DOCUMENT RESUME

ED 445 073 TM 031 750

AUTHOR	Miranda, Janet
TITLE	Moving the Bar: Transformations in Linear Regression.
PUB DATE	2000-01-00
NOTE	21p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Dallas, TX, January 27-29, 2000).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Hypothesis Testing; *Regression (Statistics); Statistical Distributions; *Transformations (Mathematics)

ABSTRACT

The assumption that is most important to the hypothesis testing procedure of multiple linear regression is the assumption that the residuals are normally distributed, but this assumption is not always tenable given the realities of some data sets. When normal distribution of the residuals is not met, an alternative method can be initiated. As an alternative, data for one or more of the variables under study can be transformed in order to increase conformity to the required distributional assumptions of linear regression. Such transformations are discussed in this paper, including: (1) transforming data by powers and roots; (2) transforming for skewness; (3) transforming for non-linearity; (4) transforming for non-constant spread; and (5) transforming proportions via probit analysis and logit analysis. Power and root transformations provide a means for improving data distributions and at the same time preserve the directionality of "X." A skewed distribution, represented by a set of scores that form a non-symmetrical curve when plotted on a frequency graph, can be transformed by ascending the ladder of powers to correct a negative skew or descending the ladder of powers to correct a positive skew. For dichotomous quantities, logit and probit are the data transformations best applied. A logit transforms both the upper and lower boundaries of the scale. The probit is similar to the logit but in a different metric. Transformations are useful in examining and modeling data when the assumptions of linear regression are not met. Such transformations do indeed change the original research question. By manipulation of the data, the question is also "transformed." An appendix contains a table of sample data. (Author/SLD)

Moving the Bar: Transformations in Linear Regression

Janet Miranda
University of North Texas

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX, January 27-29, 2000.

ABSTRACT

The assumption that is most important to hypothesis testing procedure of multiple linear regression is the assumption that the residuals are normally distributed. However, this assumption is not always the tenable given the realities of some data sets. When normal distributed of the residuals is not met, an alternative method(s) can be initiated. As an alternative, data for one or more of the variables under study can be transformed in order to increase conformity to the required distributional assumptions of linear regression. Such transformations discussed in this paper include: transforming data by powers and roots; transforming for skewness; transforming for nonlinearity; transforming for nonconstant spread; and transforming proportions via probit analysis and logit analysis.

Power and root transformations provide a means for improving data distributions and at the same time preserve the directionality of X. A skewed distribution, represented by a set of scores that form a nonsymetrical curve when plotted on a frequency graph, can be transformed by ascending the ladder of powers to correct a negative skew or descending the ladder of powers to correct a positive skew. For dichotomous quantities, logit and probit are data transformations best applied. A logit transforms both the upper and lower boundaries of the scale. The probit is similar to the logit but in a different metric.

Transformations are useful in examining and modeling of data when the assumptions of linear regression are not met. Such transformations do indeed change the original research question. My manipulation of the data, the question is also "transformed."

Moving the Bar: Transformations in Linear Regression

Introduction

A powerful method for analyzing a wide variety of statistical situations is multiple linear regression. Multiple regression is a least square general linear model technique for the use of analysis of data with a single dependent variable and one or more independent variables (Kieras, 1984). The assumption that is most important to hypothesis testing procedure of multiple linear regression is the assumption that the residuals are normally distributed (Daniel, 1999). However, this assumption is not always the tenable given the realities of some data sets. When the residuals are not normally distributed, it is inappropriate to use ordinary least squares regression, and nonlinear estimation techniques must be used which overcome statistical difficulties. As an alternative, data for one or more of the variables under study may be transformed so as to increase conformity to the required distributional assumptions of regression. Techniques that are most commonly utilized for such instances include: transforming data by powers and roots; transforming for skewness; transforming for nonlinearity; transforming for nonconstant spread; and transforming proportions via probit analysis and logit analysis.

Transforming Data

To better understand the transformation of data using computer calculations such as logit and probit values, a brief overview of data transformations is provided below. As stated in the introduction, linear regression models are based on certain assumptions about the structure of data which are more often not met than met. Transformations can prove useful in examining the data and in normalizing variable distributions when these assumptions are not met. According to Fox (1997), procedures and issues relative to data transofmations include:

(1) transforming by powers and roots;

(2) transforming for skewness;

(3) transforming for nonlinearity;

(4) transforming for nonconstant spread; and

(5) transforming for proportions.

