ON SOME METHODOLOGICAL ISSUES IN THE ANALYSIS OF SURVIVAL DATA FROM PROSPECTIVE-TYPE EXPERIMENTS

by

Regina C. Elandt-Johnson

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1340
May 1981

ON SOME METHODOLOGICAL ISSUES IN THE ANALYSIS OF SURVIVAL DATA FROM PROSPECTIVE-TYPE EXPERIMENTS

Regina C. Elandt-Johnson*

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

ABSTRACT

Let X denote age, T be the survival time since occurrence of a certain initial event at age x, and $\underset{\sim}{z}' = (z_1, z_2, \ldots, z_S)$ be a vector of characteristics measured on individual aged x. The purpose of this paper is to construct survival models for prospective-type experiments, emphasizing that the original lifetime distribution, $S_X(x; \underset{\sim}{z})$, plays an essential role and cannot be ignored. The survival distribution since initial event at age x (e.g. entry to the study, onset of a disease) is the future lifetime distribution, $S_T(t|x; \underset{\sim}{z})$, and is closely related to $S_X(x; \underset{\sim}{z})$. In the original population, this would be $S_X(x+t|X>x; \underset{\sim}{z})$. Exponential hazard rate models are suggested and justified, to some extent, in regard to their mathematical properties and simple biological interpretation of their parameters. When the covariables are linear functions of age, there is a close relation to Cox's model, and an approximate relationship with the multiple logistic linear model. The discussion is extended to competing risk analysis with several causes of death. Non-Gompertzian models are constructed by introducing non linear functional dependence of z's on x, or by allowing for interaction between the z's and x. The relationship between the z's and x can be estimated from repeated measurements of z's on each individual. The techniques of model building are illustrated by several examples.

---

*) This work was supported by U.S. National Heart, Lung and Blood Institute contract NIH-NHLI-712243 from the National Institutes of Health.

1. INTRODUCTION

1.1. Probability Models in Follow-Up Studies

In the last two decades there has been considerably increased interest in prognostic factors of mortality from chronic diseases, especially those classified as being of malignant or cardiovascular type. There is a vast number and variety of relevant studies, but here we restrict ourselves to two major types: epidemiologic prospective follow-up studies and clinical trials.

Let $\underset{\sim}{z}' = (z_1, \ldots, z_S)$ denote a vector of S characteristics, named briefly concomitant variables or covariables, which are considered as potential prognostic factors; let x be the age, and t the survival time since occurrence of a certain initial event, such as onset of a disease or entry to the study, for an individual at age x. The probabilistic and statistical problems lie in constructing a model of the survival distribution function (SDF), $S_T(t|x; \underset{\sim}{z}) = \Pr\{T>t|x; \underset{\sim}{z}\}$, or equivalently, a model for the hazard rate, $\lambda_T(t|x; \underset{\sim}{z}) = f_T(t|x; \underset{\sim}{z})/S_T(t|x; \underset{\sim}{z})$, where $f_T(t|x; \underset{\sim}{z})$ is the probability density function, and estimating the parameters, including those which represent the contributions of concomitant variables.

Two models have been commonly used:

(i) The logistic linear model

$$\log \frac{q(\tau|x; \underset{\sim}{z})}{p(\tau|x; \underset{\sim}{z})} = \gamma + \alpha_0^* x + \underset{\sim}{\beta}^{*\prime} \underset{\sim}{z} , \qquad (1.1)$$

where $q(\tau|x; z)$ and $p(\tau|x; \underset{\sim}{z}) = 1-q(\tau|x; \underset{\sim}{z})$ are the probabilities of death and survival, respectively, over a specified period τ, and

(ii) Cox's (1972) exponential hazard rate model

$$\lambda_T(t|x; \underset{\sim}{z}) = \exp(\alpha_0 x + \underset{\sim}{\beta}'\underset{\sim}{z})\lambda_0(t), \qquad (1.2)$$

where $\lambda_0(t)$ - sometimes called the underlying hazard - is solely a

function of t.

Apart from the fact that the logistic function is inappropriate to use in analysis of survival data where each individual (j) has his own potential exposure time τ_j, both models are constructed by using $T>0$ as a stochastic variable, without paying attention to what would be the original lifetime distribution $S_X(x; \underset{\sim}{z}) = \Pr\{X>x; \underset{\sim}{z}\}$. In fact, no special notation is used for age, which is regared as just one of the elements of the vector $\underset{\sim}{z}$.

The purpose of this article is to develop a methodology based on the original lifetime survival distribution, $S_X(x; \underset{\sim}{z})$.

