This research was supported by the National Institute of Health (Grants GM-70004-04 and HD-00371-02).

# A MODIFIED $\chi^2$ APPROACH FOR FITTING WEIBULL MODELS TO SYNTHETIC LIFE TABLES

By

D. H. Freeman, Jr., Jean L. Freeman, and Gary G. Koch

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 958

OCTOBER 1974

# A MODIFIED $\chi^2$ APPROACH FOR FITTING WEIBULL MODELS TO SYNTHETIC LIFE TABLES

D. H. Freeman, Jr., Jean L. Freeman, and Gary G. Koch

Department of Biostatistics, University of North Carolina
Chapel Hill, N. C. 27514, U.S.A.

## SUMMARY

Weibull models are fitted to synthetic life table data by applying weighted least squares analysis to log log functions which are constructed from appropriate underlying contingency tables. As such, the resulting estimates and test statistics are based on the linearized minimum modified $\chi^2_1$-criterion and thus have satisfactory properties in moderately large samples. The basic methodology is illustrated in terms of an example which is bivariate in the sense of involving two simultaneous, but non-competing, vital events. For this situation, the estimation of Weibull model parameters is described for both marginal as well as certain conditional distributions either individually or jointly.

## 1. INTRODUCTION

A substantial interest in the fitting of Weibull models to biological data is expressed in the current literature. Pike [1966] fits curves of this type to a class of experiments in carcinogenesis in an attempt to study the stochastic process involved. Subsequently, Peto and Lee [1973] extend

and amplify Pike's work for the same class of experiments and identify some problem areas in the estimation of the parameters. An important consideration in the fitting of these models is that the investigations have a longitudinal or cohort design. This would involve groups of subjects (for example, patients of a similar type or mice) who are followed either until the time of occurrence of a particular symptom (eg., death due to a specific cause) or until the time at which they are withdrawn from risk as a result of the termination date of the study or some unrelated cause like lost to follow-up, etc. The theoretical foundation of such studies is discussed in many places including Chiang [1968]. An important point of this work is that when the data are grouped with respect to fixed time intervals, their statistical behavior can be characterized by the multinomial distribution. This permits simplifications in the estimation of the covariance structure of survival rates which in turn provides the basis for their analysis in terms of weighted least squares techniques. In this respect, Koch, Johnson, and Tolley [1972] apply the linear model approach of Grizzle, Starmer, and Koch [1969] (hereafter GSK) for multivariate categorical data to analyze certain aspects of the survivorship function, while Gehan and Siddiqui [1973] apply somewhat different but related procedures to analyze the hazard function.

Unfortunately, not all studies are of the cohort type, usually because of the constraints of time or cost. In these circumstances, period or cross-sectional data are used to generate synthetic life tables and the corresponding survival estimates. Such an investigation may be regarded as a snapshot of the current relationships.

In this paper, we shall be concerned with statistical and methodological issues associated with fitting Weibull models to synthetic life tables.

## 2. SYNTHETIC LIFE TABLES

In a cohort study, there exists a population (often but not necessarily a birth cohort) of subjects who are at risk with respect to the occurrence of a well-defined vital event (eg., death, pregnancy, onset of some disease or physical symptom, motor vehicle accident, etc.) at the beginning of a given time period (eg., a five year period, a ten year period, a fifty year period, etc.). From this population a sample of n subjects is longitudinally followed through time. If none of these subjects is withdrawn from risk as a result of causes unrelated to the vital event of interest (eg., termination date of the study, lost to follow-up, etc.), then a conceptual data vector would have the form shown in (2.1),

$$(d_1, d_2, \ldots, d_r; S_r) \tag{2.1}$$

where $j = 1,2,\ldots,r$ refers to a set of r intervals (eg., days, months, years, etc.) into which the total time period has been partitioned. In this framework, $S_r$ denotes the number of subjects who survive the event of interest throughout the total time period, and $d_j$ for $j = 1,2,\ldots,r$ denotes the number of subjects to whom this event occurs for the first time (during the given time period) within the j-th interval (i.e., the event is survived for all intervals prior to the j-th but not survived during the j-th interval). Thus, for $j = 1,2,\ldots,r$, we have that

$$D_j = \sum_{k=1}^{j} d_k \tag{2.2}$$

represents the number of subjects to whom the event of interest has occurred for the first time by the end of the j-th interval; and for $j = 1,2,\ldots,(r-1)$,

$$S_j = \sum_{k=(j+1)}^{r} d_k + S_r \qquad (2.3)$$

represents the number of subjects who have survived the event of interest through the j-th interval. Since all subjects have either survived through the j-th interval or have experienced the event prior to the end of the j-th interval, it follows that

$$D_j + S_j = n \qquad (2.4)$$

for all $j = 1,2,\ldots,r$. All of these considerations are summarized in a life table format in Table 1.

