The following paper with its discussion was the first of two invited papers presented in the TECHNOMETRICS Session of the 131st Annual Meeting of the American Statistical Association held at Fort Collins, Colorado, August 23-26, 1971.

Linear Estimation of a Regression Relationship from Censored Data Part I—Simple Methods and Their Application

WAYNE NELSON AND GERALD J. HAHN

General Electric Corporate Research and Development
Schenectady, New York

In many regression problems, data on the dependent variable are censored; that is, the values of some observations are known only to be above or else below some value. Such data often arise in accelerated life testing where life is the dependent variable and temperature or stress is the independent variable and some test units have not failed at the time of the analysis. In such situations, the standard techniques of least squares estimation for the parameters of a linear regression model cannot be used, since the values of the censored observations are not known.

This is Part I of a two-part series on the theory and application of linear estimation methods for regression analysis using the ordered observations of censored data. The use of these methods is illustrated with analyses of censored data from an accelerated life test of motor insulation and of censored data from tandem specimens in a creep-rupture test on an alloy.

KEY WORDS

Censored Data
Regression Analysis
Linear Estimation
Best Linear Unbiased Estimators
Order Statistics
Life Data Analysis
Accelerated Life Testing

INTRODUCTION

The problem. Often one desires to estimate the relationship between some measure of product performance and one or more stress, environmental or other independent variables. Typically, performance measurements are obtained at a number of stress conditions. The resulting data are used to estimate the relationship between performance and stress. The estimated relationship, which smoothes the data, is then used to estimate performance at one or more conditions.

It may happen that the performance data are censored; that is, some performance values at some stress conditions are known only to be above or else below some value. Such data often arise when the dependent (performance) variable

is time to failure. The time to failure of each unfailed unit is known only to exceed some value, which represents its survival time at the time of the analysis. Also, censored data occur when an electrical parameter is measured using an instrument with a limited scale. Readings are obtained for all values falling within the scale of measurement. All that is known for a value outside this range is whether it was below or else above the measurement scale.

This two-part article describes linear estimation methods for regression analysis when the data on the dependent variable are censored. In such situations, censored data on some performance variable are obtained at different values of one or more independent variables, and the relationship between performance and these variables is to be estimated.

Types of censoring. A number of different types of censored data will be briefly described for background. A more detailed discussion is provided by Nelson (1969).

Data are said to be singly censored if the values of the observations in one of the distribution tails are not known and are doubly censored if the values of the observations in both tails are unknown. Life test data are frequently singly censored to the right; that is, the failure times of unfailed units are known only to be beyond their current running times. This would be the case, for example, in a life test if all units are placed on test at the same time and all unfailed units have, as a result, accumulated the same running time at the time of analysis. Instrumentation data may be doubly censored; that is, observations may be beyond the scale of measurement at either tail of the distribution. Data are said to be multiply or progressively censored if the censored values and uncensored values are intermixed. Multiply censored life data are frequently encountered in the field where different units have differing running times at the time of an analysis of the data.

Censored data are said to have Type I censoring if censored observations occur only at specified values of the dependent variable. Such censoring results, for example, in life testing when all units are put on test at the same time and the data are collected and analyzed at a specified point in time. For life data, this is called time censoring. In this type of censoring, the censoring values are fixed and the number of censored observations is random. The insulation life data in the example in Section 2 involve this type of censoring. Censored data are said to have Type II censoring if the number of censored observations is specified and their censored values are random. Such censoring results, for example, in life testing when all units are put on test at the same time and the testing is terminated when a specified number have failed. For life data, this is called failure censoring. The data in the example in Section 3 on tandem creep-rupture specimens have Type II censoring.

The model. The model and assumptions, upon which the methods are based, are described briefly here and in greater detail in the Appendix in Part II. It is assumed that the dependent variable has a specified two parameter distribution with a location parameter μ and a scale parameter σ. The location parameter is given in terms of P independent variables $x_1, \cdots, x_P$ by the linear relationship

$$\mu(x_1, \cdots, x_P) = \beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P$$

where the values of the coefficients β_0 , β_1 , $\cdots$, β_P are unknown. The scale parameter σ is assumed to be a constant whose value is unknown.

In practice, the normal or the smallest extreme value distributions is often used as the distribution for the dependent variable. For the normal distribution, the location parameter μ is the mean (also the 50'th percentile) and the scale parameter σ is the standard deviation of the distribution. For the smallest extreme value distribution, the location parameter is the mode (also the 63.2'th percentile), and the scale parameter is a multiple of the standard deviation of the distribution. The lognormal or Weibull distribution can also be used as the distribution for the dependent variable, since the logarithm of a lognormal variate or of a Weibull variate has a normal or smallest extreme value distribution, respectively. Then, for the lognormal distribution, the logarithmic mean is the location parameter, and for the Weibull distribution, the natural logarithm of the scale parameter is the location parameter. In the case of the Weibull distribution, this leads to linear unbiased estimates of the natural logarithm of the Weibull scale parameter and of the reciprocal of the shape parameter. These special models are presented in more detail by Nelson and Hahn (1971).

It is assumed that independent random samples are taken at each of K different conditions, and, for the sample at the k'th condition ($k = 1, \cdots, K$), the values of the independent variables are x_{k1} , $\cdots$, x_{kP} . Then the value of the location parameter at the k'th condition is

$$\mu_k = \beta_0 + \beta_1 x_{k1} + \cdots + \beta_P x_{kP} .$$

Suppose that the size of the k'th sample is N_k and that R_k sample values are observed and the remaining $N_k - R_k$ observations are Type II censored. Let $y_{k1} \leq \cdots \leq y_{kR_k}$ denote the ordered observed values. These need not be the R_k smallest order statistics but may be any set of R_k observed values. For example, they may be the order statistics in a progressively censored sample with Type II censoring (see Nelson (1969), Thomas and Wilson (1970), and Mann (1970)). Although the underlying theory is based on Type II censoring, the methods will also be applied to Type I censored data. The consequences of this are not serious in most practical situations. This matter is discussed further by Hahn and Nelson (1971).

Related work. Work on linear estimation for regression analysis of censored data has been done by Lieblein and Zelen (1956) and is described in Section 2F. Alternative methods to the method of linear unbiased estimation presented here for regression analysis of censored data are maximum likelihood and graphical methods. Hahn and Nelson (1971) review and compare the three methods (graphical, maximum likelihood and linear unbiased estimation) and provide a guide to help readers decide on an appropriate method in a given application.

Outline of the contents of this article. This article is published in two parts. The first part, which appears here, contains Sections 1 and 2. The second part, which will appear in a later issue of *TECHNOMETRICS*, contains Sections 3, 4, 5 and the Appendix. The theory and methods for obtaining minimum variance ("best") linear unbiased estimators of the parameters of a linear regression model are developed in the Appendix. The best linear unbiased estimators are more

precise and are therefore preferred over other linear unbiased estimators. However, the method leads to a problem in generalized least squares estimation, that is, one where the observations have unequal variances and are correlated. This presents no new theoretical problems, but requires complicated hand calculations, if a computer program for generalized regression analysis is not available. Much of the body of this paper therefore concentrates on procedures which are computationally simpler and whose basis is also described in the Appendix.

Step-by-step procedures for a method for simple (but not minimum variance) linear unbiased estimation of the parameters of a linear regression model with censored data on the dependent variable are given in Section 2 for the special case of one independent variable. These procedures are illustrated with an analysis of censored data on insulation life described by the Arrhenius model. Section 3 provides step-by-step procedures for obtaining linear estimates of the regression model parameters for the special case where the sample size at each test condition is the same and only the first order statistic is observed (i.e., uncensored). In this case, the simple method for linear unbiased estimation does not apply, but the method for best linear unbiased estimation is easy to apply. The method for this case is illustrated by an analysis of censored creep-rupture data on tandem specimens of an alloy, tested at differing stress levels. Limitations of linear estimation methods are briefly indicated in Section 4, and concluding remarks on linear estimation are made in Section 5.