We note that if $\mu_X(x; \underset{\sim}{z})$ is the hazard rate (force of mortality) of $S_X(x; \underset{\sim}{z})$, then the hazard rate for future lifetime distribution is

$$\mu_X(x+t|X>x; \underset{\sim}{z}) = \mu_X(x+t; \underset{\sim}{z}) = \lambda_T(t|x; \underset{\sim}{z}), \qquad (1.3)$$

and the distribution of the follow-up time for an individual aged x is the future lifetime survival distribution

$$S_T(t|x; \underset{\sim}{z}) = \exp[-\int_0^t \lambda_T(u|x; \underset{\sim}{z})du]$$
$$= S_X(x+t; \underset{\sim}{z})/S_X(x; \underset{\sim}{z}) . \qquad (1.4)$$

As we shall see later, it is only under certain conditions that the future lifetime hazard rate (1.3) coincides with that given by (1.2).

It is worthwhile noticing that:

(a) The hazard rate of the future lifetime distribution, $\lambda_T(t|\underset{\sim}{z})$, is the same as the hazard rate of the original (unconditional) distribution, $\mu_X(x+t; \underset{\sim}{z})$. (formula (1.3))

(b) It is usually greater than zero at t=0.

1.2. Preservation of $S_X(\cdot)$ in $S_T(\cdot)$

Suppose that $\mu_X(x; \underset{\sim}{z})$ can be factorized into two terms

$$\mu_X(x; \underset{\sim}{z}) = C(\underset{\sim}{z}_0)\psi_X(x), \tag{1.5}$$

where $C(\underset{\sim}{z}_0)$ is a constant, possibly depending on some fixed values $\underset{\sim}{z}_0$, and $\psi_X(x)$ is solely a function of x. Suppose that it is also possible to factorize the future lifetime hazard rate function into two terms

$$\lambda_T(t|x; \underset{\sim}{z}) = h(x; \underset{\sim}{z})\psi_T(t), \tag{1.6}$$

where $\psi_T(t)$ is solely a function of t. If further,

$$\psi_T(t) = \psi_X(t), \quad \text{for all } t>0, \tag{1.7}$$

then we say that the lifetime distribution $S_X(x; \underset{\sim}{z})$ is *preserved* in the future lifetime distribution $S_T(t|x; \underset{\sim}{z})$.

If the preservation property holds, the survivorship analysis based on future lifetime distribution (briefly, T-analysis) is the same as the analysis based on original lifetime distribution (briefly, X-analysis).

EXAMPLE 1. Let $\mu_X(x) = Re^{\alpha x}$, $R>0$, $\alpha>0$, $x>0$. Then $\mu_X(x+t) = \lambda_T(t|x) = Re^{\alpha x}e^{\alpha t}$. Here $\psi_X(x) = e^{\alpha x}$; and $\psi_T(t) = e^{\alpha t}$, so that $\psi_X(t) = \psi_T(t)$ - the preservation property holds. On the other hand, suppose that $\mu_X(x) = \theta x^c$, $\theta>0$, $c>0$, $x>0$. Then $\mu_X(x+t) = \theta(x+t)^c$ — here the preservation property does not hold.

We also notice that if (1.5)-(1.7) hold, we have

$$S_X(x; \underset{\sim}{z}_0) = \exp[-C(\underset{\sim}{z}_0)\int_0^x \psi_X(y)dy],$$

and

$$\begin{aligned} S_T(t|x; \underset{\sim}{z}) &= \exp[-h(x; \underset{\sim}{z})\int_0^t \psi_X(u)du] \\ &= \exp[-\frac{h(x; \underset{\sim}{z})}{C(\underset{\sim}{z}_0)} \cdot C(\underset{\sim}{z}_0)\int_0^t \psi_X(u)du] \\ &= [S_X(t|x; \underset{\sim}{z})]^{h(x; \underset{\sim}{z})/C(\underset{\sim}{z}_0)}. \end{aligned} \tag{1.8}$$

Preservation thus implies that the future lifetime distribution is in the class of Lehmann's alternative distributions.

Which distributions should be used in X-analysis? It seems that many different positively skewed families of distributions can be fitted to survival data, especially if the age range is not too wide as is often the case in clinical trials. Traditionally, however, a Gompertz distribution with exponential hazard rate, $\mu_X(x) = Re^{\alpha x}$, has been used in analysis of human mortality data, and has proved to fit life tables well, at least for the older ages. It also has some convenient mathematical properties which make it an attractive model, as is shown in the next section.