TABLE 1

THE COHORT STUDY LIFE TABLE

| Interval | Number of subjects in total study | Number of subjects who survive event through j-th interval | Number of subjects to whom event occurs by end of j-th interval | Number of subjects to whom event occurs during j-th interval | Number of subjects at risk for event during j-th interval |
|---|---|---|---|---|---|
| 1 | n | $S_1$ | $D_1$ | $d_1$ | n |
| 2 | n | $S_2$ | $D_2$ | $d_2$ | $S_1$ |
| ... | ... | ... | ... | ... | ... |
| r | n | $S_r$ | $D_r$ | $d_r$ | $S_{r-1}$ |

Under the assumption that the data vector in (2.1) has a multinomial distribution, we have that for $j = 1,2,\ldots,r$,

$$P_j = (S_j/n) \qquad (2.5)$$

is an unbiased, unrestricted maximum likelihood estimator for the probability

that a randomly chosen subject from the cohort survives the event of interest through the j-th interval. Moreover, for reasons discussed in both Chiang [1968] and Koch, Johnson, and Tolley [1972], the data vector (2.1) can be regarded as a contingency table from which the quantities $P_j$ are obtained by linear transformation operations. Thus, the weighted least squares methodology described by Grizzle, Starmer, and Koch [1969] can be used to fit various survivorship function models like the exponential, logistic, Weibull, etc. to the $P_j$'s for $j = 1,2,\ldots,r$. The resulting estimates of underlying model parameters and corresponding goodness of fit test statistics belong to the minimum modified $\chi^2_1$-class due to Neyman [1949].

Alternatively, the period study is constructed from separate samples drawn from several non-overlapping sub-populations, each of which corresponds to a specific range of values (eg., age range) for the overall time period of exposure to risk for the occurrence of the vital event of interest and hence represents a continuously grouped set of cohorts. This sampling takes place cross-sectionally at a single instant in time when subjects are observed with respect to the occurrence status (i.e., yes or no) of the vital event. Thus, a contingency table such as that shown in Table 2 may be constructed.

TABLE 2

THE PERIOD STUDY CONTINGENCY TABLE

| Exposure (age) interval range | Total sample size | Occurrence status of event of interest | |
|---|---|---|---|
| | | No | Yes |
| 1 | $n_1$ | $\bar{S}_1$ | $\bar{D}_1$ |
| 2 | $n_2$ | $\bar{S}_2$ | $\bar{D}_2$ |
| ... | ... | ... | ... |
| r | $n_r$ | $\bar{S}_r$ | $\bar{D}_r$ |

Here, $j = 1,2,\ldots,r$ indexes the set of exposure (age) interval ranges, and for the j-th interval, $n_j$, $\bar{S}_j$, and $\bar{D}_j$ respectively denote the sample size, the number of subjects to whom the vital event of interest has not yet occurred, and the number of subjects to whom the vital event has already occurred. Thus,

$$\bar{P}_j = (\bar{S}_j/n_j) \tag{2.6}$$

represents the proportion of subjects associated with the j-th exposure (age) interval range who have survived the vital event of interest through the instant in time at which the period study is conducted. Although the $\bar{P}_j$ appear to reflect the survival experience of the subjects in some underlying population, their interpretation in this context requires the following assumptions:

i. There is no remission from the event of interest; i.e., once an individual is observed with the event, he is never observed subsequently without it.

ii. There is no in or out migration, or less restrictively, migration behavior is statistically independent of the event of interest.

iii. The occurrence rates for the event of interest are constant over time; this is the usual assumption that the occurrence rates do not depend on the starting date of exposure to risk.

