PREDICTION INTERVALS FOR LOG-LINEAR REGRESSION

By

Alastair J. Scott
and
Michael J. Symons

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 714

October 1970

PREDICTION INTERVALS FOR LOG-LINEAR REGRESSION

Alastair J. Scott* and Michael J. Symons
The University of North Carolina at Chapel Hill

In situations where the log transformation is used to make the assumptions of the ordinary linear regression model more reasonable, interval estimates for future observations are often desired in the scale of the original measurements. Common practice is to transfer the usual equal tail interval in the transformed scale back to the original scale. Conditions under which this procedure gives reasonably short intervals as compared with the shortest possible interval having the same content are derived, and simple procedures that give reasonably short intervals are suggested when it is not satisfactory.

1. Introduction

Suppose that we have a set of observations on a response variable z with corresponding measurements on p independent variables $x_1,\ldots,x_p$ and that the transformed variable y = log z satisfies the assumptions of the usual linear regression model to an adequate degree of approximation. Standard regression techniques will produce estimates and predictions expressed in the transformed scale. Sometimes these new units will have a genuine physical meaning and it is preferable to work with them, but more often estimates and predictions expressed in the scale of the original measurements are desired. Most attention has been paid to estimation, particularly of the expected value of the variable z. Finney (1951), Aitchison and Brown (1957), Heien (1968),

---

*This work was done while on leave from the London School of Economics during 1969-1970.

Goldberger (1968), Land (1969), and Bradu and Mundlak (1970) all discuss this problem. For many purposes the median of z is more useful than the mean since the log-normal distribution is markedly skewed. This paper discusses interval estimates for the median of z and also deals with the problem of predicting values of new observations on the response variable z for given $x_1,\ldots,x_p$.

Both problems are somewhat simpler than estimating the expected value of z. For example, standard techniques give 1-α prediction intervals with equal upper and lower tail areas for future values of y = log z, and common practice of simply transforming the end-points of this interval gives a valid 1-α prediction interval for values of z. However, we can obtain shorter intervals in the scale of the original measurements by decreasing the lower tail area and making the upper tail larger. In the following sections we compare the commonly-used equal tail interval with the shortest possible interval and find conditions under which the inflation in length is small. We also suggest simple alternatives that give reasonable results when these conditions are not met. In particular, the shorter of the usual equal-tail area interval and one-sided interval is always fairly good.

2. Preliminary Discussion

Suppose that there are n measurements $(z_1,x_{11},\ldots,x_{p1}),\ldots,(z_n,x_{1n},\ldots,x_{pn})$. Then if $y_i = \log z_i$ (i=1,...,n) we have

$$\underset{\sim}{y} = X'\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \qquad (2.1)$$

where $\varepsilon_1,\ldots,\varepsilon_n$ are independent normal random variables with mean zero and constant variance σ^2. Let $\hat{\underset{\sim}{\beta}} = (X'X)^{-1}\underset{\sim}{y}$ be the least-squares estimate of $\underset{\sim}{\beta}$ and $s^2 = (\underset{\sim}{y}-X'\hat{\underset{\sim}{\beta}})'(\underset{\sim}{y}-X'\hat{\underset{\sim}{\beta}})/(n-p)$ be the usual unbiased estimate of σ^2. Then $\hat{y} = \underset{\sim o}{x}'\hat{\underset{\sim}{\beta}}$ is the minimum variance unbiased estimate of $\mu = \underset{\sim o}{x}'\underset{\sim}{\beta}$, the value of the regression

line at $\underset{\sim}{x}=\underset{\sim}{x}_o$, and confidence intervals for μ can be constructed from the relation

$$\frac{\hat{y}-\mu}{s[\underset{\sim}{x}_o'(X'X)^{-1}\underset{\sim}{x}_o]^{\frac{1}{2}}} \sim t(n-p), \tag{2.2}$$

