REGRESSION ON MEDIANS OF PROBABILITY DISTRIBUTIONS

Martha R. Sumler, ManTech Environmental Technology, Inc.
Henry L. Fisher, U.S. Environmental Protection Agency, HERL
Larry L. Hall, U.S. Environmental Protection Agency, HERL

Abstract

The median is a fundamental parameter in the area of lifetime and survival statistics. In toxicology, the LD50, lethal dose that results in 50% mortality, is frequently used. The median is also used to describe the incidence of cancer and other disease states. Factors such as nutritional status, age of animal, and exposure to a second chemical can cause the LD50 to shift. It is therefore desirable to determine a functional relationship between the median of a distribution and a cofactor. In this paper we use the SAS System to examine the use of median regression to predict a continuous dependent variable as a function of a single cofactor and compare these results to the standard ordinary least squares regression techniques.

Two data sets were generated using the SAS RANUNI and NORMAL functions. In one, the median line was proportional to the mean line, and both the median and mean had positive slopes with respect to a cofactor. In the other, the slope of the median line was positive while that of the mean line was negative. Thus, the values of the median increased and the mean decreased as the cofactor increased. Medians in both data sets were calculated using discrete and running intervals. Results of median and mean regressions can contradict each other in some cases, as shown with the second data set. Performing both mean and median regressions can provide a more complete picture of the statistical relationship between the dependent and independent variables.

Introduction

Probability distributions can be characterized in several ways. One way is through moments, i.e., mean, variance, skewness, kurtosis, etc. Another way is through percentiles, such as the 50th percentile (median) or other percentiles (0-100). The appropriate description of a distribution depends greatly upon its subsequent use. Also, convention, as well as ease of use, are important determinants of how a distribution is characterized.

As well as being a fundamental characteristic of a probability distribution in the percentile system, the median, where the cumulative failure reaches 50%, is an important parameter in lifetime theory. In toxicology, the LD50, lethal dose resulting in 50% mortality, is used in the initial evaluation of new compounds, since it can be obtained from a small number of animals (Trevan, 1927). More animals are required to determine the slope of the cumulative distribution than the LD50, while even more animals are required to determine the shape of the distribution or percentiles such as LD10 or LD90. Sometimes the shape of the distribution may be approximately normal, while for many cases the shape is approximately lognormal. Since the lognormal distribution is non-symmetric, the mean and median are unequal. Not only is there a probability distribution on one variable (factor), such as time or dose, but there may also be another variable (cofactor), such as body weight or age, on which the location and shape of the probability distribution depends. The median and other percentiles may then be dependent on the value that this other variable or cofactor assumes. Since the median, and perhaps other percentiles, are the major parameters of interest for an unspecified distribution shape, the objective here is to determine the median and how it changes with values taken on by a cofactor.

Data Set Construction

Two data sets with 100 points each and known properties were constructed to demonstrate the use of median regression as an analysis tool and to compare how median regression could be implemented in the SAS System. An annotated listing of the code used to generate the data sets, calculate medians, and run regressions is attached as the appendix.

Values of the covariate variables, or x-values, for both data sets were generated randomly from 0 < x < 1 by sampling from a uniform distribution using the RANUNI function. For each of these x-values, a y-value was generated from a lognormal distribution (Aitchison, 1969) by the following equation:

\[ y = 10^{\log \xi + \sigma \cdot N(0,1)} \]  

where \( \log \xi \) and \( \sigma \) are the mean and standard deviation of a normal distribution and \( N(0,1) \) is a random value from the NORMAL function with a mean of 0 and a variance of 1. Two test data sets were produced based on the assumptions described below.

In the first, the median of y was chosen to vary linearly with \( x \), with a constant standard deviation, according to the following equation:

\[ \text{median } y = \xi (x) = 10 + 90x, \quad \sigma = 0.5 \]  

This produced the following mean of y:

\[ \text{mean } y = \xi (x) 10^{\log (.52) / 2 \log e} = 1.94(10+90x) \]
Figure 1:
Lognormal data set, median = 10 + 90 x, $\sigma = 0.5$ log units, with theoretical median (-----) and mean (.....) lines, and the fitted least squares line (-----) and method 2 fitted median line (---) presented logarithmically in la and linearly in lb and lc. Panel c is an enlargement of some of the detail lost in panel b. In Id are the 10 medians ( o ) used in method 1 and the 91 medians ( X ) used in method 2, along with the theoretical median line (-----) and the calculated median lines for method 1 (- - -) and method 2 (-- --).