A company report (Nelson and Hahn (1971)) on which the present paper is based, includes three additional appendices. The first of these appendices reviews available tabulations of factors required for obtaining the various estimators. The second appendix contains some short tabulations of such factors. The third appendix briefly presents four commonly used statistical distributions and their corresponding linear regression relationships.

Methods for Simple Linear Unbiased Estimation of the Parameters of a Linear Regression Model

A. *Introduction*

This section contains a step-by-step presentation of simple methods for calculating linear unbiased estimates for the parameters of a linear regression model from censored data. It includes methods for obtaining approximate confidence intervals for the parameters as well as for estimating other quantities of interest in practical problems. The emphasis in this section is on the application of the methods, which are illustrated by a numerical example. The unbiased estimators presented here are linear functions of the ordered observations but are not the best ones (i.e., do not have minimum variance).

The theory for the simple methods for linear unbiased estimation is given in the Appendix. The discussion in the Appendix is more general than that provided here, since the presentation here is limited to certain simple, important applications of the theory.

The simple method involves obtaining the best linear unbiased estimates of

the location and scale parameters of the distribution at *each test condition,* using existing tabulations for such estimates, and then using these estimates to fit to the data the regression relationship between the independent variable and the location parameter. This method is described step by step in subsection C. The fitting involves a regression analysis based upon uncorrelated observations, since the results at the different test conditions are statistically independent of one another. However, a weighted regression analysis is required, since the variances of the estimates of the location parameters at the different test conditions will in general vary due to the differing sample sizes and differing amounts of censoring at each test condition. A weighted regression analysis on uncorrelated observations, however, is reasonably simple to perform either directly or by transforming the problem to one of a simple regression (fitted through the origin) as described in the Appendix.

A still simpler method for obtaining linear unbiased estimators is to conduct a standard (i.e., unweighted) regression analysis instead of performing a weighted regression analysis on the estimates of the location parameters at the test conditions. The simpler method is presented step by step in subsection D. This method ignores the fact that the variances of the estimates at the different test conditions vary. As a result, the computations can be carried out by essentially any available computer program for standard least squares regression analysis. This method also leads to linear unbiased estimates, but with higher variances than those of the previous simple estimates, which take into account the differences in variances by a weighted regression analysis. In the special case where the sample size and the censoring scheme are identical at each test condition, the variances of the estimates at each test condition are the same and consequently the weighted regression analysis simplifies to an unweighted regression analysis. In situations where the sample sizes and the censoring schemes are similar, but not identical, at the various test conditions, the loss in efficiency in using the unweighted as compared to the weighted regression analysis is not great. The unweighted regression analysis may be preferred due to its computational ease. This computational ease applies to obtaining point estimates only and does not carry over to calculating confidence intervals or conducting hypothesis tests.

The preceding methods for unbiased linear estimation are similar to the graphical method which is presented by Hahn and Nelson (1970) and which provided the authors with the motivation for the simple methods.

B. *Description of an Insulation Life Problem*

A problem concerning the analysis of censored data from an accelerated life test of Class B insulation in motorettes will be used to illustrate the methods of analysis presented in Part I. This problem will now be described.

In order to evaluate a new Class B insulation for electric motors, temperature accelerated life testing was conducted on 40 motorettes. Ten motorettes were put on test together at each of four temperatures (150°C, 170°C, 190°C, 220°C). The main purpose of the experiment was to obtain information about the distribution of insulation life (in particular, its median and 10% point) for the design temperature of 130°C. At the time the analysis was performed, seven

motorettes at 170°C had failed, five each had failed at 190°C and 220°C, and none had failed at 150°C. Such motorettes are inspected periodically for failure, and each recorded failure time is the midpoint of the period in which the failure occurred. The data are shown in Table 2.1. Crawford (1970) published these data and, assisted by the authors, originally used graphical and maximum likelihood methods to analyse the data.

For many products undergoing temperature accelerated life testing, the Arrhenius model has been found satisfactory for estimating life at design temperatures. This model will be used to analyze the Class B insulation data. The assumptions of the model are

i) for any temperature, the life distribution is lognormal,
ii) the standard deviation σ of the logarithmic life is a constant (i.e., independent of temperature), and
iii) the mean $\mu(x)$ of the logarithmic life is a linear function of the reciprocal $x = 1/T$ of the absolute temperature T, that is,

$$\mu(x) = \beta_0 + \beta_1 x \tag{2.1}$$

TABLE 2.1
Insulation Life Data at Various Test Temperatures

150°C	All 10 motorettes still on test without failure at 8064 hours.		
170°C	Hours to Failure	Rank i	Plotting Position $100\,(i - 0.5)/n$
	1764	1	5
	2772	2	15
	3444	3	25
	3542	4	35
	3780	5	45
	4860	6	55
	5196	7	65
	3 motorettes still on test without failure at 5448 hours.		
190°C	Hours to Failure	Rank i	Plotting Position $100\,(i - 0.5)/n$
	408	1	5
	408	2	15
	1344	3	25
	1344	4	35
	1440	5	45
	5 motorettes still on test without failure at 1680 hours.		
220°C	Hours to Failure	Rank i	Plotting Position $100\,(i - 0.5)/n$
	408	1	5
	408	2	15
	504	3	25
	504	4	35
	504	5	45
	5 motorettes still on test without failure at 528 hours.		