Similarly, standard techniques lead to confidence intervals for the value of a future observation on the transformed variable y, or more generally for the value of $\bar{y}_q$ the mean of q future observations on y at $\underset{\sim}{x}=\underset{\sim}{x}_o$, based on the relation

$$\frac{\bar{y}_q-\hat{y}}{S} \sim t(n-p) \tag{2.3}$$

where $S^2 = s^2(q^{-1}+\underset{\sim}{x}_o'(X'X)^{-1}\underset{\sim}{x}_o)$. (See Draper and Smith (1968), pp. 24 and 122.) The usual prediction interval for $\bar{y}_q$, $\hat{y} \pm St(1-\frac{1}{2}\alpha, n-p)$, with equal upper and lower tail areas is the shortest $1-\alpha$ confidence interval for $\bar{y}_q$ based on (2.3).

If the scale of the original measurements is of primary importance then we are interested in prediction intervals for values of the original variable z rather than the transformed variable $y = \log z$. The usual way of obtaining a confidence interval for z is to transform the equal tail area interval for y, producing

$$I_1 = [\exp\{\hat{y}-St(1-\tfrac{1}{2}\alpha, n-p)\}, \quad \exp\{\hat{y}+St(1-\tfrac{1}{2}\alpha, n-p)\}]. \tag{2.4}$$

This is a genuine $1-\alpha$ confidence interval for z since z is a monotone function of y, but different combinations of upper and lower tail areas may lead to $1-\alpha$ confidence intervals for z that are shorter. For example the extreme case of the one sided interval

$$I_2 = [0, \exp\{\hat{y}+St(1-\alpha, n-p)\}] \tag{2.5}$$

with lower tail area zero and upper tail area α will be shorter than I_1 if S is large enough. Similarly, the direct transform of the usual interval for $\mu=\underset{\sim}{x}_o'\underset{\sim}{\beta}$

gives a valid confidence interval for $\exp\{\mu\}$, the median (but not the mean) of observations on z at $\underset{\sim}{x}=\underset{\sim}{x}_o$. Again different combinations of upper and lower tail areas can produce shorter intervals in the scale of the original measurements.

All the discussion so far applies equally to confidence intervals in the ordinary frequency sense or to Bayes confidence intervals with conventional representations of vague prior knowledge. (See Lindley (1965).) Both types of intervals for μ or $\bar{y}_q$ are based on (2.2) and (2.3) respectively, and are completely equivalent for prespecified upper and lower tail areas. In the next section we look at the combinations of tail areas that give the shortest possible intervals for values of z or for the median $\exp\{\mu\}$, and here it is a little easier to work in the Bayesian framework since $\hat{y}$ and S are then regarded as fixed. Most of the results are immediately valid for either system of inference and the implications for confidence intervals are pointed out throughout.

In the Bayesian framework the median $\exp\{\mu\}$ is a random variable as well as future values of z and it is convenient to consider both cases together by defining $\theta = \exp\{\bar{y}_q\}$, where $\bar{y}_q$ is the mean q future observations on the transformed variable y as before. When $q=1$, θ is the value of a single observation on z, and when $q=\infty$, θ is the median $\exp\{\mu\}$ since $\bar{y}_\infty=\mu$. Intermediate values of q correspond to the geometric mean of future values of z which is sometimes of interest, particularily in economic applications.

In theory, the problem of finding the shortest interval having specified content for a quantity θ with a continuous density $p(\theta)$ is fairly simple. If L and U are the lower and upper limits of the shortest interval, then it is easy to show that L and U must satisfy

$$p(L) = p(U) \qquad (2.6)$$

with

$$\int_L^U p(\theta)d\theta = 1-\alpha \tag{2.7}$$

unless either L or U are boundary points of the region where the density is positive. If $p(\theta)$ is unimodal and is not constant over any interval the solution of (2.6) and (2.7) is unique. (See Lindley (1965) or Box and Tiao (1965).) Otherwise there may be several solutions and additional comparisons are necessary to determine which of the solutions is the shortest.