2. GOMPERTZ DISTRIBUTION AND ITS PROPERTIES. A MODEL OF MORTALITY AND AGING

2.1. Definition. Preservation Property

The hazard rate of Gompertz distribution at age x is of *exponential* form

$$\mu_X(x) = Re^{\alpha x}, \quad R>0, \ \alpha>0, \ x>0, \tag{2.1}$$

and the survival distribution function is

$$S_X(x) = \exp[-\frac{R}{\alpha}(e^{\alpha x} - 1)]. \tag{2.2}$$

It is worthwhile noticing that (2.2) is, in fact, a truncated (at x=0) extreme value Type 1 limiting distribution.

The hazard rate of future lifetime distribution is

$$\lambda_T(t|x) = \mu_X(x+t) = Re^{\alpha(x+t)} = Re^{\alpha x}.e^{\alpha t}, \quad t>0, \tag{2.3}$$

and the future lifetime distribution is

$$\begin{aligned} S_T(t|x) &= S_X(x+t)/S_X(x) \\ &= \exp[-\frac{R}{\alpha}(e^{\alpha(x+t)} - 1)]/\exp[-\frac{R}{\alpha}(e^{\alpha x} - 1)] \\ &= \exp[-\frac{R}{\alpha}e^{\alpha x}(e^{\alpha t} - 1)]. \end{aligned} \tag{2.4}$$

Clearly, (2.4) is of the same form as (2.2); the preservation property holds, with $\psi_X(x) = e^{\alpha x}$.

As mentioned earlier, Gompertz distributions fit much human mortality data, especially for older ages, above age ξ, say. Letting $x = \xi + (x-\xi)$, we obtain

$$\mu_X(x|X>\xi) = Re^{\alpha x} = Re^{\alpha\xi}e^{\alpha(x-\xi)}, \tag{2.5}$$

and the truncated (at $x=\xi$) Gompertz SDF is

$$S_X(x|X>\xi) = \exp[-\frac{R}{\alpha}e^{\alpha\xi}(e^{\alpha(x-\xi)} - 1)]. \tag{2.6}$$

The future lifetime distribution is, of course, not affected by the truncation at $x = \xi$; it is identical with (2.4), provided $x > \xi$.

2.2. Approximate Relation to Logistic Model

Let

$$q(\tau|x) = \Pr\{x<X\leq x+\tau|X>x\} = 1 - \exp[-\frac{R}{\alpha}e^{\alpha x}(e^{\alpha\tau} - 1)] \tag{2.8}$$

denote the conditional probability of death within a fixed period τ, given alive at age x, and

$$p(\tau|x) = 1 - q(\tau|x).$$

(In actuarial notation, these correspond to ${}_\tau q_x$ and ${}_\tau p_x$, respectively.)

If R is very small and τ is not too large (1 year, say), then there is an approximate logistic linear relation over the period τ,

$$\log\frac{q(\tau|x)}{p(\tau|x)} = \log\frac{1 - \exp[-\frac{R}{\alpha}e^{\alpha x}(e^{\alpha\tau} - 1)]}{\exp[-\frac{R}{\alpha}e^{\alpha x}(e^{\alpha\tau} - 1)]} \doteq \gamma + \alpha^* x, \tag{2.9}$$

or

$$\exp[\frac{R}{\alpha}e^{\alpha x}(e^{\alpha\tau} - 1)] \doteq 1 + e^{\gamma+\alpha^* x}. \tag{2.10}$$

We now derive two approximations for γ and α^*.

Approximation 1. Taking logarithms of both sides of (2.10), we obtain

$$\frac{R}{\alpha}e^{\alpha x}(e^{\alpha\tau} - 1) \doteq \log\,(1 + e^{\gamma+\alpha^* x}).$$

We note that γ is usually negative and rather large (of order -3 to -12), while α^* is small (of order 0.05-0.10), so that $e^{\gamma+\alpha^* x} = y$ is small (<1). Using then expansion of $\log(1 + y) = y$, we obtain

$$\frac{R}{\alpha}e^{\alpha x}(e^{\alpha\tau} - 1) \doteq e^{\gamma+\alpha^* x},$$

or $Ae^{\alpha x} \doteq e^{\gamma+\alpha^* x}$, where $A = \frac{R}{\alpha}(e^{\alpha\tau} - 1)$. Hence

$$e^{\log A + \alpha x} \doteq e^{\gamma+\alpha^* x},$$

so that

$$\gamma \doteq \log A = \log[\frac{R}{\alpha}(e^{\alpha\tau} - 1)] \text{ and } \alpha^* \doteq \alpha. \tag{2.11}$$