Under these conditions, the data for the period study in Table 2 can be regarded as a synthetic life table. However, even in this case, $\bar{P}_j$ does not represent a true survival rate for the j-th interval in a strict sense since the periods of exposure to risk (eg., ages) of the subjects associated with this interval are usually distributed throughout the corresponding range as opposed to being directly linked identically to a specific point (eg., the

right hand endpoint). Moreover, in reference to the underlying set of cohorts which pertain to the sample of $n_j$ subjects who are exposed to risk for time periods ranging through the end of the j-th interval, these considerations necessarily imply that

$$S_{j-1,j} \geq \bar{S}_j \geq S_{j,j} \tag{2.7}$$

where $S_{j-1,j}$ is the unknown number of subjects associated with the j-th interval for whom the vital event of interest had not occurred when they initially entered this exposure interval (i.e., at the left hand endpoint) and $S_{j,j}$ is the unknown number of subjects for whom it will not have occurred by the time they are finally withdrawn from this exposure interval (i.e., at the right hand endpoint). From this result, it follows that

$$P_{j-1,j} = (S_{j-1,j}/n_j) \geq \bar{P}_j \geq (S_{j,j}/n_j) = P_{j,j} \tag{2.8}$$

where $P_{j-1,j}$ and $P_{j,j}$ are directly analogous to $P_{j-1}$ and $P_j$ in the corresponding cohort study framework when the conditions (i) - (iii) apply. Thus, in a certain sense, $\bar{P}_j$ tends to reflect an average of survival rates which are associated with the respective time points throughout the j-th interval and which conceptually range between $P_{j-1}$ and $P_j$. For this reason, it is customary to regard $\bar{P}_j$ as an estimate of the probability that the vital event of interest is survived through the midpoint of the j-th interval. This interpretation is particularly appropriate if the following additional condition applies:

iv. The survival rates associated with the respective time points throughout the j-th interval are symmetrically distributed with respect to the midpoint of the interval.

Hence, when the conditions (i) - (iv) are applicable, the synthetic life table

associated with a period study can be analyzed in terms of the same underlying probability models which are pertinent with respect to analogous cohort studies. As a result, once an understanding of the implications of the assumptions (i) - (iv) is achieved, cohort studies and period studies as well as life table analysis and contingency table analysis can be unified into a common methodological framework.

## 3. CHARACTERIZATION OF THE WEIBULL DISTRIBUTION

The Weibull distribution has been of interest to statisticians since its introduction to the literature, partly because of its usefulness in situations suggesting increasing or decreasing hazard functions. (For example, see Johnson and Kotz [1970] or Gehan and Siddiqui [1973]). If t represents the time to the occurrence of an event of interest (eg., a death or the detection of a tumor), then the Weibull cumulative distribution function may be written as:

$$G(t|\mu,\delta,w) = 1 - \exp\{ -\mu(t-w)^{\delta} \} \text{ for } t \geq w \text{ where } \mu,\ \delta \geq 0, \qquad (3.1)$$

with the interpretation of the parameters, ($\mu$, $\delta$, w) depending on the type of data being analyzed.

Doll [1971] surveys the literature on cancer research to examine various stochastic models of cancer. He discusses the exponential time dependence of cancers as noted by Armitage and Doll [1954, 1957, 1961], Cook, Doll, and Tellingham [1967], Doll and Hill [1964], Fisher [1958], and Pike and Doll [1965]. Pike [1966] indicated that these models of the etiology of carcinomas suggest the usefulness of the Weibull distribution in analyzing longitudinal experiments in carcinogenesis. His conclusion is based

on the work of Fisher and Tippett [1928] and Gumbell [1954, 1958] on the distribution of extreme value statistics. Pike's work was enlarged upon by Lee and O'Neill [1971], Peto, Lee, and Paige [1972], and Peto and Lee [1973]. The essential point in this body of work is to note the applicability of Weibull type distributions in circumstances where a carcinogen is applied in a relatively uniform and continuous manner (for example, weekly skin paintings), and the variable of interest is the time to appearance of a tumor; or where the relative strengths of the carcinogens are of interest. In these circumstances Peto, Lee, and Paige [1972] interpret the parameters as follows: $\mu$ is a rate determining scale parameter, while $\delta$ and w characterize the process by which the tumor develops. Accordingly, hypotheses concerning $\mu$ would examine different intensities of carcinogens while differences among the $\delta$ indicate different processes. Given the argument concerning the applicability of the Weibull distribution in these circumstances, it also emerges as a plausible distribution for examining disease patterns which are the result of long term exposure to substances which are generally suspected to be deleterious to health even though their immediate results are not directly observable; that is, where an increasing hazard may be suspected with respect to some vital event of interest. Such a situation is discussed in the example of Section 5.