3. Known Variance

If σ^2 is known, $\log \theta$ is normally distributed with mean $\hat{y}$ and variance $S^2 = \sigma^2(q^{-1}+\underset{\sim}{x}_o'(X'X)^{-1}\underset{\sim}{x}_o)$ so that θ has a log-normal distribution with density

$$p(\theta) = \frac{1}{\sqrt{2\pi}S\theta} \exp\{-\tfrac{1}{2}(\frac{\log \theta-\hat{y}}{S})^2\} \tag{3.1}$$

for $\theta>0$. The distribution is unimodal with mode $\exp\{\hat{y}-S^2\}$, median $\exp\{\hat{y}\}$, and mean $\exp\{y+\tfrac{1}{2}S^2\}$. (Aichison and Brown (1957)) In this case equation (2.6) reduces to

$$\log L+\frac{(\log L-\hat{y})^2}{2S^2} = \log U+\frac{(\log U-\hat{y})^2}{2S^2} \tag{3.2}$$

which is equivalent to

$$[\log L-(\hat{y}-S^2)]^2 = [\log U-(\hat{y}-S^2)]^2. \tag{3.3}$$

Thus the shortest interval estimate for θ with probability content $1-\alpha$ has the following simple form:

$$[\exp\{\hat{y}-S^2-C_\alpha\}, \exp\{\hat{y}-S^2+C_\alpha\}] \tag{3.4}$$

where C_α is chosen so that the content is $1-\alpha$. This interval is the transform of a symmetric interval for y about $\hat{y}-S^2$, the log mode. The usual equal tail area interval is the transform of a symmetric interval about $\hat{y}$, the log median.

A comparison of the relative lengths is included as a special case of the unknown variance situation with infinite degrees of freedom in the next section. See Figure 1. Briefly, the shortest interval is approximately equal to the equal tail area interval for small S, and approaches the corresponding one-sided interval as S becomes large.

In this case the Bayesian and frequentist interval estimates coincide, since the shortest interval depends only on S and S is known when σ^2 is specified a priori.

4. Unknown Variance

When σ^2 is not known $(\log \theta-\hat{y})/S$ has a t-distribution with k=n-p degrees of freedom and it seems natural to call the resulting distribution of θ the log-t distribution by analogy with the normal case. The density is given by

$$p(\theta) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi k}\Gamma(\frac{k}{2})S\theta}\left[1+\frac{(\log \theta-\hat{y})^2}{kS^2}\right]^{-\frac{k+1}{2}} \qquad (4.1)$$

for $\theta>0$, which increases without bound as θ approaches zero.

Two cases need to be distinguished. If $S^2\geq(k+1)^2/4k$, $p(\theta)$ decreases monotonically as θ increases. As a consequence the shortest interval for θ is always one-sided with lower limit L=0 and upper limit $U = \exp\{\hat{y}+St(1-\alpha,k)\}$.

If $S^2<(k+1)^2/4k$, $p(\theta)$ decreases to a relative minimum at

$$\log \theta = \hat{y}- \frac{k+1}{2}\left[1+\left[1- \frac{4kS^2}{(k+1)^2}\right]^{\frac{1}{2}}\right],$$

increases to a relative maximum at

$$\log \theta = \hat{y} - \frac{k+1}{2}\left[1 - \left[\frac{4kS^2}{(k+1)^2}\right]^{\frac{1}{2}}\right],$$

and then decreases steadily as θ increases further. The distribution is not unimodal and there may be up to three solutions to equations (2.6) and (2.7) that need to be considered in addition to the one-sided interval discussed above. The other solutions cannot be expressed in a simple form but it is not difficult to compute them numerically. This was done for several values of α and the lengths were compared to find the shortest. Figure 1 illustrates the general situation with $\alpha=0.05$. It shows the lower tail area of the shortest interval as a function of the standard deviation for various degrees of freedom of s^2. Notice that with the degrees of freedom specified the shortest interval depends only on S. It has approximately equal tail areas for small S and approaches a one sided interval with no lower tail area as S increases. The value of S at which the one sided interval becomes a useful approximation to the shortest interval is smaller when there are fewer degrees of freedom. The shift of the shortest interval to one with a lower tail area less than $\frac{1}{2}\alpha$ is consistent with the results of Bradu and Mundlak (1970), who find a fractional multiplier of $\exp\{\hat{y}\}$ is required to produce an unbiased estimate of θ.

