Using Data to Develop Mathematical Models

Philip R. Carlson
Mathematics Department
General College
University of Minnesota

Abstract

An analysis of ordered pairs and their scatter plots leads to interesting questions related to mathematical modeling. Some statistical methods suggest ways to approach this analysis of the ordered pairs. Both high school and college methods are illustrated in this paper.

INTRODUCTION. In industry, most mathematical models are built up through an analysis of data. I would like to illustrate this procedure by looking at the plots of several data sets. For each data set I will choose a transformation to bring the plot to a linear form, and then use linear regression methods to arrive at a linear model which can, by the inverse transformation, suggest the best model for that data. For applications to high school mathematics courses, I have included an algebraic approach to finding the "best line".

AMATYC Curriculum Recommendations

In the AMATYC Curriculum Recommendations (1), Standard 2 indicates the "Students should be able to use the concept of function as a central theme throughout the study of developmental mathematics". Objective 2 indicates that "Students will be able to create and recognize a variety of function patterns." This includes the use of raw data which students are to analyze. I want to illustrate two different ways for students to engage in an analysis of data that leads to a mathematical model describing the data.

Example of a High School Approach

A book that implements the above ideas was written by the mathematics teachers of the North Carolina School of Science and Mathematics entitled: **Contemporary Precalculus through applications** (2). They start the book with a chapter entitled Data Analysis One. The first example in this book centers on the per pupil expenditure versus the graduation rate for each of the 50 states and the District of Columbia. The authors have the students do a plot of the paired data . Students quickly observe that there is no functional relationship between the expenditure and the graduation rate. They go on to other data sets where a linear pattern in the plot is very obvious. The Median- Median Line approach is then introduced by the authors. Following an introduction to this method an example will illustrate its application.

The Median-Median Line Approach (2, pp. 10-12)

The procedure works as follows:

1. Separate the data into three groups of equal size (or as close to equal as possible) according to the values of the horizontal coordinate.

2. Find the summary point for each group based on the median x-value and median y-value of the points in the group.

3. Find the equation of the line through the summary points of the outer two groups. We call this line L.

4. Slide L one-third of the way to the middle group summary point.
   a. Find the y-coordinate of the point on L with the same x-coordinate as the middle summary point.
   b. Find the vertical distance between the middle summary point and the line L by subtracting the y-values.
   c. Find the coordinates of the point P one-third of the way from the line L to the middle summary point.

5. Find the equation of the line through the point P that is parallel to the line L.

Example of Median-Median Line Approach

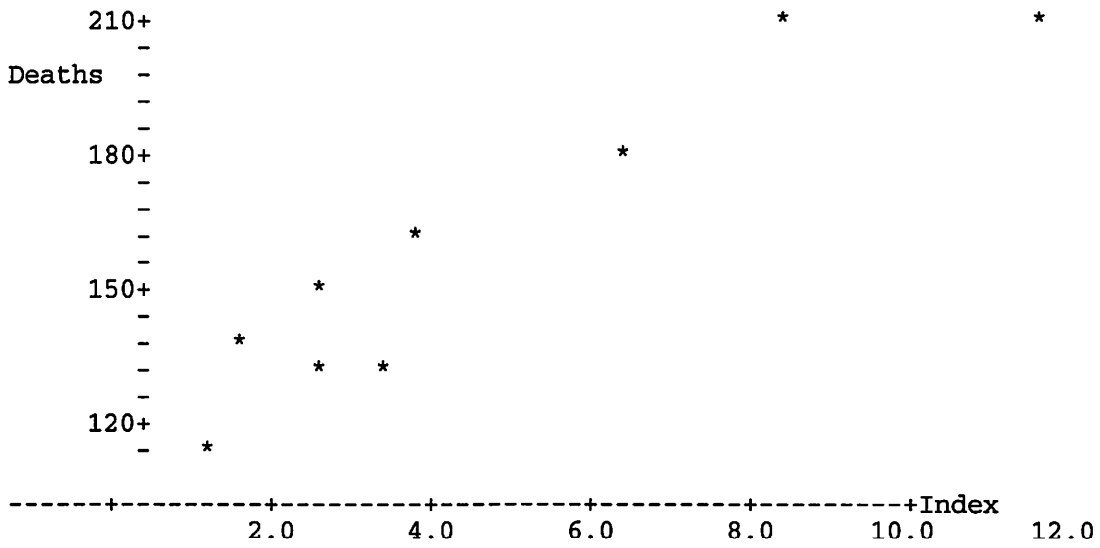
The following example illustrated this procedure. In an article published in the Journal of Environmental Health, May-June, 1965, Volume 27, Number 6, pp. 883-897, Robert Fadley listed the following data from the Columbia River Valley area: (2, p. 10)