## 4. REPEATED LOGARITHM MODELS

The Forthofer and Koch [1973] formulation of the GSK linear models approach to categorical data analysis may be used to fit the Weibull models of Section 3. From equation (3.1), it follows that

$$\ln\{1 - G(t|\mu,\delta,w)\} = -\mu(t-w)^{\delta} \qquad (4.1)$$

where t refers either to time or age. In this paper, we shall assume that w can be assumed to have a fixed known value (eg., w = 0) which can be justified in terms of the phenomenon under investigation. As a result, a linear model involving the parameters (ln $\mu$) and $\delta$ can be obtained by multiplying both sides of (3.1) by (-1) and then applying logarithmic transformations a second time as shown in (4.2).

$$\theta(t) = \ln[-\ln\{1 - G(t|\mu,\delta,w)\}] = \ln \mu + \delta \ln (t-w). \qquad (4.2)$$

The weighted least squares methodology in GSK can be used to fit the model (4.2) to sample estimates of $\theta(t)$ for which consistent estimates of variance can be obtained by methods described in Forthofer and Koch [1973]. Such results can then be used to make comparisons among corresponding parameters for different populations. In certain circumstances, such relationships may be of interest for certain types of bivariate data where there are two non-competing vital events of interest. For the synthetic life tables arising from period studies, there are limitations on the extent to which bivariate relationships can be investigated since the full bivariate life table is not available. Nevertheless, as shown in Section 5.2, predicted values for each marginal event as well as their simultaneous co-occurrence can be computed.

## 5. EXAMPLE

Ashford and Sowden [1970] analyze data from a survey of working coalminers at a representative sample of collieries distributed throughout the United Kingdom. Each subject was classified as to whether he reported the symptoms of breathlessness and wheeze. These data are shown in Table 3.

TABLE 3

OBSERVED FREQUENCY OF SYMPTOMS AND MARGINAL PROPORTIONS

| Age Group | Breathless Yes Wheeze Yes $\bar{D}_{1j}$ | Breathless Yes Wheeze No $\bar{D}_{2j}$ | Breathless No Wheeze Yes $\bar{D}_{3j}$ | Breathless No Wheeze No $\bar{D}_{4j}$ | Breathless Margin $P_{1j}$ | Wheeze Margin $P_{2j}$ | Survive Wheeze Given Breathless $P_{3j}$ |
|---|---|---|---|---|---|---|---|
| 20-24 | 9 | 7 | 95 | 1841 | 0.9918 | 0.9467 | 0.4375 |
| 25-29 | 23 | 9 | 105 | 1654 | 0.9821 | 0.9285 | 0.2813 |
| 30-34 | 54 | 19 | 177 | 1863 | 0.9654 | 0.8906 | 0.2603 |
| 35-39 | 121 | 48 | 257 | 2357 | 0.9393 | 0.8642 | 0.2840 |
| 40-44 | 169 | 54 | 273 | 1778 | 0.9019 | 0.8056 | 0.2421 |
| 45-49 | 269 | 88 | 324 | 1712 | 0.8508 | 0.7522 | 0.2465 |
| 50-54 | 404 | 117 | 245 | 1324 | 0.7507 | 0.6895 | 0.2246 |
| 55-59 | 406 | 152 | 225 | 967 | 0.6811 | 0.6394 | 0.2724 |
| 60-64 | 372 | 106 | 132 | 526 | 0.5792 | 0.5563 | 0.2218 |

They have also been analyzed in a number of other papers (Grizzle [1971], Kullback and Fisher [1973], Mantel and Brown [1973]); however, the reason for reanalyzing them here is to illustrate the procedure for applying a Weibull model to a complex synthetic life table. Thus, the fitting of a univariate Weibull model for the event of reporting breathlessness is discussed in Section 5.1, the fitting of a two symptom model using Weibull margins is discussed in Section 5.2, and the interpretation of such models in the synthetic life table framework is discussed in Section 5.3.