[Insert Figure 1,]

A comparison of the lengths of the equal tail area interval and one tail interval with that of the shortest interval showed that with given degrees of freedom the inflation corresponding to the interval I_1 (2.4) is a monotone increasing function of S; the inflation with I_2 (2.5) is a monotone decreasing

function of S. The conditions under which the usual equal tail interval results in 10%, 25%, or 50% inflation in length of the shortest interval are described in Table 1.

Table 1

Maximum Value of S for Which the Equal Tail Interval Has at Most a Specified Inflation in Length over the Shortest Interval at Selected Degrees of Freedom.

Percentage Inflation	Degrees of Freedom 1	2	3	5	10	∞
10%	0.05	0.17	0.24	0.32	0.39	0.46
25%	0.06	0.25	0.38	0.52	0.65	0.79
50%	0.09	0.34	0.53	0.77	1.00	1.32

This suggests a very simple rule: Choose the shorter of I_1 and I_2. No special tables or graphs are required for this rule and it performs well for any value of S. For example with α=0.05, the shorter of I_1 and I_2 is no more than 12% longer than the shortest possible interval. The situation is much the same for intervals of other content, as can be seen in Table 2 which shows the worst inflation compared to the length of the shortest interval for a few values of α.

Table 2

Maximum Inflation for the Rule: Choose the Shorter of I_1 and I_2.

α	0.01	0.05	0.10	0.20
Maximum Inflation	11%	12%	13%	16%

An alternative rule is to choose a symmetric interval about $\exp\{\hat{y}\}$ in the original scale; namely,

$$[\exp\{\hat{y}\}-C_\alpha, \quad \exp\{\hat{y}\}+C_\alpha],$$

where C_α is choosen so that the interval content is $1-\alpha$. This has some intuitive appeal and performs almost as well as the shorter of I_1 and I_2 rule. If S is large, it becomes the one sided interval I_2, but for smaller values of S it is similar to I_1. However, this symmetric interval cannot be calculated directly and an iterative procedure is needed to find the end-points unless special tables or graphs are constructed.

The choice of interval estimates discussed in this section depends upon the sample quantity s^2. This poses no problem for a Bayesian, but the resulting interval does not have exact $1-\alpha$ confidence in the ordinary frequency sense. Strictly speaking a frequentist should prespecify the lower tail area for his interval and not allow it to depend on the sample outcome. The information provided by Figure 1 and Tables 1 and 2 would still be useful in choosing a lower tail area if there were some prior knowledge about the magnitude of σ^2. Since the two points of view coincide when σ^2 is specified, it would be surprising if the approaches differ substantially even with only a few degrees of freedom.

5. Example

Table 3 gives values of the bacterial count z and the concentration of free chlorine x obtained from a routine surveillance of bacterial levels in swimming pools. We are grateful to Roy A. Paul (1970) for supplying the data. Although several other factors such as the number of swimmers, the size of the pool, and the acidity of the water affect the bacterial level, it was hoped that a reasonable prediction of the level could be based on the chlorine concentration alone. For small concentrations the assumption that y, the natural logarithm of the bacterial count, was distributed about the line $\alpha+\beta x$ with constant variance seemed to be a good approximation, although there may be a preference for using the inverse concentration as the independent variable.

[Insert Figure 2.]