Approximation 2. For small τ, we may write

$$q(\tau|x) \doteq \int_0^\tau \lambda_T(t|x)S_T(t|x)dt \doteq \tfrac{1}{2}\tau\lambda_T(\tau|x)S_T(\tau|x), \tag{2.12}$$

and we have

$$p(\tau|x) = S_T(\tau|x).$$

Hence

$$\frac{q(\tau|x)}{p(\tau|x)} \doteq \frac{\tfrac{1}{2}\tau\lambda_T(\tau|x)S_T(\tau|x)}{S_T(\tau|x)} = \tfrac{1}{2}\tau\lambda_T(\tau|x) = \tfrac{1}{2}\tau Re^{\alpha x}e^{\alpha\tau},$$

so that

$$\frac{q(\tau|x)}{p(\tau|x)} = \tfrac{1}{2}\tau Re^{\alpha(\tau+x)} \doteq \exp[\alpha\tau + \log(\tfrac{1}{2}\tau R) + \alpha x]$$

$$\doteq \exp(\gamma + \alpha^* x). \tag{2.13}$$

Hence

$$\gamma \doteq \alpha\tau + \log(\tfrac{1}{2}\tau R) \quad \text{and} \quad \alpha^* \doteq \alpha. \tag{2.14}$$

Estimating R and α from follow-up data, one can estimate the probability of death, $q(\tau|x)$, from (2.8) or approximate this probability by the traditional logistic prediction function

$$q(\tau|x) \doteq \exp(\gamma + \alpha^* x)[1 + \exp(\gamma + \alpha^* x)]^{-1}, \tag{2.15}$$

where γ and α^* are approximated by (2.11) or (2.14), provided that a Gompertz distribution fits the data over an appropriate range of age. In practice, this often requires a fairly large volume of the data in each age group, in order to observe enough deaths to reduce the standard errors of death rates and make these more reliable.

We now present two examples, to illustrate the techniques involved, and evaluate the applicability of the model to real situations.

EXAMPLE 2. Table 1 represents the U.S. Life Table (1969-71), White Males for ages $x \geq 35$. The force of mortality, at age $x+\frac{1}{2}$ was estimated from the approximate formula

$$\mu_{x+\frac{1}{2}} \doteq -\log p_x \doteq Re^{\alpha(x+\frac{1}{2})},$$

so that

$$y_{x+\frac{1}{2}} = \log(-\log p_x) \doteq \log R + \alpha(x+\tfrac{1}{2}). \qquad (2.16)$$

Using p_x's (for ages 35-90) from life table, we fitted by least squares a straight line yielding $R \doteq 0.00011405$ and $\alpha \doteq 0.085921$.

We then fitted the *tail* of Gompertz distribution for $x \geq 35$, that is, calculating the fitted tail, $S_X^{(1)}(x)$, of Gompertz SDF from the formula

$$S_X^{(1)}(x) = \frac{\ell_{35}}{\ell_0} \cdot \frac{S_X(x)}{S_X(35)}, \quad x \geq 35, \qquad (2.17)$$

where $S_X(x) = \exp[-\frac{R}{\alpha}(e^{\alpha x} - 1)]$ is a Gompertz survival function (Elandt-Johnson & Johnson (1980), Chapter 7). The fitted tail (column (7)) is shown in Figure 1.

The logistic parameter γ calculated from (2.11) gives $\gamma \doteq -9.0364$, and from (2.14) gives $\gamma \doteq -9.7343$. We also fitted a straight line, $y_x = \log \frac{q_x}{p_x} \doteq \gamma + \alpha^* x$, by least squares, using the same age range, 35-90. Here q_x and p_x are the life table values. We obtained $\hat{\gamma} \doteq -9.1131$, $\alpha^* \doteq 0.087605$. The points $y_x = \log \frac{q_x}{p_x}$, $x = 35,\ldots,90$