Rate of Cancer versus Exposure Index

County/City	Index	Deaths
Umatilla	2.5	147
Morrow	2.6	130
Gilliam	3.4	130
Sherman	1.3	114
Wasco	1.6	138
Hood River	3.8	162
Portland	11.6	208
Columbia	6.4	178
Clatsop	8.3	210

The plot of this data is found on the following page.

Plot of the Rate of Cancer versus Exposure Index



Median-Median Line Approach

Grouping the points: Since there are nine data points, there will be 3 in each group.
Left-most group: {(1.3, 114), (1.6, 138), (2.5, 147)}
Middle group: {(2.6, 130), (3.4, 130), (3.8, 162)}
Right-most group {(6.4, 178), (8.3, 210), (11.6, 208)}
Summary Points: Use the median x-value and median y-value to form the summary point for each group.
Left-most group (1.6, 138)
Middle group: (3.4, 130)
Right-most group: (8.3, 208) Note: this point is not one of the given points
Equation of L: Use the two outer summary points to find the equation of L.

$$\text{slope} = \frac{208 - 138}{8.3 - 1.6} = 10.4$$

$y - 138 = 10.4(x - 1.6)$ or $y = 10.4x + 121.3$

Median-Median Line: When x = 3.4, the middle group's summary point, y = 156.7 for the line L. The distance from y = 165.7 to y = 130 is -26.7. One-third of 26.7 is -8.9. The median-median line passes through the point (3.4, 156.7- 8.9) = (3.4, 147.8) and its equation is

$y - 147.8 = 10.4(x - 3.4)$ and $y = 10.4x + 112.4$.

Note: The least squares approach obtains $y = 9.27x + 115$.

TRANSFORMATIONS OF DATA. Frequently, data sets do not display a linear pattern. In these cases a transformation of the data may produce transformed data that displays a linear pattern. The difficulty lies in deciding what transformation to try. In the book, Statistics, The Exploration and Analysis of Data, the authors Devore and Peck developed a method of examining the shape of the plot of the data to determine what transformation of the x and/or y values would produce a scatter plot with a linear pattern (3, p.176). The transformation one should try is described by the combinations of changes suggested in the following table and referring to the list below the chart

type	Trend in graph	x power	y power
1	concave down, neg. slope	increase	increase
2	concave up, pos. slope	increase	decrease
3	concave up, neg. slope	decrease	decrease
4	concave down, pos. slope	decrease	increase