Figure 2 shows the scatter of observations about the fitted line: $y = 2.933-8.230x$. The estimated variance was $s^2=1.749$ and the value of S for a new observation at x_o is

$$S = s\left[1 + \frac{1}{n} + \frac{(\bar{x}-x_o)^2}{\Sigma(x_i-\bar{x})^2}\right]^{\frac{1}{2}}.$$

This varied from 1.336 to 1.402 for chlorine concentrations in the range from 0.00 mg/1 to 0.50 mg/1

[Insert Figure 3.]

Figure 3 shows the 95% equal tail intervals I_1 and 95% one tail intervals I_2 as a function of the chlorine concentration superimposed on a plot of the original data. For example, the interval estimate I_1 of a bacterial count at a chlorine concentration of 0.1mg/1 is (0.57, 119.98), a 54.9% inflation in length of the shortest interval which is essentially I_2, [0, 77.08). Since the lower limits of the intervals I_1 are close to zero as shown in Figure 3, the interval I_1 is roughly 50% longer than I_2 at all chlorine concentrations from 0.00 mg/1 to 0.50 mg/1.

The shortest interval estimate of the median using these data is interesting. The interval estimates of the future observations were markedly one sided, but for the median the equal tail interval does quite well. The value of S at x_o is

$$S = s\left[\frac{1}{n} + \frac{(\bar{x}-x_o)^2}{\Sigma(x_i-\bar{x})^2}\right]^{\frac{1}{2}}$$

and at x_o = 0.1 mg/1, S = 0.357. Although the shortest interval has a lower tail area of about 0.075 as shown in Figure 1, the inflation of the two tail

interval is less than 10%, in fact only 6.3%. Similar results hold for other values of x_o in the range of 0.00 mg/1 to 0.50 mg/1.

Table 3

Paul's Data: Bacterial Counts in Swimming Pools at Various Chlorine Concentrations.

z = Bacterial count in number of bacteria per millimeter
x = Concentration of free chlorine in milligrams per liter

z	x	z	x	z	x	z	x
12.0	0.02	150.0	0.00	0.5	0.17	49.0	0.07
3.3	0.01	96.0	0.00	1.0	0.15	20.0	0.15
2.9	0.07	86.0	0.00	5.1	0.10	1.4	0.29
1.5	0.23	32.0	0.01	0.4	0.26	2.9	0.11
4.5	0.16	2.4	0.24	7.8	0.19	1.8	0.25
35.0	0.09	5.9	0.19	0.9	0.09	8.3	0.13
5.1	0.05	21.0	0.01	1.9	0.17	212	0.08
19.0	0.01	44.0	0.02	0.7	0.36	5.6	0.00
84.0	0.01	4.4	0.07	0.2*	0.43	0.2*	0.14
93.0	0.01	0.6	0.17	1.0	0.55	3.8	0.08
98.0	0.01	0.9	0.39	0.2	0.51	2.8	0.09
18.0	0.01	0.5	0.49	17.0	0.25	63.0	0,15
11.0	0.04	2.4	0.35	6.0	0.21	39.0	0.09
43.0	0.00	3.5	0.41	5.1	0.00	20.0	0.03

*These two observations were originally zero. They were Winsorized and replaced by the smallest observed value because of measurement difficulties with small values.

References

Aitchison, J. and Brown, J.A.C. (1966). The Log-Normal Distribution. Cambridge, Mass.: Cambridge University Press.

Box, G.E.P. and Tiao, G. C. (1965). Multiparameter problems from a Bayesian point of view. Annals of Mathematical Statistics 36, 1468-1482.

Bradu, D. and Mundlak, Y. (1970). Estimation in lognormal linear models. Journal of the American Statistical Association 65, 198-211.

Draper, N. R. and Smith, H. (1966). Applied Regression Analysis, Wiley, New York.

Finney, D. J. (1951). On the distribution of a variate whose logarithm is normally distributed. Supplement to the Journal of the Royal Statistical Society 7, 155-61.

Goldberger, A. S. (1968). The interpretation and estimation of the Cobb-Douglas functions. Econometrica 36, 464-72.