TABLE 1

U.S. LIFE TABLES (1969-71), WHITE MALES

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Age x to x+1	ℓ_x/ℓ_o	q_x	p_x	$-\log p_x = \mu_{x+\frac{1}{2}}$	$\log[-\log p_x]$	$S_x^{(1)}(x)$	$\log(q_x/p_x)$
35-36	.93843	.00217	.99783	.00217	-6.1321	.93843	-6.1309
36-37	.93639	.00235	.99765	.00235	-6.0521	.93617	-6.0510
37-38	.93420	.00256	.99744	.00256	-5.9666	.93371	-5.9652
38-39	.93181	.00281	.99719	.00281	-5.8731	.93105	-5.8718
39-40	.92919	.00310	.99690	.00310	-5.7747	.92815	-5.7733
40-41	.92631	.00340	.99660	.00341	-5.6822	.92500	-5.6806
41-42	.92316	.00372	.99628	.00373	-5.5922	.92158	-5.5903
42-43	.91973	.00409	.99591	.00410	-5.4973	.91787	-5.4951
43-44	.91596	.00452	.99548	.00453	-5.3970	.91384	-5.3947
44-45	.91182	.00501	.99499	.00502	-5.2939	.90948	-5.2913
45-46	.90725	.00555	.99445	.00556	-5.1913	.90474	-5.1884
46-47	.90221	.00612	.99388	.00614	-5.0931	.89961	-5.0901
47-48	.89669	.00673	.99327	.00675	-4.9978	.98405	-4.9944
48-49	.89066	.00739	.99261	.00742	-4.9040	.88803	-4.9002
49-50	.88408	.00812	.99188	.00815	-4.8094	.88151	-4.8053
50-51	.87690	.00892	.99108	.00896	-4.7150	.87447	-4.7105
51-52	.87908	.00980	.99020	.00985	-4.6205	.86686	-4.6155
52-53	.86056	.01081	.98919	.01087	-4.5218	.85864	-4.5164
53-54	.85126	.01194	.98806	.01201	-4.4218	.84977	-4.4159
54-55	.84110	.01318	.98682	.01323	-4.3255	.84021	-4.3158
55-56	.83001	.01452	.98548	.01463	-4.2250	.82991	-4.2176
56-57	.81796	.01594	.98406	.01607	-4.1309	.81884	-4.1229
57-58	.80492	.01745	.98255	.01760	-4.0396	.80773	-4.0308
58-59	.79088	.01906	.98094	.01924	-3.9506	.79417	-3.9409
59-60	.77580	.02077	.97923	.02099	-3.8638	.78048	-3.8533
60-61	.75969	.02258	.97742	.02284	-3.7793	.76584	-3.7679
61-62	.74253	.02451	.97549	.02482	-3.6963	.75019	-3.6939
62-63	.72433	.02657	.97343	.02693	-3.6146	.73350	-3.6010
63-64	.70509	.02879	.97121	.02921	-3.5332	.71574	-3.5185
64-65	.68479	.03120	.96880	.03170	-3.4515	.69687	-3.4356
65-66	.66343	.03386	.96614	.03445	-3.3684	.67688	-3.3511
66-67	.64097	.03674	.96326	.03743	-3.2852	.65574	-3.2665
67-68	.61742	.03977	.96023	.04058	-3.2044	.63346	-3.1841
68-69	.59286	.04284	.95716	.04378	-3.1285	.61004	-3.1065
69-70	.56746	.04597	.95403	.04706	-3.0563	.59965	-3.0327
70-71	.54138	.04916	.95084	.05041	-2.9876	.55991	-2.9623
71-72	.51476	.05262	.94738	.05406	-2.9178	.53325	-2.8906
72-73	.48768	.05655	.94345	.05821	-2.8437	.50567	-2.8144
73-74	.46010	.06118	.93882	.06313	-2.7625	.47723	-2.7308
74-75	.43195	.06647	.93353	.06878	-2.6768	.44772	-2.6422
75-76	.40324	.07231	.92769	.07506	-2.5050	.41829	-2.5517
76-77	.37408	.07843	.92157	.08168	-2.5050	.38810	-2.4639
77-78	.34474	.08472	.91528	.08852	-2.4245	.35768	-2.3799
78-79	.31553	.09103	.90897	.09544	-2.3492	.32724	-2.3011
79-80	.28681	.09749	.90251	.10258	-2.2772	.29701	-2.2254
80-81	.25885	.10466	.89534	.11055	-2.2023	.26724	-2.1465
81-82	.23176	.11273	.88727	.11961	-2.1236	.23818	-2.0632
82-83	.20563	.12127	.87873	.12928	-2.0458	.21010	-1.9805
83-84	.18069	.13012	.86982	.13947	-1.9699	.18326	-1.8998
84-85	.15718	.13942	.86058	.15015	-1.8962	.15790	-1.8201
85-86	.13527	.15033	.84967	.16291	-1.8146	.13424	-1.7320
86-87	.11493	.16321	.83679	.17818	-1.7250	.11263	-1.6345
87-88	.09618	.17666	.82334	.19439	-1.6379	.09276	-1.5391
88-89	.07919	.18947	.81053	.21007	-1.5603	.07519	-1.4535
89-90	.06418	.20145	.79855	.22496	-1.4918	.05981	-1.3773
90-91	.05125	.21344	.78656	.24009	-1.4268	.04660	-1.3043
91-92	.04031	.22684	.77316	.25727	-1.3576	.03551	-1.2262
92-93	.03117	.24152	.75848	.27644	-1.2858	.02641	-1.1444
93-94	.02364	.25767	.74233	.29796	-1.2108	.01912	-1.0581
94-95	.01755	.27426	.72574	.32056	-1.1377	.01345	-0.9731