This resulted in the mean and median lines being equally distant on the log scale, but having different slopes and different intercepts on a linear scale.

In the second set, an increasing median and decreasing mean as x increased was chosen. This caused the variance of y to decrease as x increased. The median of y varied linearly with x:

$$\text{median } y = \xi (x) = 10 + 21.62 \times$$  

The mean of y was:

$$\text{mean } y = \mu (x) = 10 + (2 \times 2 \times \log \sigma) = 100 - 67.50 \times$$

This mean of y varied linearly with x, thus defining $\sigma (x)$.

Figure 1 shows the data points generated for the first data set. Curves calculated from equations 2 and 3 show the theoretical median and mean. The data are presented in log units in panel a, and linear units in panel b. Panel c is an enlargement of some of the detail lost in panel b. Figure 2 shows similar information for data set 2. The theoretical curves, shown in figure 2, were calculated using equations 4 and 5.

Method of Analysis

A discrete median method and a running median method, given by Hosteller and Tukey (1977), were used to analyse each data set. In the first method, the values of the covariate x were partitioned into a number of non-overlapping sets and the median in each set was determined. Various functions, such as spline, power series ($a + bx + cx^2 + ...$), or Fourier series, can be used to represent the sequence of medians. In
this example, which was constructed with simple linear properties, a linear function was used to fit the medians. In the second method, a discrete interval of 10 data points was selected and running medians were computed as this interval scanned the range of x-values. A linear function was fit to the medians produced by the second method also.

For the first method, discrete median method, the data set of 100 points was ordered on x with PROC SORT, and 10 data subsets of 10 points each in sequence from low to high were produced. PROC UNIVARIATE was applied to each subset to determine a median for the x-values and a median for the y-values in each subset. PROC GLM with

MODEL ymedian = xmedian

was used with these 10 median pairs, giving a first estimate for the linear function. Since the location of the probability distribution function was a function of x, the distribution changed location across a subset also. The subset in an x-interval consists of composite data from many distributions whose medians are different, i.e. the median varies across the x-interval. An iterative procedure was used to produce a correction factor to counteract this. This was done by treating the residuals of the 100 data points from the first GLM fit as values of a dependent variable in a second GLM iteration. The predicted values from the first and second GLM iterations were summed to produce the total predicted value, i.e. variable y was summed for the two lines.

For the second method, running median method, the data set of 100 points was ordered on x with PROC SORT, and a running median was calculated from sets of 10 consecutive points consisting of 1-10, 2-11, ..., 91-100 to produce 91 sets. PROC UNIVARIATE was applied to each of the 91 subsets to determine the medians on both x and y and PROC GLM was used to fit a line to these 91 medians. As described previously, a second GLM iteration provided the correction factor needed to produce the final line.
Table 1. Slope and Intercept of Regression Lines for Medians

<table>
<thead>
<tr>
<th></th>
<th>Method 1 (n = 10)</th>
<th>Method 2 (n = 91)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept ± SE</td>
<td>Slope ± SE</td>
</tr>
<tr>
<td><strong>Data Set 1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iteration 1°</td>
<td>5.5 ± 8.9</td>
<td>93.5 ± 15.0</td>
</tr>
<tr>
<td>Iteration 2°</td>
<td>-0.4 ± 9.3</td>
<td>3.0 ± 15.7</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>5.1 ± 9.3</td>
<td>96.5 ± 15.7</td>
</tr>
<tr>
<td><strong>True Value</strong></td>
<td>10.0</td>
<td>90.0</td>
</tr>
<tr>
<td><strong>Data Set 2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iteration 1°</td>
<td>14.4 ± 8.4</td>
<td>22.7 ± 14.9</td>
</tr>
<tr>
<td>Iteration 2°</td>
<td>0.1 ± 8.3</td>
<td>-1.0 ± 14.8</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>14.5 ± 8.3</td>
<td>21.7 ± 14.8</td>
</tr>
<tr>
<td><strong>True Value</strong></td>
<td>10.0</td>
<td>21.62</td>
</tr>
</tbody>
</table>

1°For iteration 1, the slope and intercept were determined from the data. For iteration 2, the slope and intercept were determined from the residuals of iteration 1 and, therefore, were corrections to the intercept and slope of the original line.