Heien, D. M. (1968). A note on the log-linear regression. Journal of the American Statistical Association 63, 1034-8.

Land, C. E. (1969). Confidence interval estimation of functions of the normal mean and variance. Invited paper at IMS meetings, New York, August 19-22, 1969. Abstract in: Annals of Mathematical Statistics 40, 1860.

Lindley, D. V. (1965). Introduction to Probability and Statistics. Cambridge University Press, London. Vol. 2, 203-214.

Paul, R. (1970). An Environmental Model for Swimming Pool Bacteriology. Masters Dissertation for the Department of Environmental Science and Engineering, The University of North Carolina at Chapel Hill.

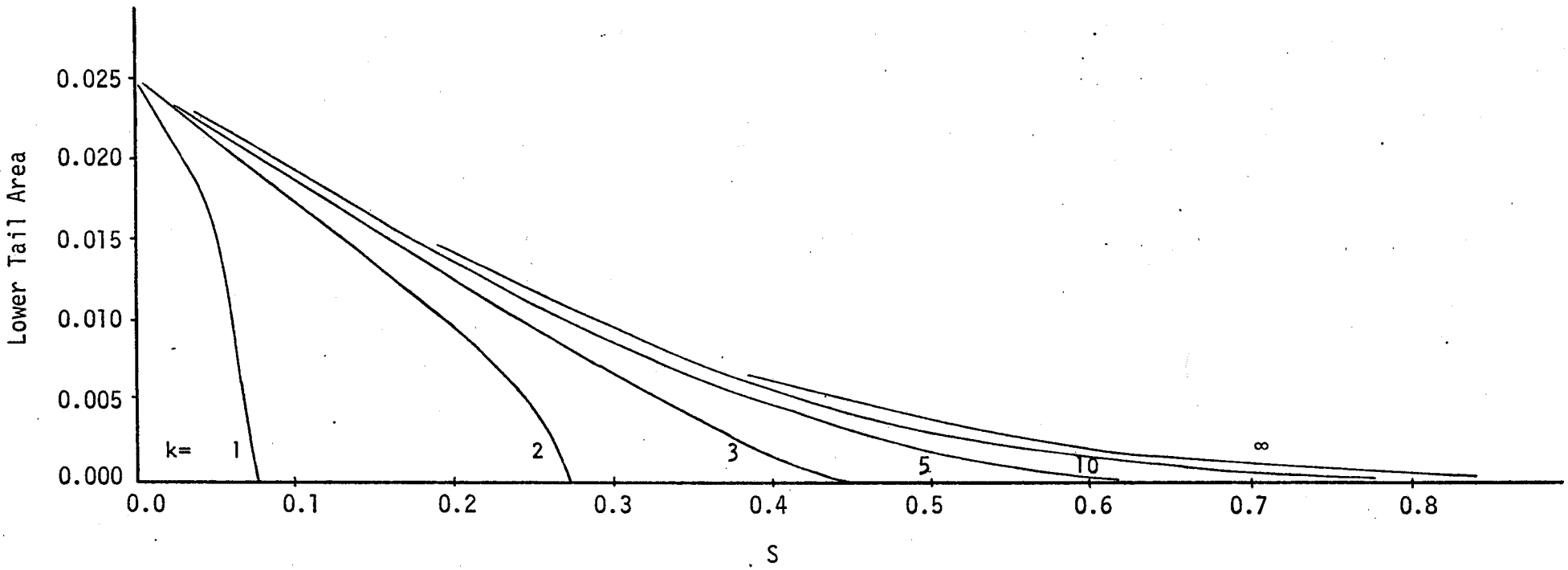

FIGURE 1 — Lower Tail Areas of Shortest Interval as a Function of S for Selected Degrees of Freedom, k