Transforming by Powers and Roots

Power and root transformations prove especially useful in improving data distributions while preserving the directionality of X. Power transformations preserve the order of the data when all values are positive and when the ratio of the largest to smallest values is itself large. When such conditions are not met, it is possible to impose them by adding a positive or negative "start" (i.e., an additive constant) to all of the data values (Fox, 1997).

A skewed distribution (Figure 1c) is defined as a set of scores that form a nonsymetrical curve when plotted on a frequency graph (Gall, Borg, and Gall, 1996). In transforming skewness, ascending the ladder of powers will likely correct a negative skew while descending the ladder of powers will likely correct a positive skew. Bimodality (Figure 1b) is one of the more difficult distribution problems to correct.

Linear relationships are improved by power transformations. It is possible to take a monotone nonlinear relationship and, by raising data values to a higher power, create a linear relationship which is more advantageous considering that a linear relationship is assumed in traditional regression analysis (See Figure 2).

The problems of skewness and nonconstant spread (i.e., conditional variance of y across values of x) usually occur together. By descending the ladder of powers, a positive association between the level of a variable in different groups and its spread, can be minimized, making the spreads more constant (Fox, 1997). The reverse, albeit less common, can be corrected by ascending the ladder of powers.

According to Fox (1997), power transformations are not usually helpful in proportions data in which quantities bounded by 0 and 1, or 0% and 10%. For these dichotomous quantities, logit and probit transformations are best applied (Conniffe, 1997). The logit transformation converts data to the "log" of the odds of a given value's probability. A logit transformation removes both upper and lower boundaries of the scale. With the tails of the distribution spread out, values are typically made symmetric about a mean of 0. The logit model is a linear, additive model for the log odds, as well as a multiplicative model for the odds. Logit transformations are based on the odds of the occurence over the odds of it not happening. For example, if there is a 30% chance of snow then the logit equation is based on a 30/70 odds ratio.

The probit transformation is the inverse distribution function for the normal distribution. Probits are like logits but in a different metric. It would be like saying that logits are in inches and probits are recorded in centimeters considering that both logits and probits provide logs of probabilities. Generalizing the probit and logit models to several independent variables is straightforward. All that is required is a linear predictor that is a function of several regressors (Fox, 1997). By using a cumulative probability distribution function, logit and probit models transform the linear predictor to the unit interval. Despite their similarity, according to Fox, the logit model is simpler to interpret since it can be written as a linear model for the log odds. Both probit and logit models can be fit to data by the method of maximum likelihood (Fox, 1997). Hence, logit and probit procedures may be useful both in converting data distributions and in developing predictive equations when original data have specific given properties. The present paper is limited to the former (i.e., data conversions).

Transformations demonstrated

The data selected for the demonstration of the effects of various transformations of data are national income and infant mortality data from Leinhardt and Wasserman's (1978) study of 103 nations of the world (Fox, 1997). These data were selected due to the lack of even distribution within the data set. Histograms are provided as means for visually inspecting the data and the effects of transformations. Also, linear regression models are recomputed following each transformation to illustrate the effects of each transformation on regression results. All data were processed using SPSS. The Leinhardt & Wasserman (1997) data set is listed in Appendix 1.

For each of the regression models illustrated herein, the variable "mort" (infant mortality per 1000 live births) served as the dependent variable, and the variable income (per capita income in U.S. dollars) served as the single predictor variable. All transformation involved the data only on the infant mortality variable as the income variable was already relatively normally distributed. Figure 3 presents a histogram of the original data on the mortality variable. Obviously, the distribution is positively skewed, violating the normality assumption necessary to linear regression analysis.

Findings

Data histograms based on squares of the original values and square roots of the original values are presented, respectively, in Figures 4 and 5. As noted earlier, power transformations can sometimes be useful when data are skewed, with higher power transformations most effective for correcting negative skews, and lower power transformations (i.e., roots) most effective for correcting positive skews. Because the original data were positively skewed, it would be expected that the square root transformation would more effectively move the distribution toward normality than would the square transformation. This is indeed the case; however, the square root values still fall into a somewhat positively skewed and oddly shaped distribution.