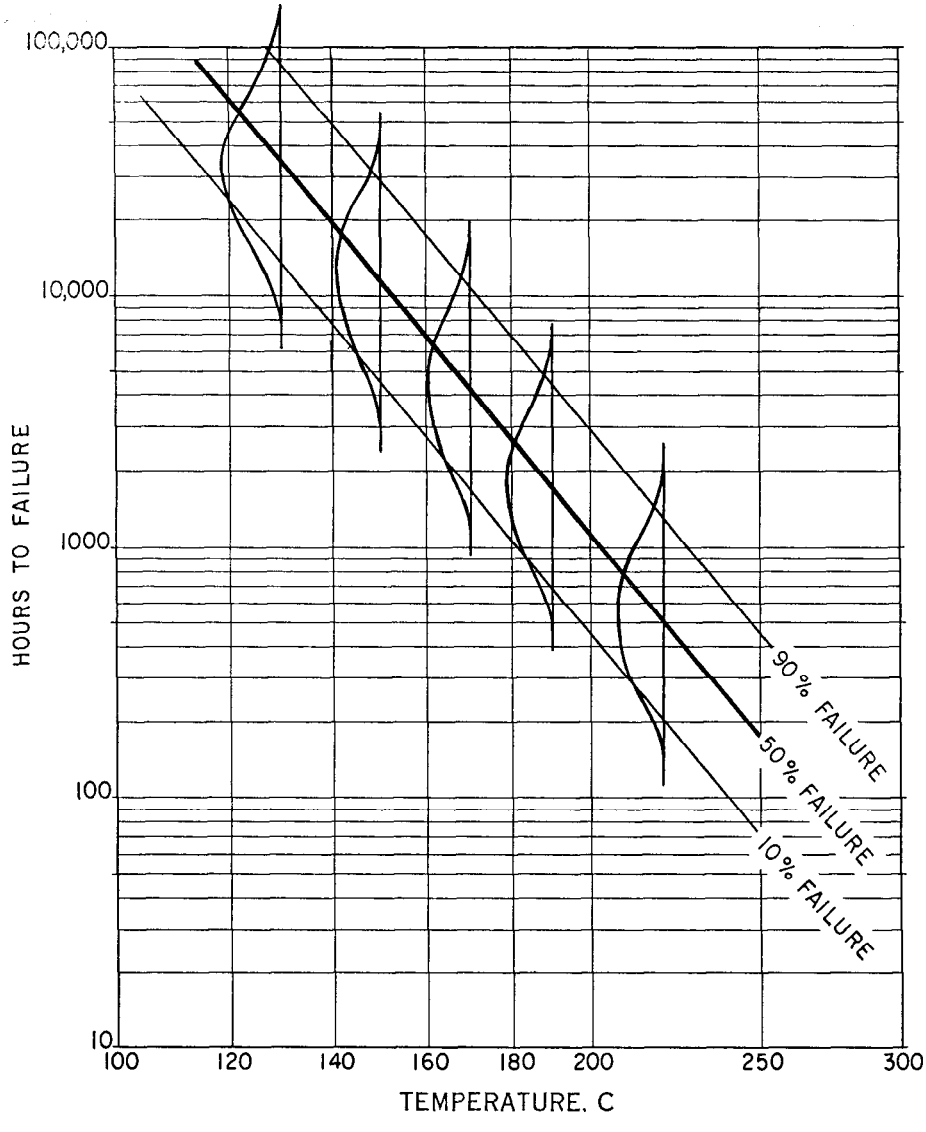

FIGURE 2.1—The Arrhenius Model on Arrhenius Paper

where β_0 and β_1 are parameters characteristic of the product and the test method. (It is convenient to work with $x = 1000/T$.)

The antilogarithm of the mean logarithmic life $\mu(x)$ is the median life and is regarded as a nominal life. Equation (2.1) is called the Arrhenius relationship. An example of such a relationship is depicted in Figure 2.1 on Arrhenius plotting paper. Such paper has a horizontal scale for reciprocal absolute temperature and a vertical logarithmic scale for time. The model is also depicted in Figure 2.2 on lognormal probability paper. Methods for analyzing complete data with this model are given by Nelson (1970a).

The grouped nature of the data will not be taken into account in the analysis. That is, it will be assumed that the times midway between inspections in Table 2.1 are the actual times to failure. The actual failure could have occurred any time between the time it was detected and the time of the preceding inspection. The times between inspections are short, and the effect of the grouping on the results is therefore small.

Linear methods for estimating the life distribution at the design temperature from censored data are given in the next subsection. These methods can also be

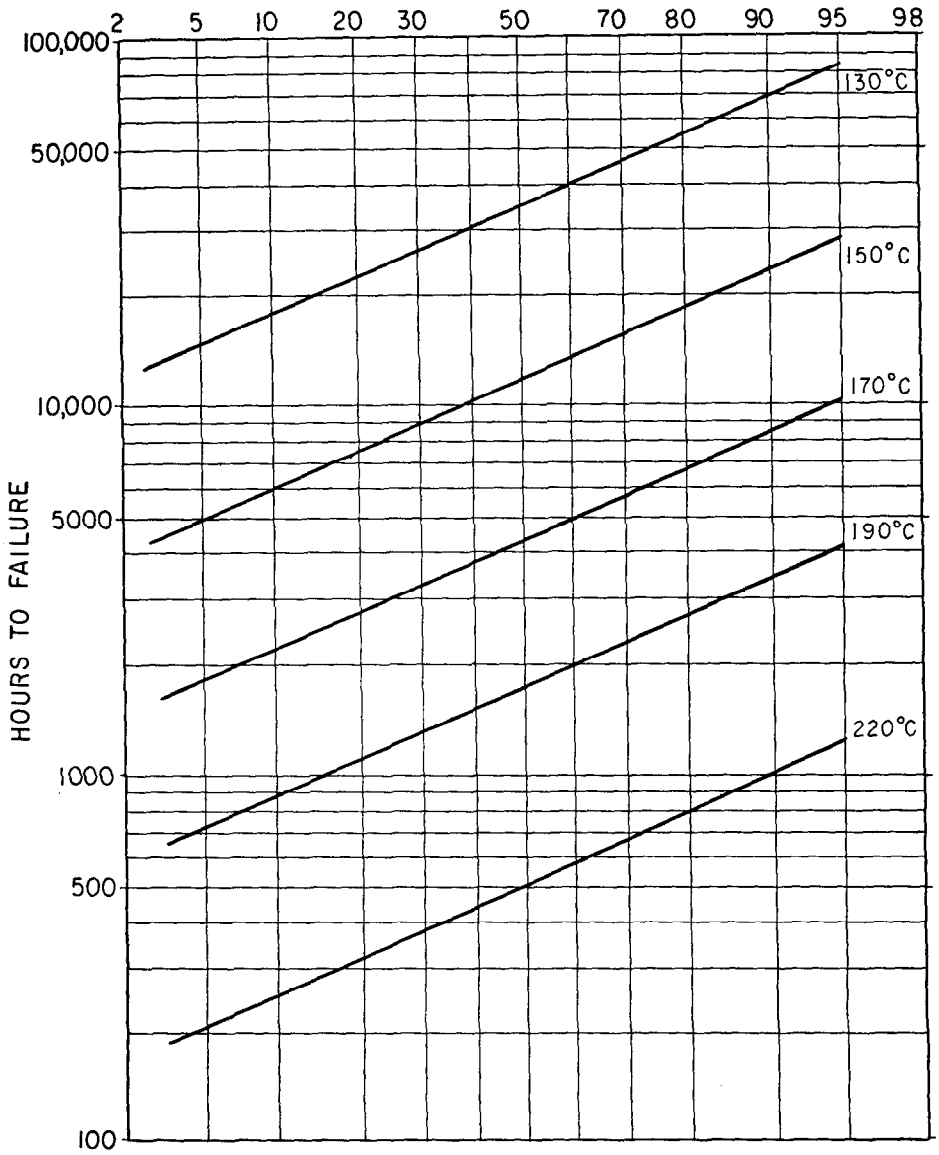

FIGURE 2.2—The Arrhenius Model on Lognormal Probability Paper

used for other distributions and regression relationships. Corresponding linear methods for complete data are presented by Nelson (1970b) for the inverse power law model, which involves the Weibull distribution and a simple linear regression relationship.

C. *Simple Linear Unbiased Estimators for Regression with One Independent Variable and Censoring to the Right (Using Weighted Regression Analysis)*

Given next are the specific expressions for the simple linear unbiased estimators of the parameters in the linear regression model with one independent variable and data censored to the right. They are applied to the insulation life problem just presented. This method requires that there be at least two observed (i.e., uncensored) values at each of the test conditions which are to be included in the analysis. This requirement is met at the three test temperatures (test conditions) of 170°C, 190°C and 220°C in the insulation life example, but not at 150°C. Therefore that information cannot be used in the following analysis. A step-by-step presentation of the method is given below.

Calculational Steps

Notation

Let k denote a typical test condition for which at least the 2 smallest sample values are observed, $k = 1, \cdots, K$.

Let N_k denote the total sample size at the k'th condition, $k = 1, \cdots, K$.

Let R_k denote the number of (uncensored) observations at the k'th condition, $k = 1, \cdots, K$, (assumed here to be the R_k smallest values, where $R_k \geq 2$).

Let x_k denote the value of the independent variable x at the k'th condition, $k = 1, \cdots, K$.

Let $y_{k1}, \cdots, y_{kR_k}$ denote the values of the dependent variable y for the R_k observed values at the k'th condition, arranged in order of magnitude, i.e., y_{k1} is the smallest observation, y_{k2} is the second smallest observation, etc., where $k = 1, \cdots, K$.

Insulation Life Example (see Section 2B)

There are 3 temperatures at which 2 or more failures were obtained. Thus, there are $K = 3$ test conditions.