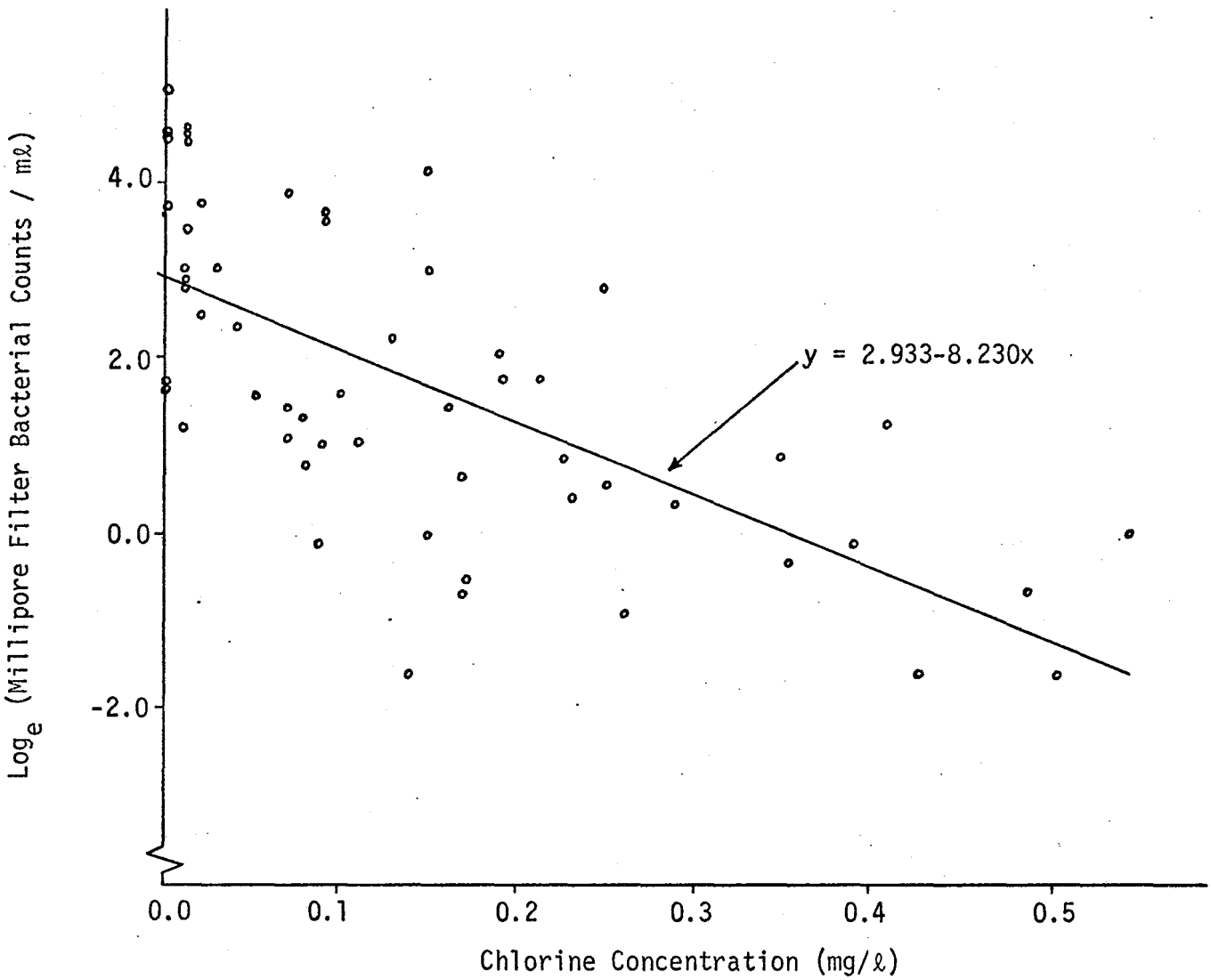

FIGURE 2 — Scatter Diagram and Estimated Regression for Paul's Data