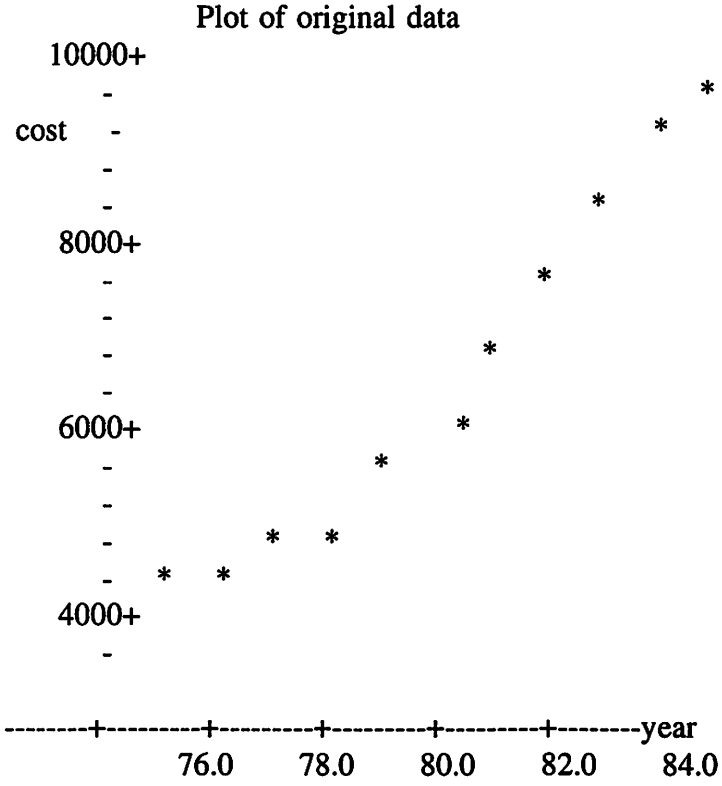
The list below shows the options that relate to the suggestions in the chart. Note that no change is indicated as power 1.

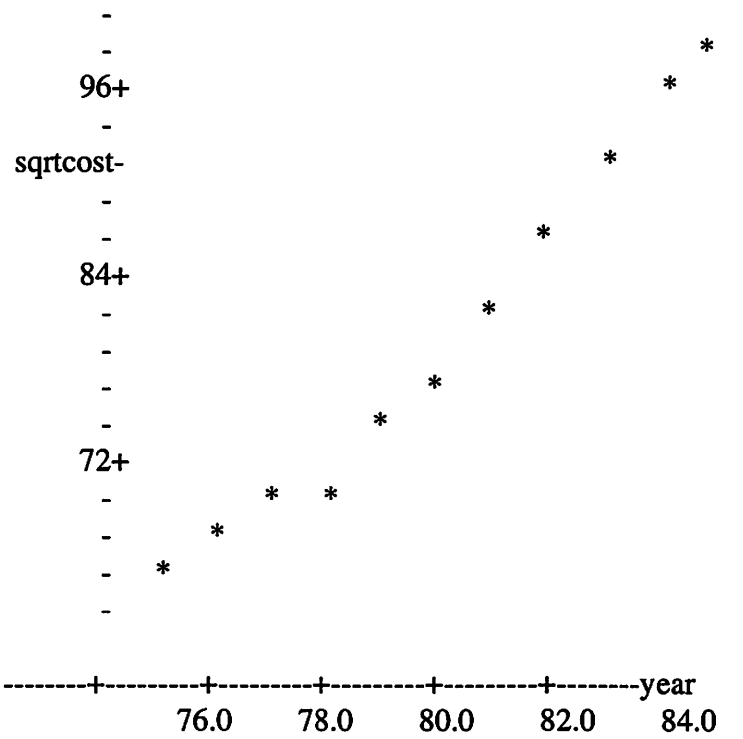
Power	Transformed Value	Name
3	$(\text{original value})^3$	Cube
2	$(\text{original value})^2$	Square
1	original value	No transformation
1/2	$\sqrt{\text{original value}}$	Square root
1/3	$\sqrt[3]{\text{original value}}$	Cube root
0	Log (original value)	Logarithm
-1	1/ (original value)	Reciprocal

COLLEGE COSTS EXAMPLE. This college cost data was obtained from The College Board News in Fall, 1985. It is used in the Contemporary Precalculus book to introduce the concept of transforming the data to bring it into a linear form. (2, p. 19,20)

Academic Year	Total Annual Cost
1975	$4205
1976	$4460
1977	$4680
1978	$4960
1979	$5510
1980	$6060
1981	$6845
1982	$7600
1983	$8435
1984	$9000
1985	$9659

The plot of this data shows a curvilinear pattern. The curvature is concave up with a positive slope so y should go down in power and/or x should go up in power. Taking the square root of y, we plot the original data and transformed data





The following Linear Regression table displays the least squares equation and statistical information about the coefficients. The R-sq (adj) is a correlation kind of measure regarding the amount of variation accounted for by the equation. R-sq = 97.9% is an excellent confirmation that this transformed data is well represented by this equation.