U.S. Life Table (1969-71), White Males

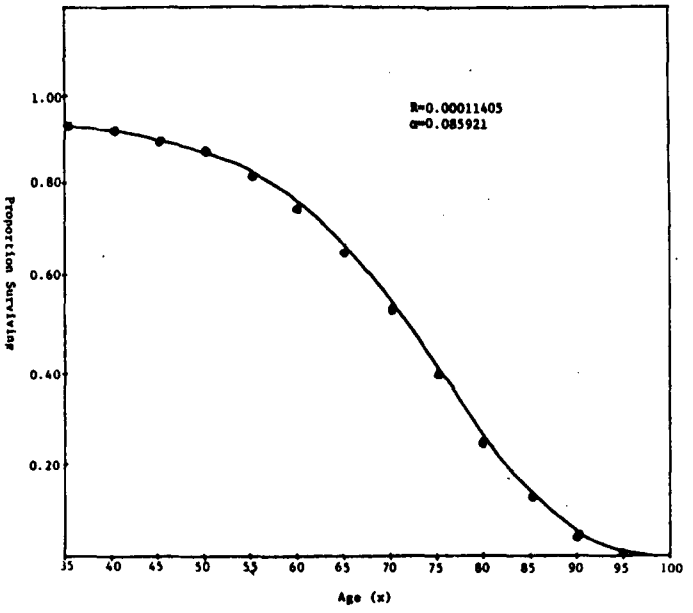

Fig. 1. Fitted tail of Gompertz distribution

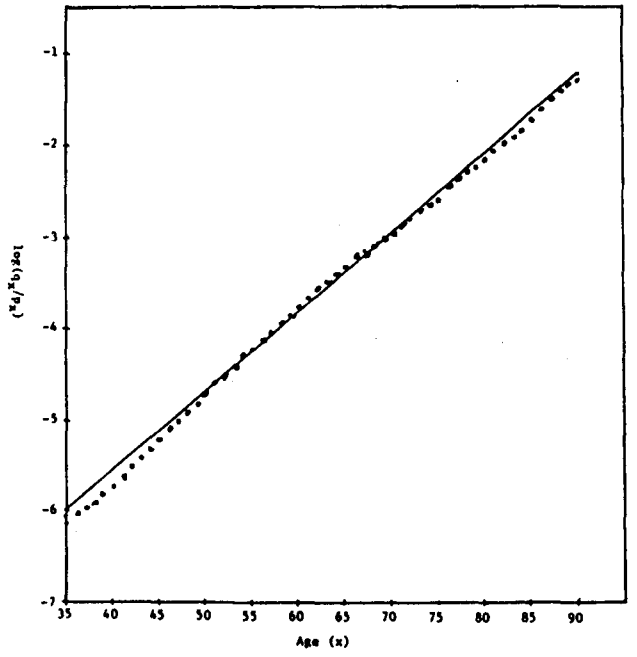

Fig. 2. Fitting logistic linear function

and the fitted line $y_x \doteq -9.1137 + 0.087605x$ are shown in Fig. 2.

The line $y_x \doteq -9.0364 + 0.085921x$ (not displayed in Fig. 2) is very close to the least squares fitted line. Approximation 1 is quite good, since R and τ are small. A still better fit can be obtained by fitting two Gompertz distributions piecewise.

