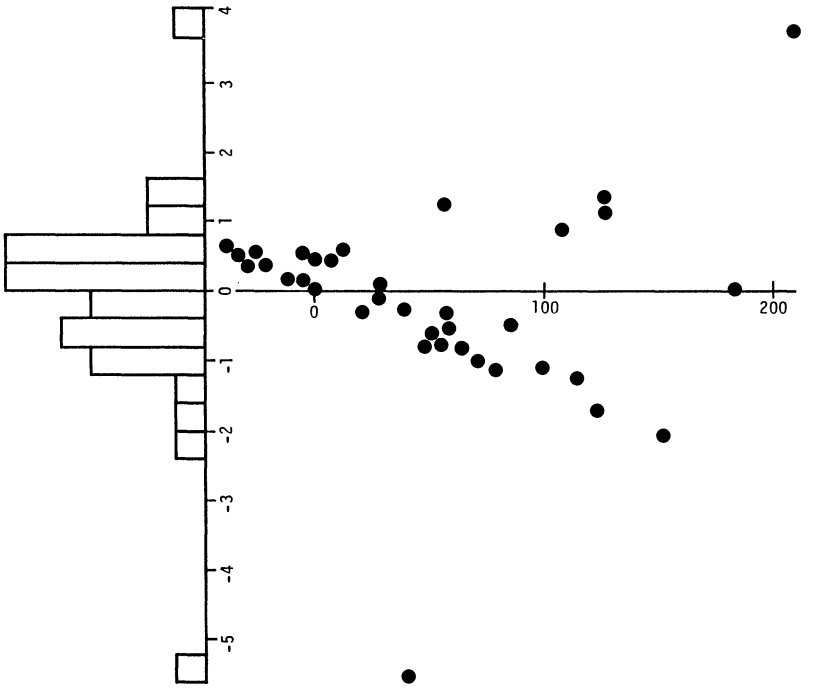

Figure 7. Standardized residuals vs fitted values and histogram of residuals; $c = \infty$ (untransformed case).

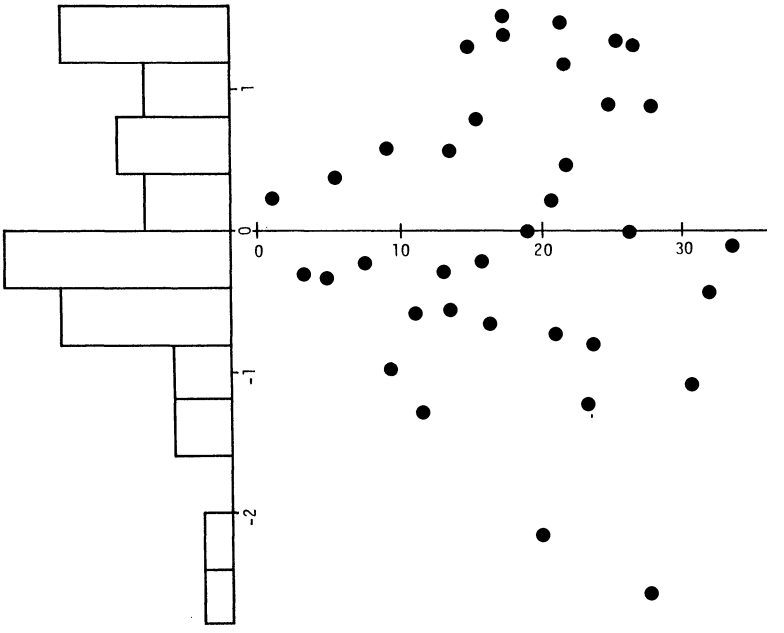

Figure 8. Standardized residuals vs fitted values and histogram of residuals for Example 3; RT-1.

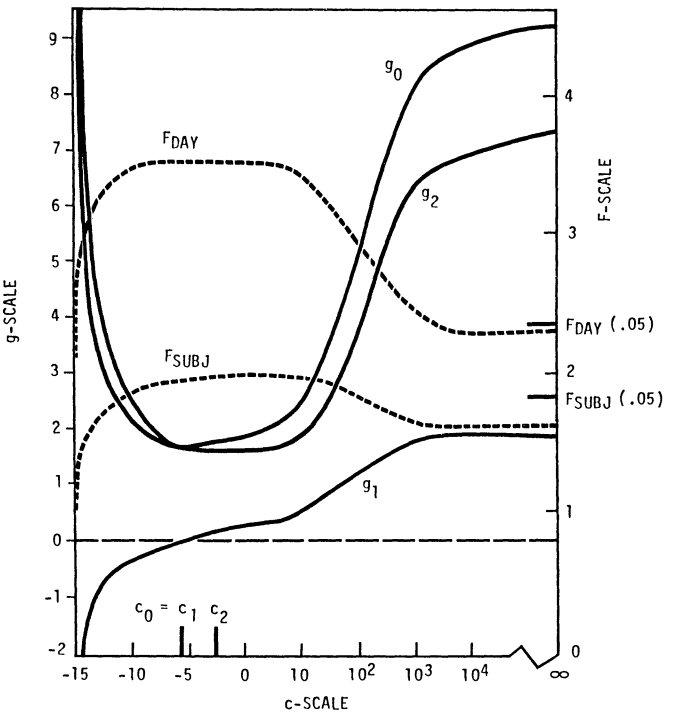

Figure 9. Graphical analysis for Example 4.