5.1. A Univariate Model

Let x denote age as it ranges from birth (x = 0) to 100 years in five year intervals. Let j index these five year age groups by corresponding to the right endpoint of the age group so that $j = \{\frac{x}{5}\} = 1,2,3,\ldots$ for $x = 5, 10,15,\ldots$, respectively. Since this is a period study, the discussion of

Section 2 requires that the parameter w be used to shift the survival probabilities back to the midpoint of the age interval. In the notation of Section 2, $\bar{P}_j$ denotes the proportion of subjects in the j-th age interval range who have survived (in the sense of not reporting) breathlessness through the instant in time at which the survey was conducted. Thus, from (3.1), the fitted Weibull model for surviving breathlessness is

$$\bar{P}_j \hat{=} \exp\{-\mu(j - 0.5)^{\delta}\}, \tag{5.1}$$

where "$\hat{=}$" means "is an estimate of." Using the approach outlined in Section 4, we then have,

$$\hat{\theta}_j = \ln\{-\ln\{\bar{P}_j\}\} \hat{=} \ln \mu + \delta \ln(j - 0.5). \tag{5.2}$$

In matrix notation, let $\hat{\underset{\sim}{H}}' = (\hat{\theta}_5, \hat{\theta}_6, \ldots, \hat{\theta}_{13})$ be our vector of functions of estimated survival probabilities. Then it may be written as

$$\hat{\underset{\sim}{H}} = \underset{\sim}{L}(\underset{\sim\sim}{\log_e}\{\underset{\sim}{K}\ \underset{\sim\sim}{\log_e}[\ \underset{\sim}{\bar{P}}\ ]\}), \tag{5.3}$$

where $\underset{\sim}{\bar{P}}' = (\bar{P}_5, \bar{P}_6, \ldots, \bar{P}_{13})$, $\underset{\sim}{K} = -\underset{\sim}{I}_9$, $\underset{\sim}{L} = \underset{\sim}{I}_9$, $\underset{\sim}{I}_9$ is a 9 x 9 identity matrix, and $\underset{\sim\sim}{\log_e}$ is a natural log function of the vector applied to each component. If one starts with the vector of observed frequencies in Table 3, then $\underset{\sim}{\bar{P}}$ is written, $\underset{\sim}{\bar{P}} = \underset{\sim}{A}\ \underset{\sim}{p}$, where

$$\underset{\sim}{p}' = (p_{5,1}, p_{5,2}, p_{5,3}, p_{5,4}, p_{6,1}, \cdots, p_{13,1}, p_{13,2}, p_{13,3}, p_{13,4}),$$

and $p_{jk}$ denotes the proportion of subjects in age interval j in symptom class k (k is 1,2,3,4 for breathless and wheeze, breathless and no wheeze, no breathless and yes wheeze, and no symptom, respectively), and $\underset{\sim}{A} = [0\ 0\ 1\ 1] \otimes \underset{\sim}{I}_9$. Here, $\otimes$ denotes Kronecker product. Thus, $\hat{\underset{\sim}{H}}$ may then be fitted to the following linear model,

$$E_A\{\hat{\underset{\sim}{\Theta}}\} = \underset{\sim}{X}\,\underset{\sim}{\beta} = \begin{bmatrix} 1 & \log 4.5 \\ 1 & \log 5.5 \\ \cdots & \cdots \\ 1 & \log 12.5 \end{bmatrix} \begin{bmatrix} \log \mu \\ \delta \end{bmatrix},$$

where $E_A\{\cdot\}$ means asymptotic expectation. The numerical results and fitted values between age 0 and 89 are shown in Table 4. The goodness of fit

TABLE 4

OBSERVED AND FITTED PROPORTIONS SURVIVING BREATHLESSNESS

| Age Group | Total Subjects $n_j$ | Observed Surviving Breathlessness $P_j$ | Fitted Surviving Breathlessness $\hat{P}_j$ |
|---|---|---|---|
| 0-4 | ---- | ---- | 1.0000 |
| 5-9 | ---- | ---- | 0.9999 |
| 10-14 | ---- | ---- | 0.9994 |
| 15-19 | ---- | ---- | 0.9975 |
| 20-24 | 1952 | 0.9918 | 0.9928 |
| 25-29 | 1791 | 0.9821 | 0.9831 |
| 30-34 | 2113 | 0.9654 | 0.9660 |
| 35-39 | 2783 | 0.9393 | 0.9385 |
| 40-44 | 2274 | 0.9019 | 0.8978 |
| 45-49 | 2393 | 0.8508 | 0.8413 |
| 50-54 | 2090 | 0.7507 | 0.7678 |
| 55-59 | 1750 | 0.6811 | 0.6780 |
| 60-64 | 1136 | 0.5792 | 0.5749 |
| 65-69 | ---- | ---- | 0.4643 |
| 70-74 | ---- | ---- | 0.3539 |
| 75-79 | ---- | ---- | 0.2520 |
| 80-84 | ---- | ---- | 0.1658 |
| 85-89 | ---- | ---- | 0.0996 |