3. EXTENDED GOMPERTZ DISTRIBUTION WITH CONCOMITANT VARIABLES

3.1. Interpretation of the Parameters

Assume for a moment that a Gompertz distribution fits the mortality data of a biological population over the whole lifespan. Note that at $x=0$, $\mu_X(0) = R\ (>0)$, which implies that at birth there is already a positive force of mortality, for each individual. It is not unreasonable to speculate that this is distinct for each member of a given population, and depends on the individual's genetic makeup; the population value can be regarded as a kind of an average of individual R's. It is worthwhile to notice that, in contrast, the power hazard rate of the Weibull distribution, $\mu_X(x) = \theta x^c$, is equal to zero at $x=0$. This might be not an unreasonable assumption for inanimate populations; in fact, the Weibull distribution is commonly used in reliability theory of industrial devices and systems. Some statisticians are very keen on using this distribution in survival analysis. Since this is a skew distribution, it also fits much mortality data (e.g. life table, see Elandt-Johnson & Johnson (1980), Chapter 7), but it is rather difficult to give a biological interpretation to its parameters. In my opinion, Weibull distribution should be avoided in analysis of mortality data.

The parameter α in the Gompertz distribution can be interpreted as a rate of using up vital resources, or as a "rate of living". Though it is specific for a given population - it is different, for example,

for males and females in the same ethnic population - it also depends on environmental factors. However, for members of the same population living in similar environmental conditions, it might be practically the same.

Let us represent the hazard rate function (2.1) in the form

$$\mu_X(x) = Re^{\alpha x} = e^{-\phi}e^{\alpha x} = e^{-(\phi-\alpha x)}, \qquad (3.1)$$

where ϕ (>0) may be interpreted as some initial level of "living capacity", while $R = e^{-\phi}$ may be thought of as an initial level of "unfitness" or toxicity". The bigger ϕ (or smaller R), the smaller the force of mortality for a given α, that is, the better chance of surviving. At age ξ (>0), say,

$$\mu_X(\xi) = e^{-(\phi-\alpha\xi)},$$

and then for $x = \xi + (x-\xi)$,

$$\mu_X(x) = \mu_X(x|X>\xi) = e^{-(\phi-\alpha\xi)}e^{\alpha(x-\xi)}, \qquad (3.2)$$

- that is, the "living resources" decrease as x increases.

Our interest is in determining these resources. This is a very difficult problem in cell biology and one may have some doubts whether the aging problem can ever be completely solved. At a less ambitious level, we may select a set of S specific observable characteristics ($\underset{\sim}{z}$), which are expected to make some contribution to the aging (mortality) process. In general, the z's might be functions of x, so that we consider a vector $\underset{\sim}{z}(x)' = (z_1(x),\ldots,z_S(x))$.

To be consistent with (3.1), we represent the hazard rate function in an exponential form

$$\mu_X(x;\ \underset{\sim}{z}(x)) = \exp[g(x;\ \underset{\sim}{z}(x))], \qquad (3.3)$$

where $g(x;\ \underset{\sim}{z}(x))$ is an unknown (negative) function of x and $z(x)$'s.

As a first approximation, we take a *linear* form for function $g(\cdot)$,

$$g(x;\ \underset{\sim}{z}(x)) = [\beta_0 + \beta_1 z_1(x) + \ldots + \beta_S z_S(x) + \alpha_0 x]$$

$$= \beta_0 + \underset{\sim}{\beta}'\underset{\sim}{z}(x) + \alpha_0 x, \tag{3.4}$$

so that

$$\mu_X(x;\ \underset{\sim}{z}(x)) = R_0 \exp[\underset{\sim}{\beta}'\underset{\sim}{z}(x)]e^{\alpha_0 x}, \tag{3.5}$$

where $R_0 = \exp(\beta_0)$.

This approximation is not unreasonable noticing that the z's need not represent direct measurements, but each can be a function of a specific observable characteristic such as, for example, logarithm or square root, etc.

If the z's are not constants (e.g. sex or race) then we still need to establish the relationship (if any) between the z's and x. Since the exponential hazard rate function (3.1) has proved to be appropriate for fitting many lifetime distributions, and its exponent is a linear function of x (see (3.5)), it must be that z_s is either a constant allowing, perhaps, for a random variation and using an average of several replications, or z_s is *linear* function of x, that is,

$$z_s(x) = z_s(0) + m_s x. \tag{3.6}$$

Hence

$$g(x;\ \underset{\sim}{z}(x)) = \{\beta_0 + \beta_1[z_1(0) + m_1 x] + \ldots + \beta_S[z_S(0) + m_S x] + \alpha_0 x\}$$

$$= \{\beta_0 + [\beta_1 z_1(0) + \ldots + \beta_S z_S(0)] + (\beta_1 m_1 + \ldots + \beta_S m_S + \alpha_0)x\}$$

$$= [\beta_0 + \underset{\sim}{\beta}'\underset{\sim}{z}(0) + \alpha x], \tag{3.7}$$