The regression equation is: sqrtcost = - 205 + 3.56 year

Predictor	Coef	Stdev	t-ratio	p
Constant	-205.19	13.10	-15.67	0.000
year	3.5619	0.1636	21.78	0.000

s = 1.715 R-sq = 98.1% R-sq(adj) = 97.9%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	1395.6	1395.6	474.25	0.000
Error	9	26.5	2.9		
Total	10	1422.1			

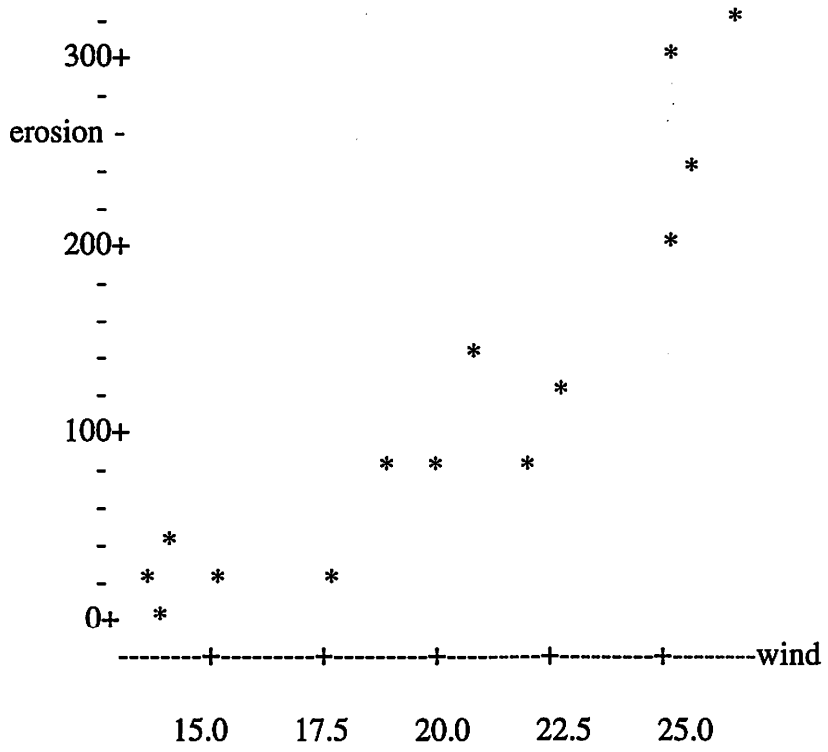
Since $\sqrt{y} = 3.56x - 205$, we obtain the cost = $12.673x^2 - 1459.6x + 42025$

SOIL EROSION BY WIND: This example also displays a concave upward pattern with a positive slope. Since it has more curvature than the previous example we may need less power than the square root for the change in y. The following data is reported in the paper "Soil Erosion by Wind from Bare Sandy Plains in Western Rajasthan, India", **Journal of Arid Environment**, 1981, p. 15-20 (3, p. 181). This paper "reported on a study of the relationship between wind velocity x (km/h) and soil erosion y (kg/d) in this very dry environment, where erosion control is especially important".