The sample sizes are $N_1 = N_2 = N_3 = 10$.

The numbers of observed (uncensored) values are $R_1 = 7$, $R_2 = 5$, $R_3 = 5$.

The values of the independent variable are

$$x_1 = 1000/(170 + 273.2) = 2.256 \quad \text{for} \quad 170°\text{C},$$

$$x_2 = 1000/(190 + 273.2) = 2.159 \quad \text{for} \quad 190°\text{C},$$

$$x_3 = 1000/(220 + 273.2) = 2.028 \quad \text{for} \quad 220°\text{C}.$$

Since the assumed distribution is lognormal, we use the logarithms of the times

to failure:

$$y_{11} = 3.2465, \cdots, y_{17} = 3.7157,$$
$$y_{21} = 2.6107, \cdots, y_{25} = 3.1584,$$
$$y_{31} = 2.6107, \cdots, y_{35} = 2.7024.$$

Times are tabulated in Table 2.1 and are plotted on lognormal probability paper in Figure 2.3. Plotting positions are given in Table 2.1.

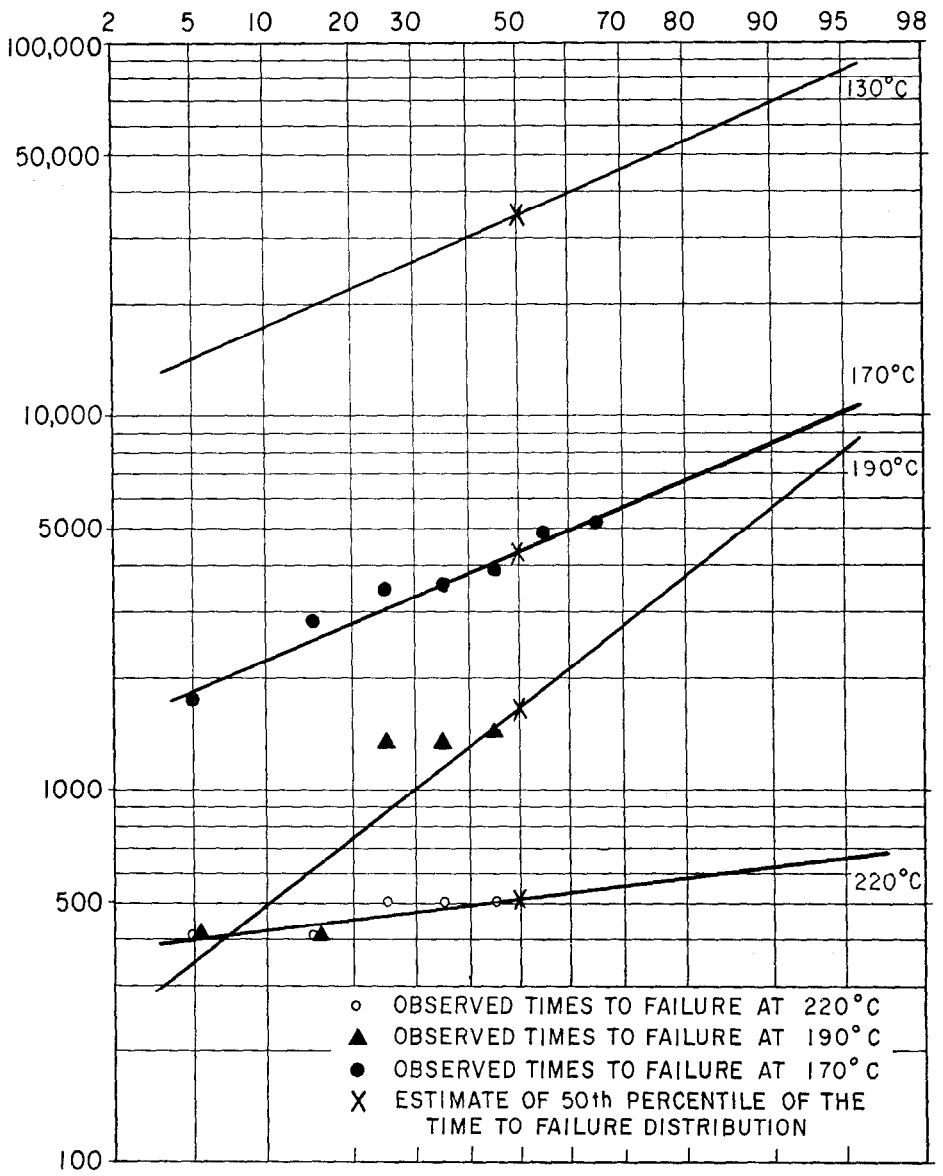

FIGURE 2.3—Lognormal Plots of the Insulation Life Data at the Different Test Temperatures

The Simple Estimation Procedure

Step 1. Obtain the tabulated coefficients

$$a_1(R_k , N_k), \cdots , a_{R_k}(R_k , N_k),$$
$$b_1(R_k , N_k), \cdots , b_{R_k}(R_k , N_k), \qquad k = 1, \cdots , K,$$

for the best linear unbiased estimates of the location and scale parameters μ and σ, respectively. These coefficients depend on the number R_k of observed values, the size N_k of the sample at the test condition and the assumed distribution for the random variation. Tabulations of such coefficients are given by Sarhan and Greenberg (1962).

Example. Since the random variation in the logarithmic y_{kr}'s is assumed to be normally distributed, the required following coefficients are obtained from Table 10C.1 of Sarhan and Greenberg (1962). For $k = 1$, $R_1 = 7$, $N_1 = 10$:

i	$a_i(7, 10)$	$b_i(7, 10)$
1	.0244	−.3252
2	.0636	−.1758
3	.0818	−.1058
4	.0962	−.0502
5	.1089	−.0006
6	.1207	.0469
7	.5045	.6107

For $k = 2, 3$, $R_2 = R_3 = 5$, $N_2 = N_3 = 10$;

i	$a_i(5, 10)$	$b_i(5, 10)$
1	−.1240	−.4919
2	−.0016	−.2491
3	.0549	−.1362
4	.0990	−.0472
5	.9718	.9243

Step 2. Calculate the best linear unbiased estimates of the location and scale parameters at each test condition as

$$\mu_k^* = \sum_{i=1}^{R_k} a_i(R_k , N_k)y_{ki} , \qquad \sigma_k^* = \sum_{i=1}^{R_k} b_i(R_k , N_k)y_{ki} , \quad \text{for} \quad k = 1, \cdots , K,$$

Example. The estimates are

$$\mu_1^* = (0.0244)(3.2465) + \cdots + (0.5045)(3.7157) = 3.6381,$$
$$\mu_2^* = (-0.1240)(2.6107) + \cdots + (0.9718)(3.1584) = 3.2233,$$
$$\mu_3^* = (-0.1240)(2.6107) + \cdots + (0.9718)(2.7024) = 2.7142,$$

$$\sigma_1^* = (-0.3252)(3.2465) + \cdots + (0.6107)(3.7157) = 0.2265,$$

$$\sigma_2^* = (-0.4919)(2.6107) + \cdots + (0.9243)(3.1584) = 0.4110,$$

$$\sigma_3^* = (-0.4919)(2.6107) + \cdots + (0.9243)(2.7024) = 0.0677.$$

The fitted distribution lines on the lognormal plots of these data in Figure 2.3 were drawn using these parameter estimates rather than by a visual fit. The antilogarithm of the estimate of the logarithmic mean at a temperature is the estimate of the median life at that temperature. For example, the estimate of the median life at 170°C is antilog (3.6381) = 4350 hours. This and the other estimates of the median lives for the test temperatures are shown as crosses in Figures 2.3 and 2.4. The antilogarithm of the sum of the estimates of the logarithmic mean and standard deviation at a temperature is the estimate of the 84th percentile at that temperature. For example, the estimate of the 84th percentile at 170°C is antilog (3.6381 + 0.2265) = 7320 hours. In Figure 2.3, the fitted distribution line at each test temperature is drawn through the corresponding estimate of the median and the 84th percentile.