Because the mortality data represent a proportion (i.e., instance of mortality per 1,000 live births), data transformations suitable to proportions may appropriately be applied to the data. A data histogram based on conversion of the original data to probit values is presented in Figure 6. Note that this distribution is an improvement over the square root transformed distribution (Figure 5), considering the tendency toward minimization of extreme values when logs of odds are taken into consideration. A histogram based on an "arcsine square root" transformation (similar in distribution shape to the probit distribution - Fox, 1997) of the mortality values is presented in Figure 7. This transformation is not nearly as effective as the probit transformation, leading to the conclusion that the probit values are the most appropriate transformation of the data.

Table 1 presents the results of five regression analyses, each employing one of the data transformation. Using the probit transformed values of the dependent variable increased the Rsquare from .109 to .321, increasing the effect from small to moderate. Hence, by changing the metric in which the data are specivied, the assumptions of regression analysis are better substantiated and, additionally, the statistical effect may be appreciably modified.

Summary

Transformations are useful in examining and modeling of data when the assumptions underlying linear regression are not met. Power transformations preserve order when all values are positive and are most effective when the ratio of largest to smallest data values is itself large. If these conditions are not met, it is possible to impose them by adding a positive or negative start to all data values. Descending the ladder of powers is likely to correct a positive skew, while ascending is likely to correct a negative skew. Monotone nonlinearity can sometimes be corrected by a power transformation of one or both variables. Spreads can be made more constant by descending the ladder of powers when there is a positive association and ascending the ladder of powers is the spread is negative. Probit, logit and other proportional data of transformations can be useful to the researcher when exploring research questions using proportional data.

It should be noted that transformation of the data for linear regression purposes does in fact change the original research question. The research question is also "transformed" by the manipulation of the data and should be adjusted to the transformation performed. The researcher should now correctly preface all hypotheses with a statement such as, "If the data were more normally distributed...". However, considering that social science data scales are arbitrarily determined, transformations may be viewed simply as a way to specify data such that the assumption of normality may be met.

References

Conniffe, D. (1997). Improving a linear regression through joint estimation with a probitmodel. *The Statistician*, 46, (4), 487-493.

Daniel, L. (1999). *Logistic Regression*, Handout to EDER 6210, University of North Texas, Denton, Texas, May-mester 1999.

Fox, J. (1997). Applied regression analysis, linear models, and related methods, Thousand Oaks, CA. Sage Publications.

Gall, M., Borg, W.R., & Gall, J.P, (1996). Educational research, White Plains, NY: Longman Publishing.

Kieras, D. E. & Just, M.A. (1984). New methods in reading comprehension research, Hillsdale, N.J.: Lawrence Erlbaum Associates.