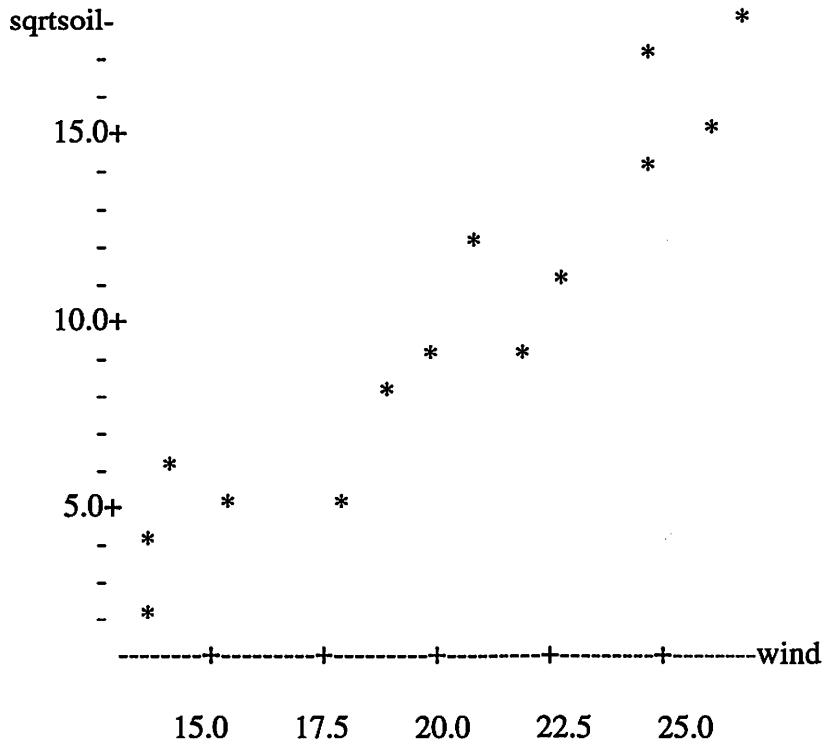
Observation	x	y	Observation	x	y
1	13.5	5	8	21	140
2	13.5	15	9	22	75
3	14	35	10	23	125
4	15	25	11	25	190
5	17.5	25	12	25	300
6	19	70	13	26	240
7	20	80	14	27	315

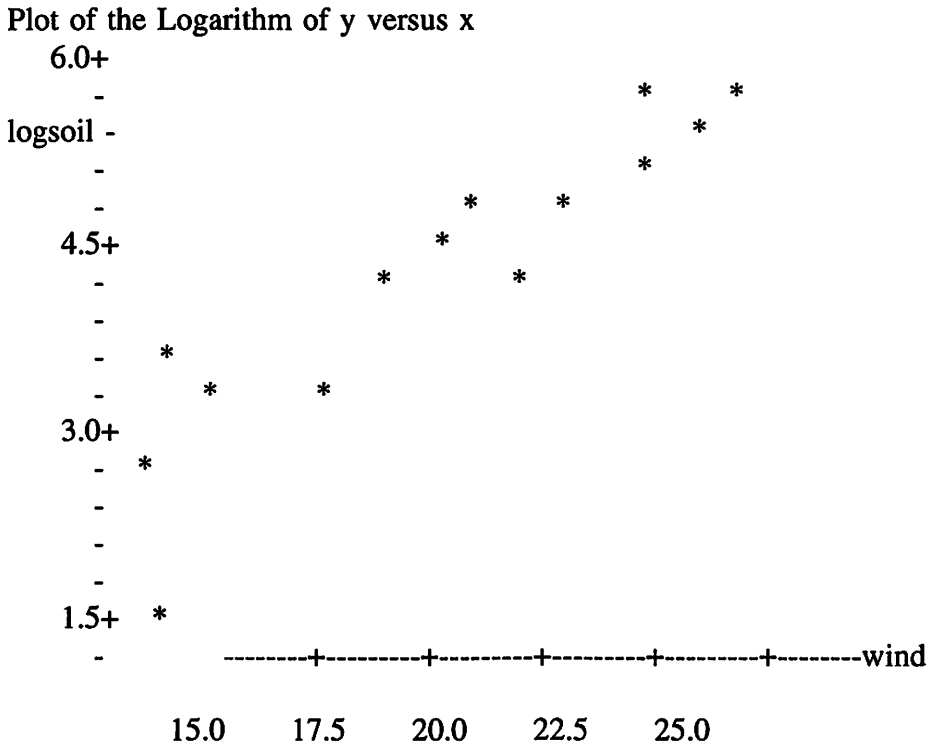
To produce a linear pattern I will try two power reduction transformations of y, the square root of y and the logarithm of y.

Plot of the original data



Plot of the Square Root of y versus x





The linear regression equation for the square root of y accounts for more of the variation than the logarithm linear regression equation as revealed in the following tables. Thus, the square root transformation will be selected for this data.