Step 3. For each test condition, obtain from the tabulations the standardized variances

$$V(\mu_k^*) = V_{\mu k}(R_k, N_k), \qquad V(\sigma_k^*) = V_{\sigma k}(R_k, N_k), \qquad k = 1, \cdots, K,$$

of the best linear unbiased estimators of the location and scale parameters. These depend on R_k, N_k, and the assumed distribution of the random variation. Selected tabulations of these standardized variances are given by Sarhan and Greenberg (1962).

Example. Using Table 10C.2 of Sarhan and Greenberg (1962) for $R_1 = 7$ and $N_1 = 10$, one obtains for $k = 1$

$$V(\mu_1^*) = 0.1167 \text{ and } V(\sigma_1^*) = 0.0989.$$

For $k = 2, 3$, $R_2 = R_3 = 7$, $N_2 = N_3 = 10$, one obtains

$$V(\mu_2^*) = V(\mu_3^*) = 0.1664$$

and

$$V(\sigma_2^*) = V(\sigma_3^*) = 0.1613.$$

Step 4. Obtain the pooled estimate σ^* of the scale parameter σ as

$$\sigma^* = V(\sigma^*) \sum_{k=1}^{K} (\sigma_k^*/V(\sigma_k^*))$$

where

$$V(\sigma^*) = \left[\sum_{k=1}^{K} (1/V(\sigma_k^*))\right]^{-1}.$$

The estimated standard error of σ^* is

$$s(\sigma^*) = \sqrt{V(\sigma^*)}\sigma^*.$$

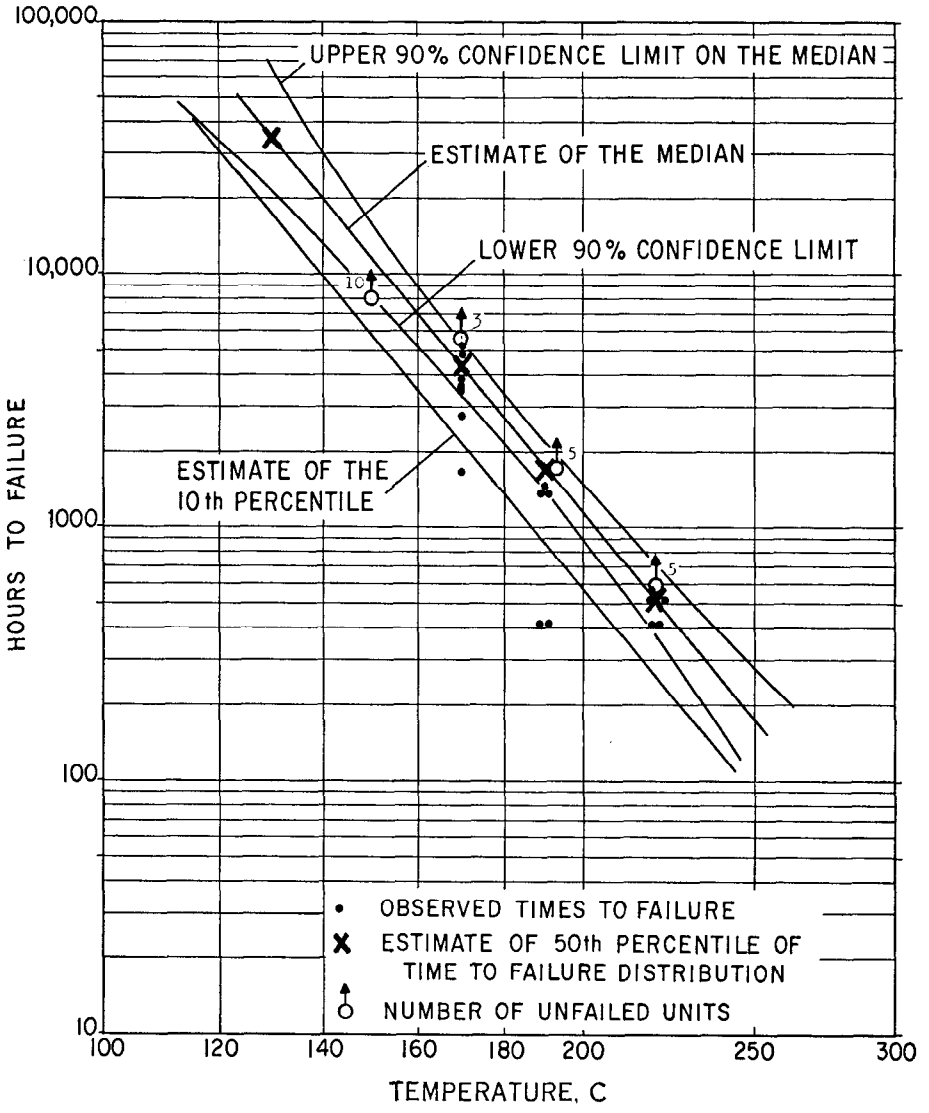

FIGURE 2.4—Arrhenius Plot of the Insulation Life Data

Example. The calculations for the estimate of σ and its estimated standard error are

$$V(\sigma^*) = [(1/0.0989) + (1/0.1613) + (1/0.1613)]^{-1} = 0.04442,$$

$$\sigma^* = 0.04442[(0.2265/0.0989) + (0.4110/0.1613) + (0.0677/0.1613)] = 0.2336,$$

$$s(\sigma^*) = \sqrt{V(\sigma^*)}\sigma^* = \sqrt{0.04442}\,(0.2336) = 0.0492.$$

Step 5. The linear unbiased estimators of the regression coefficients β_0 and β_1 are

$$\beta_1^* = \left\{\left[\sum_{k=1}^{K}\frac{\mu_k^* x_k}{V(\mu_k^*)}\right]\left[\sum_{k=1}^{K}\frac{1}{V(\mu_k^*)}\right] - \left[\sum_{k=1}^{K}\frac{\mu_k^*}{V(\mu_k^*)}\right]\left[\sum_{k=1}^{K}\frac{x_k}{V(\mu_k^*)}\right]\right\} \Bigg/ \left\{\left[\sum_{k=1}^{K}\frac{1}{V(\mu_k^*)}\right]\left[\sum_{k=1}^{K}\frac{x_k^2}{V(\mu_k^*)}\right] - \left[\sum_{k=1}^{K}\frac{x_k}{V(\mu_k^*)}\right]^2\right\},$$

$$\beta_0^* = \left\{\sum_{k=1}^{K}(\mu_k^*/V(\mu_k^*)) - \beta_1^*\sum_{k=1}^{K}(x_k/V(\mu_k^*))\right\} \Bigg/ \sum_{k=1}^{K}[1/V(\mu_k^*)].$$

Note that a large number of significant figures, say, six or more, should be carried in the intermediate results in the evaluation of these formulas, to ensure satisfactory accuracy of the final results to three or four figures. Alternate equivalent expressions are given in Step 7.