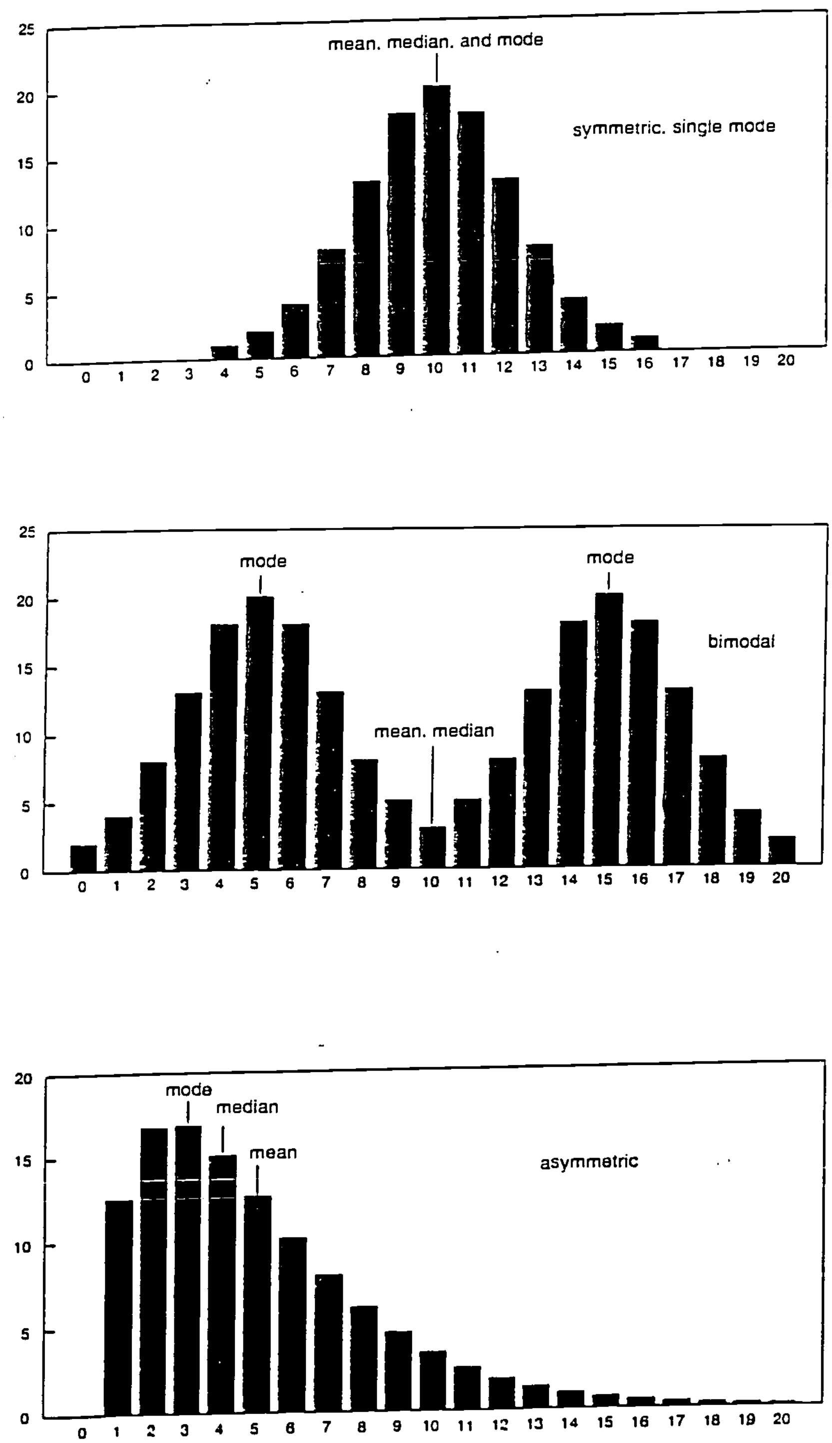
Leinhardt, S., & Wasserman, S.S. (1978). Exploratory data analysis: An introduction to selected methods. In K.F. Schuessler (Ed.), *Sociological methods* 1979 (pp. 311-365). San Francisco, CA: Jossey-Bass.

Table 1
Results of Five Simple Regression Analyses*

Analysis	Transformation	R	R2
1	None	.331	.109
2	Squared Values	.079	.006
3	Square Root of Values	.501	.251
4	Probit	.567	.315
5	Arcsine Square Root	.474	.225

*Based on data from Leinhardt and Wasserman's 1978 study (n=101). For all analyses, the single predictor variable was per capita income in US dollars. The dependent variable was based on various transformations of the infant mortality variable.

Figure 1:
Data distributions representing sysmmetry (a); bimodality (b): and skewness (c).



(a)

(b)

(c)

BEST COPY AVAILABLE

Figure 2: Data relationships displayed as monotone and linear.