2°The standard errors for method 2 were not calculated because of correlations among the 91 values but would be asymptotic to the standard errors found by method 1.

3°Uncorrected estimate using the data.

4°Correction value obtained from using the residuals.

5°Corrected estimate obtained from sum of values from iterations 1 and 2.

6°Parameters from the theoretical distribution.

Results

Figures 1a, 1b, and 1c show the data from set 1, in log10 scale and linear scale. Figure 1c is a detail of 1b. Also shown in these figures are the theoretical mean and median lines of the underlying population, and the GLM least squares (mean) line and the running median line, i.e. method 2. The theoretical mean and median lines are separated by a factor of 1.94 for this lognormal distribution with large variance. A comparison of the theoretical mean line (intercept = 19.4, slope = 174.6) with the GLM least squares line (intercept = 44.8 ± 28.5, slope = 103.8 ± 48.4) shows that there is a large difference. This is due to the large variance of the distribution and small sample size of 100 points, which consequently produces a relatively large standard error of at least 47% for both the slope and intercept of the GLM line. The median line from method 2 is shown in figures 1a, 1b, 1c, and 1d and lies near the theoretical one for the population. The parameter estimates for both methods 1 and 2 are shown in Table 1 and the lines are shown in figure 1d. Since both methods perform equally well, there is little to recommend either method 1, discrete medians, or method 2, running medians, as the data in table 1 and figure 1d show. There are no standard errors for the parameters for method 2 because the 91 moving medians are not independent. However, these standard errors would be asymptotic to those calculated from method 1 because the 10 discrete medians on which the fit is based are statistically independent. The correction factors are small, as shown in table 1 and not significantly different from zero. Thus, using only the first iteration, uncorrected analysis, may be preferable in this case.

Figures 2a, 2b, 2c, and 2d show the results for data set 2. Here, the theoretical mean of the underlying population decreases, while the theoretical median increases. Analysis of the sample of 100 points confirms this behavior through the least squares line and the median lines respectively. Although the theoretical median line (intercept = 100, slope = -67.5) and the GLM least squares line (intercept = 166.6 ± 88.5, slope = -166.7 ± 156.4) appear quite different in figures 1a, 1b, and 1c, the intercept and slope of the theoretical mean line are within the confidence limits of the GLM least squares line. The two methods of finding
the median line produced approximately the same results, as shown in table 1 and figure 2d. Intercepts determined by the two methods were closer to each other, 14.5 and 13.0, than to the expected intercept of 10.0. A larger sample than 100 points may be necessary to make analysis and theory agree as there must be variability due to sampling as well as analysis. As in the first data set, the correction factor was small and not significantly different from zero.

Conclusions

- Median regression was implemented in the SAS System and applied to two non-symmetric test distributions. The obtained results approximated the known values.
- The Mosteller-Tukey methods of determining running medians or median in discretized intervals were shown to approximate the known theoretical values when applied to two highly skewed lognormal distributions.
- The accuracy of the linear regression line was asymptotically the same for running and discrete medians. Thus these examples suggest that both methods perform equally well.
- The running median graph has finer detail than the discrete median, but contains equally appropriate information for determining regression lines.

Acknowledgments

The authors wish to express their thanks to Dr. Woody Setzer for his beneficial consultation.