| Parameter | Parameters for Breathlessness Fitted Model | |
|---|---|---|
| | Est. | S.E. |
| $\ln \mu$ | -11.303 | 0.245 |
| $\delta$ | 4.241 | 0.105 |
| **Hypothesis Test** | $\chi^2$ | d.f. |
| Model: $\delta = 0$ | 1628.022 | 1 |
| Goodness of Fit | 6.138 | 7 |

$\chi^2$-statistic indicates a satisfactory fit and the test of degeneracy of the model ($H_0$: $\delta = 0$) is highly significant.

## 5.2. A Two Symptom Model

The original data given by Ashford and Sowden [1970] was for two symptoms, wheeze as well as breathlessness. Using the formulation for the breathlessness margin, it is possible to fit Weibull models to each margin and some appropriate measure of association. It was observed in preliminary work that the proportion of subjects who reported wheeze given they reported breathlessness was nearly constant, so a model using this as a measure of association was fitted. The full data set for these three margins is shown in Table 3. Let $\bar{P}_{1j}$, $\bar{P}_{2j}$, and $\bar{P}_{3j}$ denote the proportions surviving breathlessness, wheeze, and wheeze given reporting breathlessness. As before,

$$\left.\begin{aligned} \hat{\theta}_{1j} &= \ln\{-\ln \bar{P}_{1j}\} \cong \ln \mu_1 + \delta_1 \ln(j - 0.5) \\ \hat{\theta}_{2j} &= \ln\{-\ln \bar{P}_{2j}\} \cong \ln \mu_2 + \delta_2 \ln(j - 0.5) \\ \hat{\theta}_{3j} &= \ln\{-\ln \bar{P}_{3j}\} \cong \ln \mu_3 + \delta_3 \ln(j - 0.5) \end{aligned}\right\} (j = 5,6,\ldots,13) \quad (5.4)$$

or in matrix notation $\hat{\underset{\sim}{H}}_1' = (\hat{\theta}_{1,5},\hat{\theta}_{1,6},\ldots,\hat{\theta}_{1,13})$, $\hat{\underset{\sim}{H}}_2' = (\hat{\theta}_{2,5},\hat{\theta}_{2,6},\ldots,\hat{\theta}_{2,13})$, $\hat{\underset{\sim}{H}}_3' = (\hat{\theta}_{3,5},\hat{\theta}_{3,6},\ldots,\hat{\theta}_{3,13})$ and

$$\hat{\underset{\sim}{H}} = \begin{bmatrix} \hat{\underset{\sim}{H}}_1 \\ \hat{\underset{\sim}{H}}_2 \\ \hat{\underset{\sim}{H}}_3 \end{bmatrix}$$

where $\hat{\underset{\sim}{H}}_i = \underset{\sim}{L}_i(\underset{\sim}{\log_e}\{\underset{\sim}{K}_i \underset{\sim}{\log_e}[\underset{\sim}{A}_i \underset{\sim}{p}]\})$, $i = 1,2,3$. Here $\underset{\sim}{p}$ is as described in 5.1, $\underset{\sim}{A}_1 = [0\ 0\ 1\ 1] \otimes \underset{\sim}{I}_9$, $\underset{\sim}{A}_2 = [0\ 1\ 0\ 1] \otimes \underset{\sim}{I}_9$, $\underset{\sim}{A}_3 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \otimes \underset{\sim}{I}_9$, $\underset{\sim}{K}_1 = \underset{\sim}{K}_2 = -\underset{\sim}{I}_9$, $\underset{\sim}{K}_3 = [-1\ \ 1] \otimes \underset{\sim}{I}_9$, $\underset{\sim}{L}_1 = \underset{\sim}{L}_2 = \underset{\sim}{L}_3 = \underset{\sim}{I}_9$. Thus, $\hat{\underset{\sim}{H}}$ may then be fitted to the linear model,