Example. For the insulation life data

$$\begin{aligned}\beta_1^* = \Bigg\{&\left[\frac{3.6381(2.256)}{0.1167} + \frac{3.2233(2.159)}{0.1664} + \frac{2.7142(2.028)}{0.1664}\right] \\ &\cdot\left[\frac{1}{0.1167} + \frac{1}{0.1664} + \frac{1}{0.1664}\right] - \left[\frac{3.6381}{0.1167} + \frac{3.2233}{0.1664} + \frac{2.7142}{0.1664}\right] \\ &\cdot\left[\frac{2.256}{0.1167} + \frac{2.159}{0.1664} + \frac{2.028}{0.1664}\right]\Bigg\} \Bigg/ \Bigg\{\left[\frac{1}{0.1167} + \frac{1}{0.1664} + \frac{1}{0.1664}\right] \\ &\cdot\left[\frac{(2.256)^2}{0.1167} + \frac{(2.159)^2}{0.1664} + \frac{(2.028)^2}{0.1664}\right] - \left[\frac{2.256}{0.1167} + \frac{2.159}{0.1664} + \frac{2.028}{0.1664}\right]^2\Bigg\} \\ = 4.05365.\end{aligned}$$

$$\beta_0^* = \left\{\left[\frac{3.6381}{0.1167} + \frac{3.2233}{0.1664} + \frac{2.7142}{0.1664}\right] - 4.05365\left[\frac{2.256}{0.1167} + \frac{2.159}{0.1664} + \frac{2.028}{0.1664}\right]\right\} \Bigg/ \left\{\frac{1}{0.1167} + \frac{1}{0.1664} + \frac{1}{0.1664}\right\} = -5.513.$$

The heavy line shown on Arrhenius paper in Figure 2.4 has these coefficient estimates, which may also be calculated by the equivalent matrix method in Steps 6 and 7.

Step 6. By using simple matrix algebra methods one can obtain the preceding estimates of the regression coefficients and, in addition, their standard errors. The latter may be used to obtain approximate confidence intervals for the coefficients.

Develop the following matrices:

$$V = \begin{bmatrix} V(\mu_1^*) & 0 & \cdots & 0 \\ 0 & V(\mu_2^*) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V(\mu_K^*) \end{bmatrix},$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_K \end{bmatrix}, \qquad \mu^* = \begin{bmatrix} \mu_1^* \\ \mu_2^* \\ \vdots \\ \mu_K^* \end{bmatrix}.$$

Example. The necessary matrices are

$$V = \begin{bmatrix} 0.1167 & 0 & 0 \\ 0 & 0.1664 & 0 \\ 0 & 0 & 0.1664 \end{bmatrix},$$

$$X = \begin{bmatrix} 1 & 2.256 \\ 1 & 2.159 \\ 1 & 2.028 \end{bmatrix}, \qquad \mu^* = \begin{bmatrix} 3.6381 \\ 3.2233 \\ 2.7142 \end{bmatrix}.$$

Using matrix algebra, carry out the following steps:

Step 7. a. Obtain the estimates of the regression coefficients:

$$\beta^* = \begin{bmatrix} \beta_0^* \\ \beta_1^* \end{bmatrix} = (X'V^{-1}X)^{-1}(X'V^{-1}\mu^*).$$

Six figure accuracy is shown in the intermediate results in the example below, since six or more figures were carried throughout to ensure four figure accuracy in the final results.

Example.

a.

$$X'V^{-1}X = \begin{bmatrix} 1 & 1 & 1 \\ 2.256 & 2.159 & 2.028 \end{bmatrix} \begin{bmatrix} 1/0.1167 & 0 & 0 \\ 0 & 1/0.1664 & 0 \\ 0 & 0 & 1/0.1664 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2.256 \\ 1 & 2.159 \\ 1 & 2.028 \end{bmatrix} = \begin{bmatrix} 20.5882 & 44.4939 \\ 44.4939 & 96.3409 \end{bmatrix}.$$

$$[X'V^{-1}X]^{-1} = \frac{1}{(20.8582)(96.3409) - (44.4939)(44.4939)} \cdot \begin{bmatrix} 96.3409 & -44.4939 \\ -44.4939 & 20.5882 \end{bmatrix} = \begin{bmatrix} 25.4780 & -11.7667 \\ -11.7667 & 5.44468 \end{bmatrix}.$$

$$X'V^{-1}\mu^* = \begin{bmatrix} 1 & 1 & 1 \\ 2.256 & 2.159 & 2.028 \end{bmatrix} \begin{bmatrix} 1/0.1167 & 0 & 0 \\ 0 & 1/0.1664 & 0 \\ 0 & 0 & 1/0.1664 \end{bmatrix} \cdot \begin{bmatrix} 3.6381 \\ 3.2233 \\ 2.7142 \end{bmatrix} = \begin{bmatrix} 66.8569 \\ 145.2312 \end{bmatrix}.$$

$$\beta^* = (X'V^{-1}X)^{-1}(X'V^{-1}\mu^*)$$
$$= \begin{bmatrix} 25.4780 & -11.7667 \\ -11.7667 & 5.44468 \end{bmatrix}\begin{bmatrix} 66.8569 \\ 145.2312 \end{bmatrix} = \begin{bmatrix} -5.513 \\ 4.054 \end{bmatrix},$$

i.e., $\beta_0^* = -5.513$ and $\beta_1^* = 4.054$ (as before).

Step 7. b. Obtain the matrix of the estimated variances and covariances of the estimates of the regression coefficients

$$\Sigma_{\beta^*}^* = \begin{bmatrix} s^2(\beta_0^*) & \operatorname{cov}(\beta_0^*, \beta_1^*) \\ \operatorname{cov}(\beta_0^*, \beta_1^*) & s^2(\beta_1^*) \end{bmatrix} = (\sigma^*)^2(X'V^{-1}X)^{-1}.$$

Example.

$$\Sigma_{\beta^*}^* = (\sigma^*)^2(X'V^{-1}X)^{-1} = (0.2336)^2\begin{bmatrix} 25.4780 & -11.7667 \\ -11.7667 & 5.44468 \end{bmatrix}.$$

Thus,

$$s(\beta_0^*) = [(0.2336)^2(25.4780)]^{\frac{1}{2}} = 1.179,$$
$$s(\beta_1^*) = [(0.2336)^2(5.44468)]^{\frac{1}{2}} = 0.545.$$

Step 7. c. Obtain two-sided approximate $100\gamma\%$ confidence intervals for β_0 and β_1 as

$$\beta_0^* \pm z[(1+\gamma)/2]s(\beta_0^*)$$

and

$$\beta_1^* \pm z[(1+\gamma)/2]s(\beta_1^*),$$

respectively, where $z[(1+\gamma)/2]$ is the $100(1+\gamma)/2$'th percentile of the standard normal distribution.

Example.

An approximate 90% confidence interval for β_0 is

$$-5.513 \pm 1.645(1.179) = -5.513 \pm 1.940.$$

An approximate 90% confidence interval for β_1 is

$$4.054 \pm 1.645(0.545) = 4.054 \pm 0.897.$$

Estimation and Prediction at a Specified x_0

Obtain the estimate and confidence interval for the location parameter and a prediction interval for a future observation corresponding to the value x_0 of the independent variable as follows.

Step 8. a. The linear estimator for the location parameter of the distribution of the dependent variable at x_0 is

$$\mu^*(x_0) = \beta_0^* + \beta_1^* x_0 .$$

This estimator is unbiased.

Example.

a. The estimate of the logarithmic mean life (i.e., location parameter) at the design temperature of 130°C ($x_0 = 1000/(130 + 273.2) = 2.480$) is

$$\mu^*(2.480) = -5.513 + 4.054(2.480) = 4.540.$$

The antilogarithm of this value is 34,700 hours and is the estimate of the median life at 130°C.

Step 8. b. The estimated standard error of $\mu^*(x_0)$ is

$$s[\mu^*(x_0)] = (\sqrt{\text{Var}\,[\mu^*(x_0)]},$$

where

$$\text{Var}\,[\mu^*(x_0)] = [1\ x_0]\Sigma_{\beta^*}^{*}\begin{bmatrix}1\\ x_0\end{bmatrix}.$$

Example.