7. Discussion

The method proposed here generalizes analysis of variance. It uses the data to decide automatically when a logarithmic transformation is appropriate.

Researchers should learn as much as possible from their data. This includes looking at the data in various ways and using assorted parametric and nonparametric estimates and tests. But, especially when regulatory agencies are involved, some statisticians must, in advance of the experiment, announce *the* test they will use. What happens when they say "t test" and they obtain unexpectedly large deviations? Perhaps they should transform the data first, or use a nonparametric procedure. But they leave themselves open to criticism that the transform or procedure used was chosen to show what they wanted to show.

There is no canonical alternative to the standard analysis. Spelling out in advance all possible contingencies, depending on the data, seems hopeless. No procedure can be a panacea, but the method I propose handles most contingencies that arise in real data settings by simply specifying in advance that $\log(x + c)$ will be used, with c chosen to minimize g_0 and thus make the residuals close to a sample from a normal distribution.

Acknowledgements

I thank Professors R. D. Cook, S. Das Gupta, M. L. Eaton, B. M. Hill, and S. Weisberg for helpful discussions. Two referees gave very helpful suggestions. K. Samaranayake helped with programming the simulations. Riker Laboratories, 3M Corporation, provided the data sets.

Résumé

Une méthode d'estimation de la constante additive c est proposée lorsque l'on veut utiliser une transformation des données du type, x devient $y = \log(x + c)$. Sous l'hypothèse qu'il existe c^* telle que les $x + c^*$ suivent une distribution lognormale, cette méthode préserve le niveau et la puissance des tests pratiqués en analyse de la variance. Ses avantages sont semblables à ceux des transformations de rang, en particulier, elle est facile à mettre en oeuvre et est robuste aux valeurs extrêmes. De plus, comme lorsque c tend vers l'infini, cela équivaut à la transformation $y = x$, pour l'analyse de variance, il s'agit d'une généralisation intéressante de la méthode des moindres carrés.

References

Andrews, D. F. (1971). A note on the selection of data transformations. *Biometrika* **58,** 249–254.

Atkinson, A. C. (1973). Testing transformations to normality. *Journal of the Royal Statistical Society, Series B* **35,** 473–479.

Atkinson, A. C. (1985). *Plots, Transformations and Regression.* Oxford: Clarendon.

Beckman, R. J. and Cook, R. D. (1983). Outlier.....s (with Discussion). *Technometrics* **25,** 119–163.

Berry, D. A. (1985). Interim analysis in clinical trials: Classical vs Bayesian approaches. *Statistics in Medicine* **4,** 521–526.

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association* **76,** 296–311.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with Discussion). *Journal of the Royal Statistical Society, Series B* **26,** 211–252.

Box, G. E. P. and Cox, D. R. (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association* **77,** 209–210.

Carroll, R. J. (1980). A robust method for testing transformations to achieve normality. *Journal of the Royal Statistical Society, Series B* **42,** 71–78.

Carroll, R. J. and Ruppert, D. (1984). Power transformations when fitting theoretical models to data. *Journal of the American Statistical Association* **79,** 321–328.

Conover, W. J. and Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics (with Discussion). *The American Statistician* **35,** 124–133.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* **19,** 15–18.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association* **74,** 169–174.

Cramer, H. (1946). *Mathematical Methods of Statistics.* Princeton, New Jersey: Princeton University Press.

Draper, N. R. and Hunter, W. G. (1969). Transformations: Some examples revisited. *Technometrics* **11,** 23–40.

Griffiths, D. A. (1980). Interval estimator of the three-parameter log-normal distribution via the likelihood function. *Journal of the Royal Statistical Society, Series C* **29,** 58–68.

Harris, E. K. and DeMets, D. L. (1972). Estimation of normal ranges and cumulative proportions by transforming observed distributions to Gaussian form. *Clinical Chemistry* **18,** 605–612.

Hill, B. M. (1963). The three-parameter log-normal distribution and Bayesian analysis of a point-source epidemic. *Journal of the American Statistical Association* **58,** 72–84.

Hinkley, D. V. (1975). On power transformations to symmetry. *Biometrika* **62,** 101–111.

Hinkley, D. V. and Runger, G. (1984). The analysis of transformed data (with Discussion). *Journal of the American Statistical Association* **79,** 302–320.

Hoyle, M. H. (1973). Transformations—An introduction and a bibliography. *International Statistical Review* **41,** 201–223.

Snedecor, G. W. and Cochran, W. G. (1967). *Statistical Methods,* 6th edition. Ames, Iowa: The Iowa State University Press.

Taylor, J. M. G. (1985). Power transformations to symmetry. *Biometrika* **72,** 145–152.

Received April 1986; *revised December* 1986 *and February* 1987.