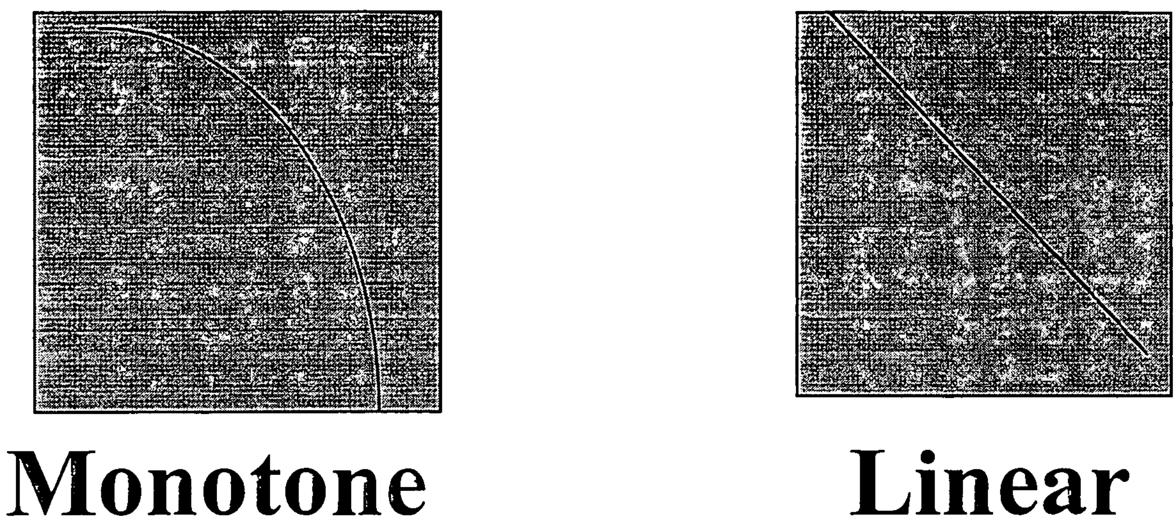


Figure 3: Linhardt & Wasserman infant mortality data.

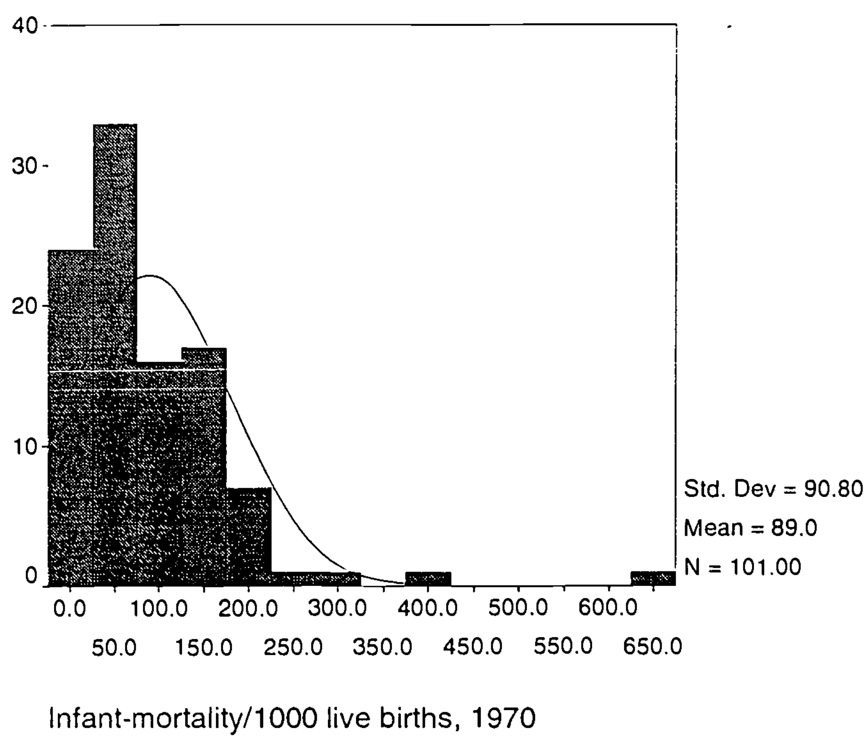


Figure 4: Infant mortality values squared.