The estimated variance of $\mu^*(2.480)$ is

$$\text{Var}\,(\mu^*(2.480)) = (0.2336)^2[1 \quad 2.480]\begin{bmatrix}25.4780 & -11.7667\\ -11.7667 & 5.44468\end{bmatrix}\begin{bmatrix}1\\ 2.480\end{bmatrix}$$

$$= 0.03285.$$

The estimate of the standard error of $\mu^*(2.480)$ is thus

$$s(\mu^*(2.480)) = (0.03285)^{\frac{1}{2}} = 0.181.$$

Step 8. c. A two-sided approximate $100\gamma\%$ confidence interval for the true value of the location parameter of the distribution of the dependent variable at x_0 is

$$\mu^*(x_0) \pm z[(1 + \gamma)/2]s[\mu^*(x_0)]$$

where $z[(1 + \gamma)/2]$ is the $100(1 + \gamma)/2$'th percentile of the standard normal distribution.

Example.

An approximate 90% confidence interval for $\mu(2.480)$ is

$$\mu^*(2.480) \pm 1.645s(\mu^*(2.480)) = 4.540 \pm 1.645(0.181) = 4.540 \pm 0.298.$$

The antilogarithms of these limits are 17,500 and 68,900 hours and are the corresponding approximate limits for the median life at 130°C and differ from the estimate, 34,700 hours, by a factor of about 2.

Corresponding estimates and approximate 90% confidence intervals for the logarithmic means are

$$3.632 \pm 0.120 \quad \text{for} \quad 170°\text{C},$$

$$3.239 \pm 0.080 \quad \text{for} \quad 190°\text{C},$$

$$2.708 \pm 0.146 \quad \text{for} \quad 220°\text{C},$$

and the antilogarithms of these quantities give the corresponding estimates and confidence limits for the median lives at those temperatures, namely,

4,280 hours (3,250 to 5,650 hours)

1,730 hours (1,440 to 2,080 hours)

510 hours (365 to 715 hours)

Curves have been drawn through these limits in Figure 2.4.

Step 8. d. If the underlying distribution of the dependent variable is normal or lognormal, an approximate $100\gamma\%$ prediction interval on a single future observation of the dependent variable at x_0 is

$$\mu^*(x_0) \pm z[(1 + \gamma)/2]\{\text{Var } [\mu^*(x_0)] + (\sigma^*)^2\}^{\frac{1}{2}}.$$

where $z[(1 + \gamma)/2]$ is the $100(1 + \gamma)/2$'th percentile of the standard normal distribution.

Example.

An approximate 90% prediction interval for a single future observation at 130°C ($x_0 = 2.480$) is

$$\mu^*(2.480) \pm 1.645[\text{Var } (\mu^*(2.480)) + (\sigma^*)^2]^{\frac{1}{2}}$$

$$= 4.540 \pm 1.645[0.03285 + (0.2336)^2]^{\frac{1}{2}} = 4.540 \pm 0.485.$$

The antilogarithms of these limits are 11,400 and 106,000 hours and are the corresponding approximate limits of an interval to contain the life of a single future unit at 130°C. Similar prediction limits could be calculated at the other temperatures.

Step 8. e. A linear estimate of the $100P$'th percentile of the distribution of the dependent variable at x_0 is

$$y^*(P, x_0) = \mu^*(x_0) + z(P)\sigma^*$$

where $z(P)$ is the $100P$'th percentile of the standardized form of the assumed distribution, that is, the assumed distribution with the location parameter equal to zero and the scale parameter equal to one. This estimator is unbiased.

Example.

An estimate of the 10% point of the distribution of logarithmic life at the design temperature of 130°C ($x_0 = 2.480$) is

$$y^*(0.10, 2.480) = 4.540 + (-1.2816)0.2336 = 4.241$$

where $z(0.10) = -1.2816$ is the 10% point of the standard normal distribution. The antilogarithm of this is 17,400 hours and is the estimate of the 10% point of the insulation life distribution at the design temperature. The dependence of the 10th percentile on temperature is shown as a straight line in Figure 2.4.

D. *Simpler Linear Unbiased Estimators (Using Unweighted Regression Analysis)*

The procedure in the previous section employed weighted regression analysis to obtain linear unbiased estimators for the coefficients of a linear regression model. Given in this section is a simpler procedure for obtaining linear unbiased estimators using a standard (unweighted) regression analysis. The first four steps are the same as those for the previous method. The discussion is limited to estimation procedures. The construction of confidence intervals and hypothesis tests is not simple computationally, and such methods are therefore discussed only in the Appendix in Part II. A comparison of the estimates from this simpler method for the example problem with those previously obtained using a weighted regression analysis shows very close correspondence. This is as expected, since the sample sizes are identical and the censoring schemes are similar at the three test conditions.

The Simpler Estimation Procedure

Notation and Steps 1 to 4. These are identical to those for the simple estimation procedure.

Example.

For the insulation life example, one obtains as previously the estimates

$$\mu_1^* = 3.6381,$$

$$\mu_2^* = 3.2233,$$

$$\mu_3^* = 2.7142,$$

$$\sigma^* = 0.2336.$$

As before,

$$x_1 = 2.256,$$

$$x_2 = 2.159,$$

$$x_3 = 2.028.$$

Step 5. Using a computer program for simple regression analysis, enter the values $x_1, \cdots, x_K$ of the independent variable and $\mu_1^*, \cdots, \mu_K^*$ of the estimates of the location parameter of each distribution of the dependent variable, and obtain the resulting computer estimates β_{0U}^* and β_{1U}^* of the regression coefficients β_0 and β_1. If hand calculations are used, the expressions for the estimates are

$$\beta_{1U}^* = \left\{\sum_{k=1}^{K} \mu_k^* x_k - \frac{1}{K}\left[\sum_{k=1}^{K} \mu_k^*\right]\left[\sum_{k=1}^{K} x_k\right]\right\} \Big/ \left\{\sum_{k=1}^{K} x_k^2 - \frac{1}{K}\left[\sum_{k=1}^{K} x_k\right]^2\right\}$$

$$\beta_{0U}^* = \left\{\sum_{k=1}^{K} \mu_k^* - \beta_{1U}^* \sum_{k=1}^{K} x_k\right\} \Big/ K$$

Example. By hand calculation,

$$\beta_{1U}^* = \{[3.6381(2.256) + 3.2233(2.159) + 2.7142(2.028)] - (1/3)[3.6381 + 3.2233 + 2.7142][2.256 + 2.159 + 2.028]\} / \{[(2.256)^2 + (2.159)^2 + (2.028)^2] - (1/3)[2.256 + 2.159 + 2.028]^2\}$$

$$= 4.043,$$

$$\beta_{0U}^* = \{[3.6381 + 3.2233 + 2.7142] - (-4.043)[2.256 + 2.159 + 2.028]\}/3$$

$$= -5.491.$$

These estimates of the coefficients would be given by a regression program. The corresponding estimates with the simple method using weighted regression analysis were $\beta_0^* = -5.513$ and $\beta_1^* = 4.054$.

Estimation and Prediction at a Specified x_0

Estimate the value of the location parameter and the $100P$'th percentile of the distribution of the dependent variable at the value x_0 of the independent variable as follows.

Step 6. a. The linear estimator of the location parameter of the distribution of the dependent variable at x_0 is

$$\mu_U^*(x_0) = \beta_{0U}^* + \beta_{1U}^* x_0 .$$

This estimator is unbiased.