$$\underset{\sim}{E}_A\{\hat{\underset{\sim}{H}}\} = \underset{\sim}{X}\,\underset{\sim}{\beta} = \begin{bmatrix} \underset{\sim}{X}_1 & \underset{\sim}{0}_{9,2} & \underset{\sim}{0}_{9,1} \\ \underset{\sim}{0}_{9,2} & \underset{\sim}{X}_2 & \underset{\sim}{0}_{9,1} \\ \underset{\sim}{0}_{9,2} & \underset{\sim}{0}_{9,2} & \underset{\sim}{X}_3 \end{bmatrix} \begin{bmatrix} \ln \mu_1 \\ \delta_1 \\ \ln \mu_2 \\ \delta_2 \\ \ln \mu_3 \end{bmatrix},$$

where,

$$\underset{\sim}{X}_1 = \underset{\sim}{X}_2 = \begin{bmatrix} 1 & \ln 4.5 \\ 1 & \ln 5.5 \\ \cdots & \cdots \\ 1 & \ln 12.5 \end{bmatrix}, \quad \underset{\sim}{X}_3 = \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix},$$

and $\underset{\sim}{0}_{9,2}$ and $\underset{\sim}{0}_{9,1}$ are appropriate matrices of zeroes. The resulting estimates and some hypotheses of interest are given in Table 5.

TABLE 5

PARAMETER ESTIMATES, STANDARD ERRORS, AND FIT OF MODEL FOR JOINT MODELS

| | Joint Model | | | | | |
|---|---|---|---|---|---|---|
| | 1 | | 2 | | 3 | |
| Parameter | Est. | S.E. | Est. | S.E. | Est. | S.E. |
| $\ln \mu$ | -11.173 | 0.237 | -6.700 | 0.143 | 0.335 | 0.025 |
| $\delta$ | 4.185 | 0.102 | 2.420 | 0.063 | ----- | ----- |
| Hypothesis Test | | | $\chi^2$ | | d.f. | |
| Model: Parameter Degeneracy: $\ln \mu_1 - \ln \mu_2 = 0$; $\ln \mu_1 - \ln \mu_3 = 0$; $\delta_1 = 0$; $\delta_2 = 0$ | | | 6458.161 | | 4 | |
| Uniformity given model: An overall mean is equal to ½: $\ln \mu = \ln(-\ln ½)$ | | | 816.149 | | 1 | |
| Error or goodness of fit | | | 26.295 | | 22 | |
| Total variation from uniformity | | | 7300.605 | | 27 | |
| Process Difference: $\delta_1 - \delta_2 = 0$ | | | 365.640 | | 1 | |
| Equal Rates: $\ln \mu_1 - \ln \mu_2 = 0$ | | | 459.054 | | 1 | |
| Equal Margins: $\ln \mu_1 - \ln \mu_2 = 0$; $\delta_1 - \delta_2 = 0$ | | | 742.125 | | 2 | |
| Age Independence: $\delta_1 = 0$; $\delta_2 = 0$ | | | 2182.163 | | 2 | |

## 5.3. Interpretation of a Synthetic Life Table

Sections 5.1 and 5.2 demonstrate the procedure for fitting Weibull models to synthetic life tables. The interpretation of such fitted models depends largely on the applicability of assumptions (i) - (iv) in Section 2 to data sets viewed in this manner. It would seem reasonable to view the population that was sampled as a conceptual population defined by its occupation of coal mining. As such, the subjects are continuously exposed to coal dust and other material which may have effects on the respiratory system; furthermore, there may be certain constraints on the occupational mobility of coal miners which would tend to limit the likelihood of subjects seeking alternative employment. Thus, the interpretation of the results of this analysis should be governed by the extent to which the following assumptions are approximately valid:

i. Subjects are reasonably aware of the potential nature of respiratory symptoms and as such are able to report accurately the presence of the symptoms; thus, remission of such reporting of symptoms is not likely to occur as long as a miner continues in the coal mining occupation.

ii. The occupational age structure is relatively fixed; i.e., migration, in an occupational sense, is relatively minimal and independent of the age-symptom structure.

iii. Exposure is dependent on length of employment, and thus occurrence rates are unchanging.

iv. The five year age groups are sufficiently short in the light of the argument in Section 2 to ensure near symmetry for the occurrence rates in each conceptual exposure interval.

It is our opinion that these assumptions are not unrealistic for this study, and hence, we feel that epidemiological inferences can be drawn to some extent from this synthetic life table for an occupationally defined population. Nevertheless, at a minimum, the fit of the models suggests the usefulness of Weibull models for further examination of respiratory symptoms, possibly in cohort or longitudinal studies.