Square Root Transformation Regression table

The regression equation is sqrtsoil = - 10.3 + 0.994 wind

Predictor	Coef	Stdev	t-ratio	p
Constant	-10.322	1.980	-5.21	0.000
wind	0.99417	0.09593	10.36	0.000

s = 1.666 R-sq = 90.0% R-sq(adj) = 89.1%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	298.08	298.08	107.41	0.000
Error	12	33.30	2.78		
Total	13	331.39			

Logarithm Transformation Regression Table

The regression equation is logsoil = - 0.560 + 0.238 wind.

Predictor	Coef	Stdev	t-ratio	p
Constant	-0.5600	0.5485	-1.02	0.327
wind	0.23820	0.02658	8.96	0.000

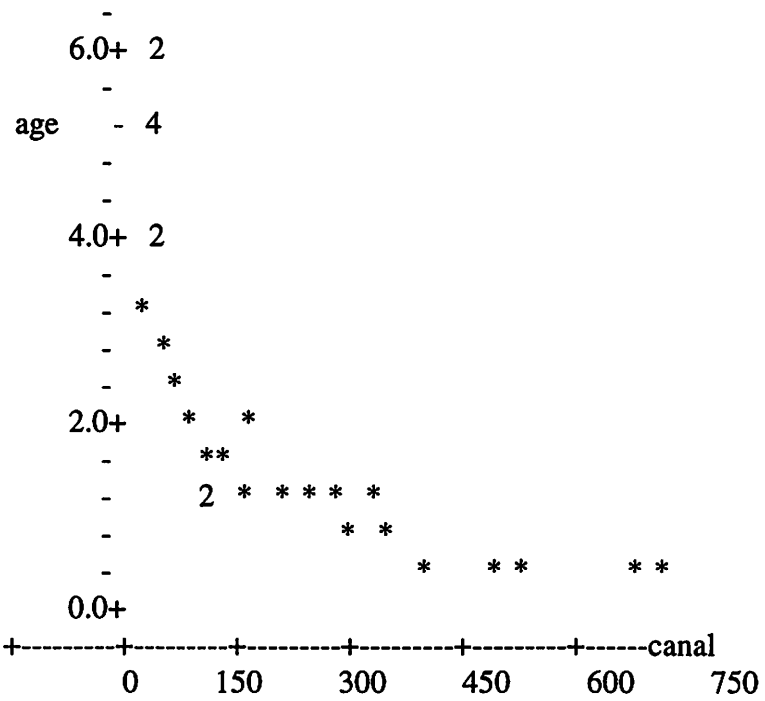
s = 0.4616 R-sq = 87.0% R-sq(adj) = 85.9%

Analysis of Variance

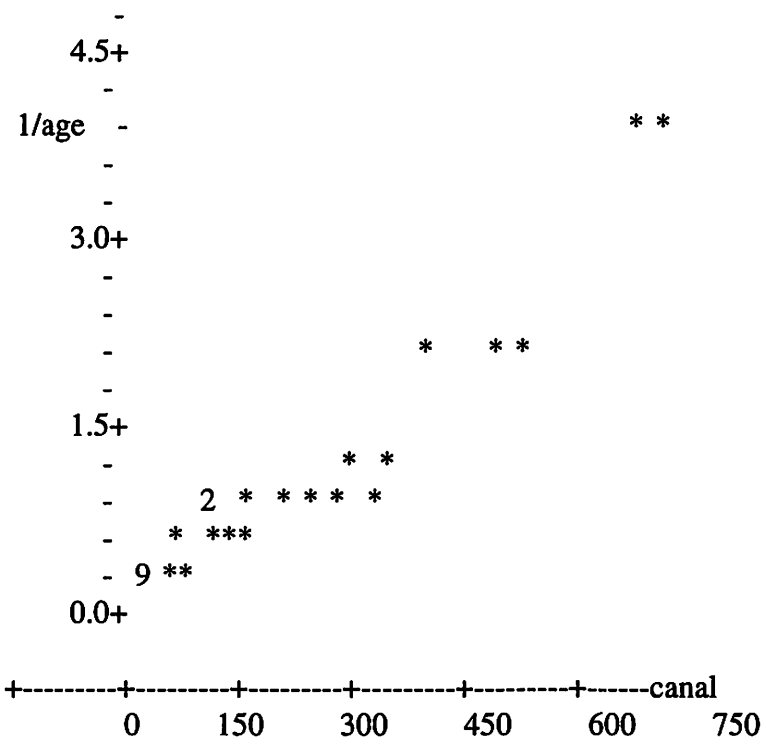
SOURCE	DF	SS	MS	F	p
Regression	1	17.111	17.111	80.32	0.000
Error	12	2.556	0.213		
Total	13	19.668			