Example.

a. The estimate of the logarithmic mean life (i.e., location parameter) at the design temperature of 130°C ($x_0 = 1000/(130 + 273.2) = 2.480$) is

$$\mu_U^*(2.480) = -5.491 + 4.043(2.480) = 4.536.$$

The antilogarithm of this value is 34,400 hours and is the corresponding estimate of the median life. Similarly, the estimates for the logarithmic mean and median lives at the design temperature and the three test temperatures are:

Temperature	Estimate of the Logarithmic Mean Life	Estimate of the Mean Life
130°C	4.536	34,400 hours
170°C	3.630	4,260 hours
190°C	3.238	1,730 hours
220°C	2.708	510 hours

Step 6. b. The linear estimator for the $100P$'th percentile of the distribution of the dependent variable at x_0 is

$$y_U^*(P, x_0) = \mu_U^*(x_0) + z(P)\sigma^*$$

where $z(P)$ is the $100P$'th percentile of the standardized form of the assumed

distribution, that is, the assumed distribution with the location parameter equal to zero and the scale parameter equal to one. This estimator is unbiased.

Example.

The estimate of the 10% point of the logarithmic life distribution at the design temperature of 130°C ($x_0 = 2.480$) is

$$y_U^*(0.10, 2.480) = 4.536 + (-1.2816)0.2336 = 4.237$$

where $z(0.10) = -1.2816$ is the 10% point of the standard normal distribution. The antilogarithm of this value is 17,300 hours and is the estimate of the 10% point of the insulation life distribution at the design temperature.

E. *Further Comments*

As can be seen from the preceding discussion, the most laborious aspect of obtaining the simple linear unbiased estimators arises in fitting the regression relationship to the estimates μ_k^* of the location parameter at each test condition. For those with limited calculational capabilities, one possible compromise approach would be to obtain the estimates of μ_k^* and possibly of σ_k^* at each test condition and then to resort to plotting techniques along the lines described by Hahn and Nelson (1970) to estimate the regression relationship.

Also, it should be re-emphasized that the procedures described in this part of this paper are only special cases of the more general methods presented in the Appendix in Part II. Thus, the general methods are not restricted to regression with a single independent variable, censoring only to the right and normally distributed random variation, as they have been in this part.

The estimation methods are based on an assumed form for the distribution of the dependent variable and may give unsatisfactory results if the true distribution differs appreciably from the assumed one. Nelson (1971) provides methods for analysis of residuals for checking the distributional assumption and also the adequacy of the assumed relationship and the validity of the data.

F. *Related Work*

Lieblein and Zelen (1956) present a method for linear unbiased estimation for the coefficients in a linear regression model, when the data are singly censored. The regression model (Palmgren's equation) which they used for analysis of bearing life is based on an assumed underlying Weibull distribution of life. They express selected percentiles of the distribution as the product of powers of the independent variables where the powers are unknown parameters. By a logarithmic transformation of the data, they convert the underlying distribution into an extreme value distribution. Then a selected percentile is a linear function of the logarithms of the independent variables and of the unknown power parameters, which are the regression coefficients. The method they present for estimating the model regression coefficients from censored data is essentially equivalent to the simple method presented here. To obtain the estimates of the regression coefficients from the best linear unbiased estimates of a selected percentile for each combination of independent variables, they used a weighted regression where the weight for a particular estimate is proportional to the

corresponding sample size. This appears to be a satisfactory approximation to our exact simple method which uses as weights the reciprocals of the variances of the estimates of the location parameters. They also present approximate tests of hypotheses on the regression coefficients, where the tests employ the usual quadratic forms used for testing general linear hypotheses. However, they use a quadratic estimate of the mean square for error about the regression in contrast to the simple linear estimate presented here. While not as statistically efficient as the simple method, theirs appears to be robust and well suited to their practical problems.

An application of regression analysis with linear estimation using order statistics is given by Nelson (1970b). He uses the simple method for an analysis of complete data with the inverse power law and the Weibull Distribution.

Acknowledgements

The authors wish to express their appreciation to Dr. Richard L. Shuey, of General Electric Corporate Research and Development, for his support and encouragement of this work. We are indebted to Mr. Delmar Crawford, of the GE Small AC Motor and Generator Department, Fort Wayne, Indiana, and to Mr. Lars Sjodahl, of the GE Flight Propulsion Department, Evendale, Ohio, for their kind permission to use their data to illustrate the methods presented here. It was through questions raised by Mr. Sjodahl on the analysis of his data that the authors were stimulated to develop the methods presented here.

The authors also wish to thank Professor Harry Smith, Jr., former Editor of TECHNOMETRICS and Dean of the School of Management of Rensselaer Polytechnic Institute, for providing the opportunity to present this work at the Technometrics Session at the Fort Collins meeting, and to thank Dr. Donald A. Gardiner, Editor of TECHNOMETRICS, for his deligent personal attention and beneficial comments on this manuscript.

References

[1] Crawford, D. E. (1970). "Analysis of Incomplete Life Test Data on Motorettes," *Insulation/Circuits, 16,* No. 11, 43–8.

[2] Hahn, G. J. and Nelson, W. B. (1970). "Regression Analysis for Censored Data—Graphical Methods," General Electric Company Research and Development Center TIS Report 70-C-383.* Also this appeared in *Insulation/Circuits,* September 1971, 79–84.

[3] Hahn, G. J. and Nelson, W. B. (1971). "Regression Analysis for Censored Data—A Comparison of Maximum Likelihood, Graphical, and Linear Estimation Methods." General Electric Corporate Research and Development TIS Report 71-C-196.*

[4] Lieblein, J. and Zelen, M. (1956). "Statistical Investigation of the Fatigue Life of Deep-Groove Ball Bearings," *Journal of Res. National Bureau of Standards, 57,* 273–316.

[5] Mann, Nancy, R. (1970). "Estimation of Location and Scale Parameters Under Various Models of Censoring and Truncation" containing "Best Linear Invariant Estimation for Weibull Parameters Under Progressive Censoring," Aerospace Research Laboratories Report 70-0026.

[6] Nelson, W. B. (1969). "Theory and Applications of Hazard Plotting for Censored Failure Data," General Electric Research and Development Center TIS Report 69-C-378.* Tentatively accepted for publication in *Technometrics.*

[7] Nelson, W. B. (1970a). "Planning and Statistical Analysis of Accelerated Life Tests—

Methods for Complete Data," General Electric Research and Development Center TIS Report 70-C-294.* This also appeared in the *IEEE Transactions on Electrical Insulation*, EI-6, Dec. 1971, 165–81.

[8] NELSON, W. B. (1970b). "Statistical Methods for Accelerated Life Test Data—The Inverse Power Law Model," General Electric Corporate Research and Development TIS Report 71-C-011.* Graphical methods from this appeared in the *IEEE Transactions on Reliability*, R-21, Feb. 1972, 2-11.

[9] NELSON, W. B. (1971). "Analysis of Residuals from Censored Data—With Applications to Life and Accelerated Test Data," General Electric Corporate Research and Development TIS Report 71-C-120.*

[10] NELSON, W. B. and HAHN, G. J. (1971). "Regression Analysis of Censored Data—Linear Estimation Using Ordered Observations," General Electric Corporate Research and Development TIS Report 71-C-122.*

[11] SARHAN, A. E. and GREENBERG, B. G. (editors) (1962). *Contributions to Order Statistics*, John Wiley and Sons, New York.

[12] THOMAS, D. R. and WILSON, W. M. (1970). "Linear Order Statistic Estimation for the Two-Parameter Weibull and Extreme-Value Distributions from Type II Progressively Censored Samples," Available from the Authors at Oregon State University, Corvallis, Oregon.

---

* General Electric Company Corporate Research & Development TIS Reports may be requested from:

Distribution Unit
Building #5—Room #237
GE Research & Development
Schenectady, New York 12301