A last point to be noted is that the joint model chosen in Section 5.2 is one of many different alternatives. It was selected for illustration because of the interesting relationship of the marginals to a constant conditional model. However, it is recognized that it generates inadmissable probability estimates for ages greater than 92.5. Alternative measures of association which avoid this difficulty, such as the log cross product ratio, were also examined.

## 6. REFERENCES


Armitage, P. and Doll, R. [1954]. The age distribution of cancer and a multi-stage theory of carcinogenesis. British Journal of Cancer 8, 1-12.

Armitage, P. and Doll, R. [1957]. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. British Journal of Cancer 11, 161-69.

Armitage, P. and Doll, R. [1961]. Stochastic models for carcinogenesis. Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability 4, 19-38.

Ashford, J. R. and Sowden, R. R. [1970]. Multi-variate probit analysis. Biometrics 26, 535-46.

Chiang, Chin Long. [1961]. A stochastic study of the life table and its applications: III. The follow-up study with the consideration of competing risks. Biometrics 17, 57-78.

Cook, P., Doll, R. and Fellingham, S. A. [1961]. A mathematical model for the age distribution of cancer in man. International Journal of Cancer 4, 93-112.

Doll, Richard. [1971]. The age distribution of cancer: implications for models of carcinogenesis. Journal of the Royal Statistical Society, A 134, 133-66.

Doll, R. and Hill, A. B. [1964]. Mortality in relation to smoking: ten years' observations of British doctors. British Medical Journal i, 1399-410, 1460-67.

Fisher, J. C. [1958]. Multiple-mutation theory of carcinogenesis. Nature 181, 651-52.

Fisher, R. A. and Tippett, L. H. C. [1928]. Limiting forms of the frequency distribution of the largest and smallest member of a sample. Proceedings of the Cambridge Philosophical Society 24, 180-90.

Forthofer, R. N. and Koch, G. G. [1973]. An analysis for compounded functions of categorical data. Biometrics 29, 143-57.

Gehan, E. A. and Siddiqui, M. M. [1973]. Simple regression methods for survival time studies. Journal of the American Statistical Association 68, 848-56.

Grizzle, J. E. [1971]. Multivariate logit analysis. Biometrics 27, 1057-62.

Grizzle, J. E., Starmer, C. F. and Koch, G. G. [1969]. Analysis of categorical data by linear models. Biometrics 25, 489-504.

Gumbel, E. J. [1954]. Statistical Theory of Extreme Values and Some Practical Applications, National Bureau of Standards Applied Mathematics Series 33.

Gumbel, E. J. [1958]. Statistics of Extremes. Columbia University Press, New York.

Johnson, N. L. and Kotz, S. [1970]. Distributions in Statistics: Continuous Univariate Distributions, Volume I. Houghton Miffin Company, Massachusetts.

Koch, G. G., Johnson, W. D. and Tolley, H. D. [1972]. A linear models approach to the analysis of survival and extent of disease in multidimensional contingency tables. Journal of the American Statistical Association 67, 783-96.

Kullback, S. and Fisher, M. [1973]. Partitioning second-order interaction in three-way contingency tables. Journal of the Royal Statistical Society C 22, 172-84.

Lee, P. N. and O'Neill, J. A. [1971]. The effect both of time and dose applied on tumour incidence rate in benzpyrene skin painting experiments. British Journal of Cancer 25, 759-70.

Mantel, N. and Brown, C. [1973]. A logistic reanalysis of Ashford and Sowden's data on respiratory symptoms in British coalminers. Biometrics 29, 649-66.

Neyman, J. [1949]. Contributions to the theory of the $\chi^2$ test. Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, Berkeley and Los Angeles: University of California Press, 239-272.

Peto, R. and Lee, P. [1973]. Weibull distributions for continuous-carcinogenesis experiments. Biometrics 29, 457-70.

Peto, R., Lee, P. N. and Paige, W. S. [1972]. Statistical analysis of continuous carcinogenesis. British Journal of Cancer 26, 258-61.

Pike, M. C. [1966]. A method of analysis of a certain class of experiments in carcinogenesis. Biometrics 22, 142-61.

Pike, M. C. and Doll, R. [1965]. Age at onset of lung cancer: significance in relation to effect of smoking. Lancet i, 665-8.