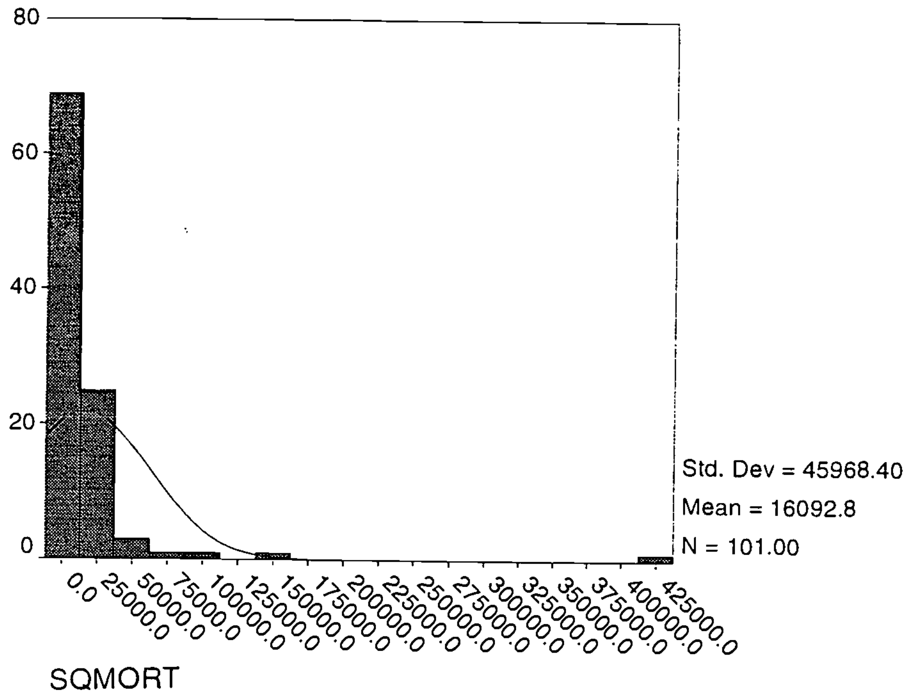


Figure 5: Square Root Transformation of Infant mortality values.