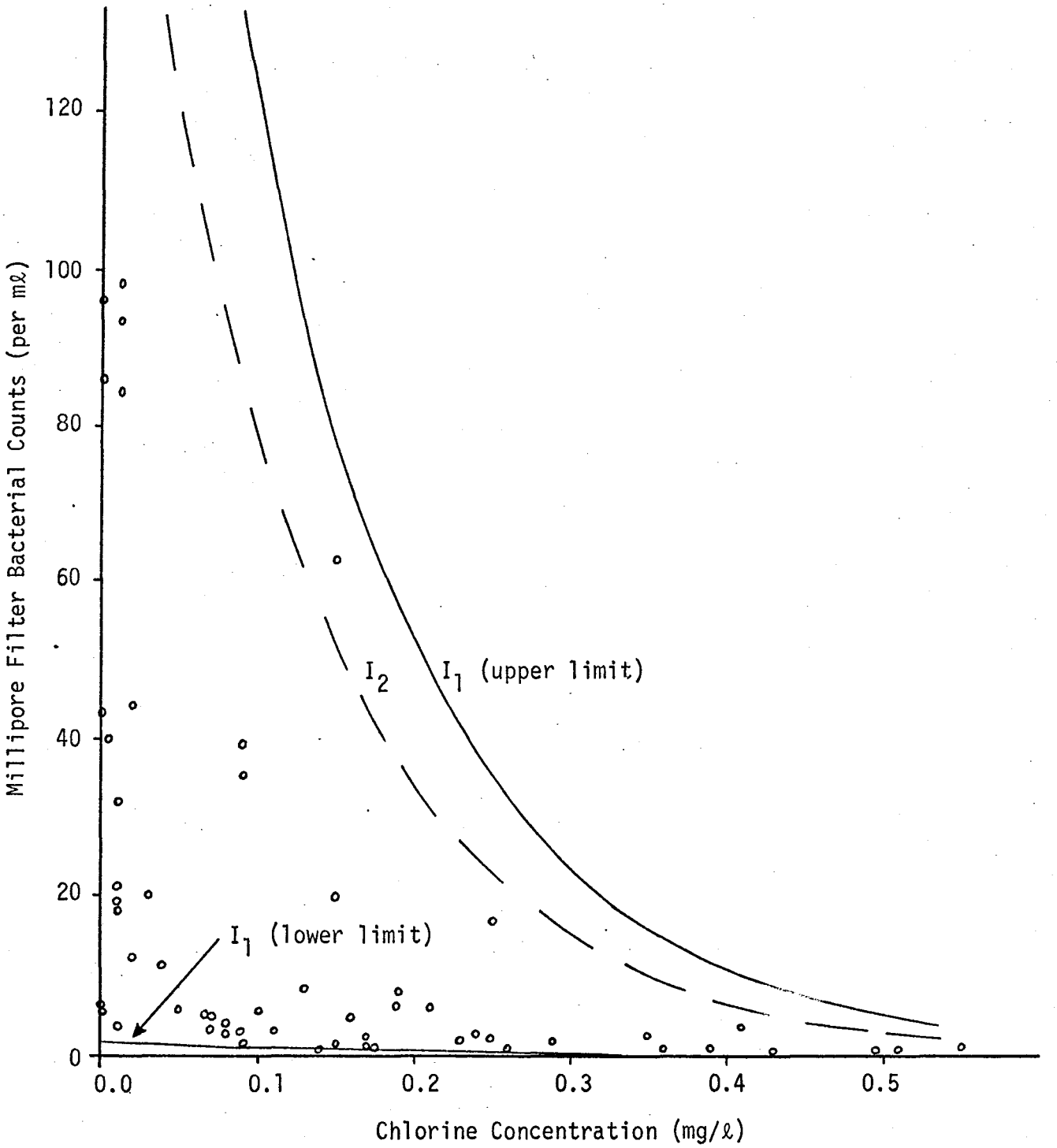

FIGURE 3 — Distribution of Paul's Data with 95% Prediction Intervals I_1 and I_2 Superimposed