Note that the R-sq value for the square root transformed is 90.0% whereas the R-sq value for the logarithmic transformed data is 87.0%. Since the square root transformed data best accounts for the variation, as measured by the R-sq value, the following model will be chosen for the original data (i.e. the erosion due to wind is a quadratic model). $y = (.994x - 10.3)^2 = .998x^2 - 20.48x + 106.1$

FINAL EXAMPLE: This example comes from (3, p. 186,7) and regards a paper entitled "Validation of age estimation in Harp Seals Using Dentinal Annuli", **Canadian Journal of Fisheries and Aquatic Science**, 1983: p 1430-1441. For seals whose age was known they tested this procedure to estimate age x (years) by the measure of the pulp canal y in the seals canine teeth. The plot for this data is given below. It is clear that this plot is concave up with a negative slope. A transformation decreasing the power of y will be tried. The shape suggests the 1/y transformation.



The Transformed Plot with 1/y versus x



The regression analysis indicates that this transformed data is well represented by the linear equation 1/age = 0.102 + 0.468 canal.

The regression equation is C3 = 0.102 + 0.00468 canal.

Predictor	Coef	Stdev	t-ratio	p
Constant	0.10243	0.07683	1.33	0.194
canal	0.0046833	0.0002797	16.74	0.000

s = 0.3034 R-sq = 91.2% R-sq(adj) = 90.9%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	25.807	25.807	280.33	0.000
Error	27	2.486	0.092		
Total	28	28.292			

The model that best relates age to canal length is age = $\frac{1}{0.102 + 0.468\text{canal}}$

In each of the above examples I have tried to illustrate how data obtained in an experiment can often be modeled by transforming the data, according to the guidelines given earlier in the paper, to produce a nearly linear data set. If the linear regression

(or median-median line) provides a good model for the transformed data, the inverse transformation will identify the appropriate model for the original data.

All of the transformations I have used are available to a student in an Intermediate Algebra class. Using the median-median line would provide a means to apply these ideas to such a group of students. For collegiate students who have had some statistics, the linear regression analysis provides more power.

In closing I want to note that linear regression requires that the data be normally distributed and that the variance is equal at all x levels. If these assumptions are not met one should use non-parametric methods. In addition, an analysis of residuals would provide another measure of the adequacy the model.

Summary

1. Building up mathematical models through differential equations is only one aspect of model building.
2. In industry, most models are developed from data that is gathered in experiments.
3. Starting with a plot of the data acquaints students with the reality of random error and "noise" in data.
4. The linear regression model is very useful in mathematical model building when it is combined with a transformation of data.
5. At the pre-calculus level, the median-median line approach can be combined with transformations to have a similar experience of building mathematical models.
6. These statistical methods of model buliding are useful to the extent that the assumptions underlying the statistical procedure are met.

Bibliography

1. AMATYC, **Standards for Curriculum and Pedagogical Reform in Two-Year Colleges and Lower Division Mathematics** (draft), AMA, 1993

2. Dept. of Mathematics, North Carolina School of Science and Mathematics, **Contemporary Precalculus through applications**, Janson Publications, Dedham, MA, 1992

3. Devore and Peck, **Statistics, The Exploration and Analysis of Data**, Duxbury Press, Belmont, CA, 1993