References


SAS and SAS/GRAPH are registered trademarks of SAS Institute Inc., Cary, NC, U.S.A.
/* TAKE OUT ORIGINAL DATA & CALCULATE RESIDUALS */
DATA RESIDU;
  SET RESID; IF RES=., AND CT NE .1, RESY=2-PRED; 
  PROC SORT DATA=RESID; BY CT X2; /* SORT & PRINT BY X2 */ 
  PROC PRINT DATA=RESID; TITLE "RESID"; 
  /* SORT AND PRINT BY RES */ 
  PROC SORT DATA=RESID OUT=RESORD; BY CT RESY; 
  PROC PRINT DATA=RESID; TITLE "RESORD"; 
  /* BY CT RESY */ 
  PROC PRINT DATA=RESORD; ID CT; TITLE "RESORD"; 
END MACRO /* MACRO TO PROCESS RESIDUALS */
/* END MACRO */
/* MACRO TO CONCATENATE DATA SETS CONTAINING MEDIANS */
MACRO CONCAT;
  DATA CONCAT;
  SET CREATE FIRST+1; RETAIN LAST 
  ZOO Z1 ZTO 91; 
  SET XY; VAR X Y; /* OUTPUT OUT=CONCAT */ 
  DATA CONCAT;
  SET CREATE FIRST+1; 
  /* CONCATENATE CREATED DATA, MED OF RES, & RESIDUALS */ 
DATA PASS2;
  SET CALC2 MEDIB RESID2; 
  PROC PRINT DATA=PASS2; TITLE "PASS2"; 
  /* SECOND GLM - ON MEDIANS OF RESIDUALS */ 
  PROC GLM DATA=PASS2; 
  MODEL YR=WEIGHT HEIGHT; 
  OUTPUT OUT=PASS2 RESID=R predicted=PRED_R; 
  PROC PRINT DATA=PASS2; TITLE "PASS2"; 
END MACRO /* END MACRO */
MACRO CONCAT;
  DATA CONCAT;
  SET CREATE FIRST+1; RETAIN LAST 
  ZOO Z1 ZTO 91; 
  SET CREY; VAR X Y; /* OUTPUT OUT=CONCAT */ 
  DATA CONCAT;
  SET CREATE FIRST+1; 
  /* CONCATENATE CREATED DATA, MED OF RES, & RESIDUALS */ 
DATA PASS3;
  SET CALC2 MEDIB RESID2; 
  PROC PRINT DATA=PASS3; TITLE "PASS3"; 
  /* SECOND GLM - ON MEDIANS OF RESIDUALS */ 
  PROC GLM DATA=PASS3; 
  MODEL YR=WEIGHT HEIGHT; 
  OUTPUT OUT=PASS3 RESID=R predicted=PRED_R; 
  PROC PRINT DATA=PASS3; TITLE "PASS3"; 
END MACRO /* END MACRO */
MACRO CONCAT;
  DATA CONCAT;
  SET CREATE FIRST+1; RETAIN LAST 
  ZOO Z1 ZTO 91; 
  SET CREY; VAR X Y; /* OUTPUT OUT=CONCAT */ 
  DATA CONCAT;
  SET CREATE FIRST+1; 
  /* CONCATENATE CREATED DATA, MED OF RES, & RESIDUALS */ 
DATA PASS4;
  SET CALC2 MEDIB RESID2; 
  PROC PRINT DATA=PASS4; TITLE "PASS4"; 
  /* SECOND GLM - ON MEDIANS OF RESIDUALS */ 
  PROC GLM DATA=PASS4; 
  MODEL YR=WEIGHT HEIGHT; 
  OUTPUT OUT=PASS4 RESID=R predicted=PRED_R; 
  PROC PRINT DATA=PASS4; TITLE "PASS4"; 
END MACRO /* END MACRO */
MACRO CONCAT;
  DATA CONCAT;
  SET CREATE FIRST+1; RETAIN LAST 
  ZOO Z1 ZTO 91; 
  SET CREY; VAR X Y; /* OUTPUT OUT=CONCAT */ 
  DATA CONCAT;
  SET CREATE FIRST+1; 
  /* CONCATENATE CREATED DATA, MED OF RES, & RESIDUALS */ 
DATA PASS5;
  SET CALC2 MEDIB RESID2; 
  PROC PRINT DATA=PASS5; TITLE "PASS5"; 