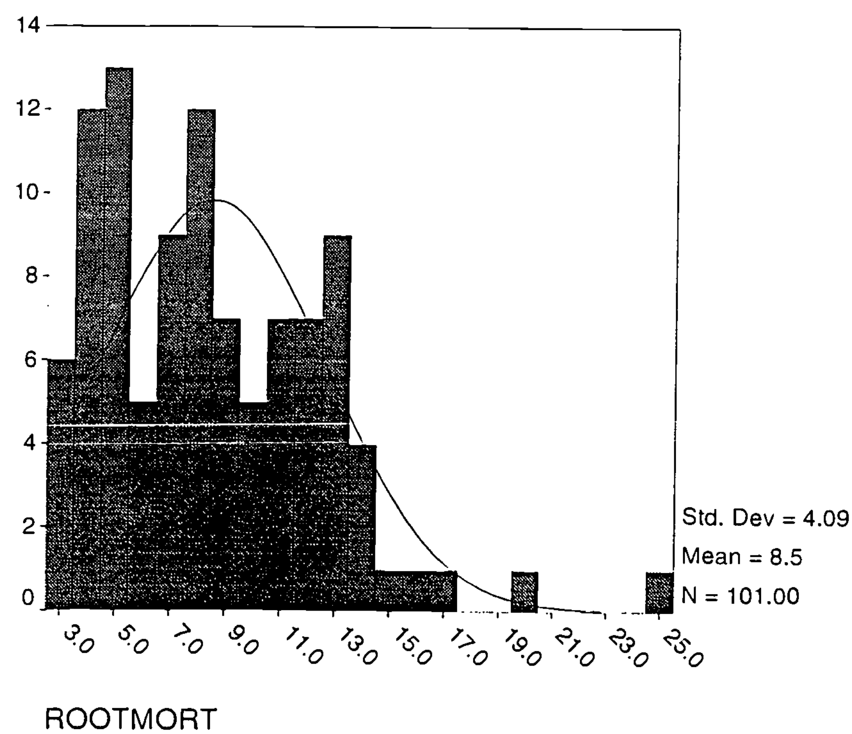


Figure 6: Probit Transformed Infant Mortality Data

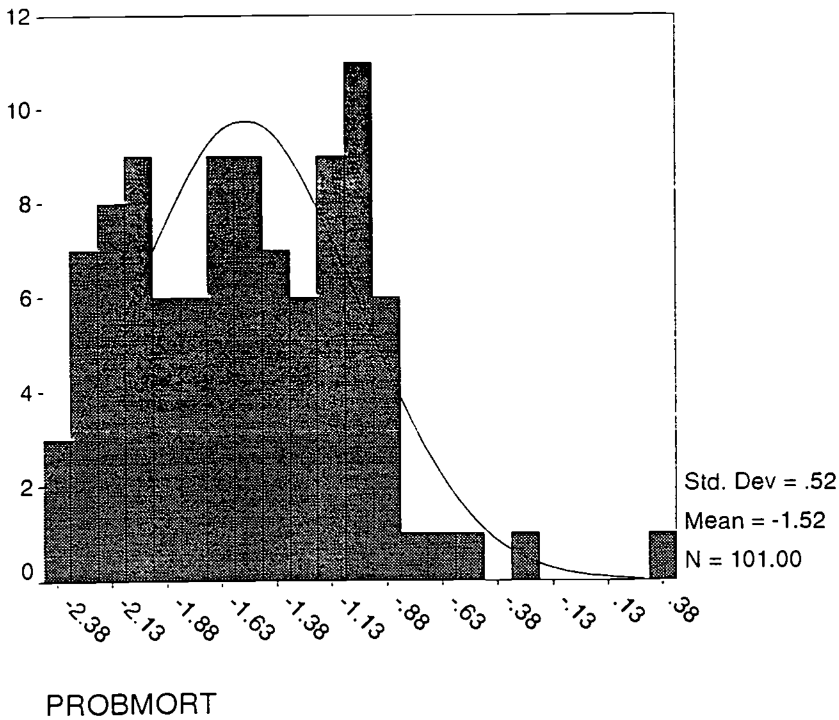
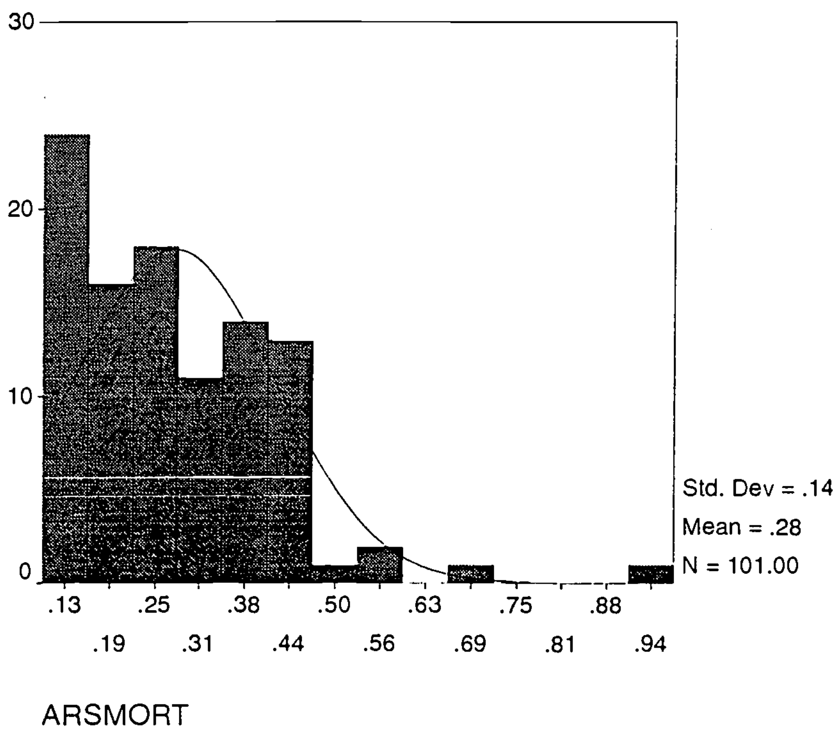


Figure 7: Arcsine Sqare Root Transformation of Mortality Data.