  /* SECOND GLM - ON MEDIANS OF RESIDUALS */ 
  PROC GLM DATA=PASS5; 
  MODEL YR=WEIGHT HEIGHT; 
  OUTPUT OUT=PASS5 RESID=R predicted=PRED_R; 
  PROC PRINT DATA=PASS5; TITLE "PASS5"; 
END MACRO /* END MACRO */
MACRO CONCAT;
  DATA CONCAT;
  SET CREATE FIRST+1; RETAIN LAST 
  ZOO Z1 ZTO 91; 
  SET CREY; VAR X Y; /* OUTPUT OUT=CONCAT */ 
  DATA CONCAT;
  SET CREATE FIRST+1; 
  /* CONCATENATE CREATED DATA, MED OF RES, & RESIDUALS */ 
DATA PASS6;
  SET CALC2 MEDIB RESID2; 
  PROC PRINT DATA=PASS6; TITLE "PASS6"; 
  /* SECOND GLM - ON MEDIANS OF RESIDUALS */ 
  PROC GLM DATA=PASS6; 
  MODEL YR=WEIGHT HEIGHT; 
  OUTPUT OUT=PASS6 RESID=R predicted=PRED_R; 
  PROC PRINT DATA=PASS6; TITLE "PASS6"; 
END MACRO /* END MACRO */
MACRO CONCAT;
  DATA CONCAT;
  SET CREATE FIRST+1; RETAIN LAST 
  ZOO Z1 ZTO 91; 
  SET CREY; VAR X Y; /* OUTPUT OUT=CONCAT */ 
  DATA CONCAT;
  SET CREATE FIRST+1; 
  /* CONCATENATE CREATED DATA, MED OF RES, & RESIDUALS */ 
DATA PASS7;
  SET CALC2 MEDIB RESID2; 
  PROC PRINT DATA=PASS7; TITLE "PASS7"; 
  /* SECOND GLM - ON MEDIANS OF RESIDUALS */ 
  PROC GLM DATA=PASS7; 
  MODEL YR=WEIGHT HEIGHT; 
  OUTPUT OUT=PASS7 RESID=R predicted=PRED_R; 
  PROC PRINT DATA=PASS7; TITLE "PASS7"; 
END MACRO /* END MACRO */
MACRO CONCAT;
  DATA CONCAT;
  SET CREATE FIRST+1; RETAIN LAST 
  ZOO Z1 ZTO 91; 
  SET CREY; VAR X Y; /* OUTPUT OUT=CONCAT */ 
  DATA CONCAT;
  SET CREATE FIRST+1; 
  /* CONCATENATE CREATED DATA, MED OF RES, & RESIDUALS */ 
DATA PASS8;
  SET CALC2 MEDIB RESID2; 
  PROC PRINT DATA=PASS8; TITLE "PASS8"; 
  /* SECOND GLM - ON MEDIANS OF RESIDUALS */ 
  PROC GLM DATA=PASS8; 
  MODEL YR=WEIGHT HEIGHT; 
  OUTPUT OUT=PASS8 RESID=R predicted=PRED_R; 
  PROC PRINT DATA=PASS8; TITLE "PASS8"; 
END MACRO /* END MACRO */
MACRO CONCAT;
  DATA CONCAT;
  SET CREATE FIRST+1; RETAIN LAST 
  ZOO Z1 ZTO 91; 
  SET CREY; VAR X Y; /* OUTPUT OUT=CONCAT */ 
  DATA CONCAT;
  SET CREATE FIRST+1; 
  /* CONCATENATE CREATED DATA, MED OF RES, & RESIDUALS */ 
DATA PASS9;
  SET CALC2 MEDIB RESID2; 
  PROC PRINT DATA=PASS9; TITLE "PASS9"; 
  /* SECOND GLM - ON MEDIANS OF RESIDUALS */ 
  PROC GLM DATA=PASS9; 
  MODEL YR=WEIGHT HEIGHT; 
  OUTPUT OUT=PASS9 RESID=R predicted=PRED_R; 
  PROC PRINT DATA=PASS9; TITLE "PASS9"; 
END MACRO /* END MACRO */
MACRO CONCAT;
  DATA CONCAT;
  SET CREATE FIRST+1; RETAIN LAST 
  ZOO Z1 ZTO 91; 
  SET CREY; VAR X Y; /* OUTPUT OUT=CONCAT */ 
  DATA CONCAT;
  SET CREATE FIRST+1; 
  /* CONCATENATE CREATED DATA, MED OF RES, & RESIDUALS */