Appendix 1
Leinhardt and Wasserman (1978) Income and Infant Mortality Data

	nation	income	mort
1	Australia	3426.00	26.70
2	Austria	3350.00	23.70
3	Belgium	3346.00	17.00
4	Canada	4751.00	16.80
5	Denmark	5029.00	13.50
6	Finland	3312.00	10.10
7	France	3403.00	12.90
8	West Germany	5040.00	20.40
9	Ireland	2009.00	17.50
10	Italy	2298.00	25.70
11	Japan	3292.00	11.70
12	Netherlands	4103.00	11.60
13	New Zeland	3723.00	16.20
14	Norway	4102.00	11.30
15	Portugal	956.00	44.80
16	South Africa	1000.00	71.50
17	Sweden	5596.00	9.60
18	Switzerland	2963.00	12.80
19	Britain	2503.00	17.50
20	United States	5523.00	17.60
21	Algeria	400.00	86.30
22	Ecuador	250.00	78.50
23	Indonesia	110.00	125.00

24	Iran	1280.00	
25	Iraq	560.00	28.10
26	Libya	3010.00	300.00
27	Nigeria	220.00	58.00
28	Saudi Arabia	1530.00	650.00
29	Venezuela	1240.00	51.70
30	Argentina	1191.00	59.60
31	Brazil	425.00	170.00
32	Chile	590.00	78.00
33	Colombia	426.00	62.80
34	Costa Rica	725.00	54.40
35	Dominican Republic	406.00	48.80
36	Greece	1760.00	27.80
37	Guatemala	302.00	79.10
38	Israel	2526.00	22.10
39	Jamaica	727.00	26.20
40	Lebanon	631.00	13.60
41	Malaysia	295.00	32.00
42	Mexico	684.00	60.90
43	Nicaragua	507.00	46.00
44	Panama	754.00	34.10
45	Peru	335.00	65.10
46	Singapore	1268.00	20.40
47	Spain	1256.00	15.10
48	Taiwan	261.00	19.10
49	Trinidad and Tobago	732.00	26.20

50	Tunisia	434.00	76.30
51	Uruguay	799.00	40.40
52	Yugoslavia	406.00	43.30
53	Zambia	310.00	259.00
54	Bolivia	200.00	60.40
55	Cameroon	100.00	137.00
56	Congo	281.00	180.00
57	Egypt	210.00	114.00
58	El Salvador	319.00	58.20
59	Ghana	217.00	63.70
60	Honduras	284.00	39.30
61	Ivory Coast	387.00	138.00
62	Jordan	334.00	21.30
63	South Korea	344.00	58.00
64	Liberia	197.00	159.20
65	Moroco	279.00	149.00
66	Papua New Guinea	477.00	10.20
67	Paraguay	347.00	38.60
68	Philippines	230.00	67.90
69	Syria	334.00	21.70
70	Thailand	210.00	27.00
71	Turkey	435.00	153.00
72	South Vietnam	130.00	100.00
73	Afganistan	75.00	400.00
74	Bangladesh	100.00	124.30
75	Burma	73.00	200.00

76	Burundi	68.00	150.00
77	Cambodia	123.00	100.00
78	Central African Republic	122.00	190.00
79	Chad	70.00	160.00
80	Dahomey	81.00	109.60
81	Ethiopia	79.00	84.20
82	Guinea	79.00	216.00
83	Haiti	100.00	
84	India	93.00	60.60
85	Kenya	169.00	55.00
86	Laos	71.00	
87	Madagascar	120.00	102.00
88	Malawi	130.00	148.30
89	Mali	50.00	120.00
90	Mauritania	174.00	187.00
91	Nepal	90.00	
92	Niger	70.00	200.00
93	Pakistan	102.00	124.30
94	Rwanda	61.00	132.90
95	Sierra Leone	148.00	170.00
96	Somalia	85.00	158.00
97	Sri Lanka	162.00	45.10
98	Sudan	125.00	129.40
99	Tanzania	120.00	162.50
100	Togo	160.00	127.00
101	Uganda	134.00	160.00

102	Upper Volta	82.00	180.00
103	Southern Yemen	96.00	80.00
104	Yemen	77.00	50.00
105	Zaire	118.00	104.00

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM031750

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Moving the Bar	
Author(s): Janet Miranda	
Corporate Source: University of North Texas Trinity Christian Academy	Publication Date: 1/27/00

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
[X]	[ ]	[ ]
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Janet Miranda	Printed Name/Position/Title: Janet Miranda, Counselor	
Organization/Address: 17001 Addison Road	Telephone: 972-931-8325	FAX: 972-931-9436
	E-Mail Address: jmiranda@trinitychristian.com	Date: 1/27/00

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com