where

$$\alpha = \beta_1 m_1 + \ldots + \beta_S m_S + \alpha_0. \tag{3.8}$$

Substituting (3.7) into (3.3), we obtain

$$\mu_X(x;\ \underset{\sim}{z}(x)) = \exp[\beta_0 + \underset{\sim}{\beta}'\underset{\sim}{z}(0)]e^{\alpha x}$$

$$= R_0 \exp(\underset{\sim}{\beta}'\underset{\sim}{z}(0))e^{\alpha x} = R[\underset{\sim}{z}(0)]e^{\alpha x}, \tag{3.9}$$

where

$$R[\underset{\sim}{z}(0)] = R_0 \exp[\underset{\sim}{\beta}'\underset{\sim}{z}(0)] \text{ or } -\phi = \beta_0 + \underset{\sim}{\beta}'\underset{\sim}{z}(0). \qquad (3.10)$$

Of course, the z(0)'s (and so z(x)'s) are distinct for different individuals, because the "living capacity" for each individual in a given population is specific, determined by the individual's genetic composition.

The coefficients (β's) can be interpreted as contributions of the corresponding z's to the "vital material". If $\beta_s > 0$, the effect of z_s is harmful, and the characteristic z_s can be considered as a '*risk factor*'; if $\beta_s < 0$, the z_s is a prognostic factor, with a beneficial effect. The coefficient β_0 is almost always negative; it may be regarded, perhaps, as the genetic contribution specific for a given biological population and confounded with other factors, not controlled in a particular study. On the other hand, the term $\beta_s m_s$ depends on the change (slope m_s) in z_s, and on the contribution (β_s) of z_s to the aging process. When $\beta_s = 0$ or $m_s = 0$, $\beta_s m_s = 0$. The parameter α_0 accounts for the contribution of age, defined formally as time since day of birth; it may vary in different calendar periods.

Of course, it is not claimed that this explanation gives a complete and true picture of biological processes; but it seems that, at least, it is not inconsistent with observations.

3.2. Future Lifetime Distribution

We now derive the hazard rate for future lifetime distribution, using the measurements of z's and age x, the z(x)'s. We have

$$z_s(x) = z_s(0) + m_s x,$$

so that

$$\begin{aligned} z_s(x+t) &= z_s(0) + m_s(x+t) = z_s(0) + m_s x + m_s t \\ &= z_s(x) + m_s t. \end{aligned} \qquad (3.11)$$

Hence

$$\lambda_T(t|x;\ \underset{\sim}{z}(x)) = \mu_X(x+t) = R_0\exp[\underset{\sim}{\beta}'\underset{\sim}{z}(x+t)]e^{\alpha_0(x+t)}$$

$$= R_0\exp[\underset{\sim}{\beta}'\underset{\sim}{z}(x) + t(\underset{\sim}{\beta}'\underset{\sim}{m})]\exp[\alpha_0(x+t)]$$

$$= R_0\exp[\underset{\sim}{\beta}'\underset{\sim}{z}(x) + \alpha_0 x]\exp[(\underset{\sim}{\beta}'\underset{\sim}{m} + \alpha_0)t]$$

$$= R_0\exp[\underset{\sim}{\beta}'\underset{\sim}{z}(x) + \alpha_0 x]e^{\alpha t}, \qquad (3.12)$$

where $\alpha = \underset{\sim}{\beta}'\underset{\sim}{m} + \alpha_0$ is already defined in (3.8).

It is easy to see that the preservation property holds. The future lifetime distribution is

$$S_T(t|x;\ \underset{\sim}{z}(x)) = \exp\{-\frac{R_0}{\alpha}\exp[\underset{\sim}{\beta}'\underset{\sim}{z}(x) + \alpha_0 x](e^{\alpha t} - 1)\}. \qquad (3.13)$$

Also, we notice that (3.12) resembles a Cox's (1972) model, with the 'underlying' hazard rate $\lambda_0(t) = R_0e^{\alpha t}$.

It should be also noticed that if a Gompertz SDF is fitted for $x > \xi$, (3.12) and (3.13) are not affected by the truncation point, provided $x > \xi$.

Likelihood Function

Construction of the likelihood function is straightforward.

Let N be the number of individuals who were observed for some time during the investigation period. Suppose that for individual (j) the following data are available: age at entry, x_j; values of z's at entry, z_{sj}, s=1,2,...,S; and time last seen, t_j.

Let

$$\delta_j = \begin{cases} 1 \text{ if individual (j) died at time } t_j, \\ 0 \text{ otherwise.} \end{cases}$$

The likelihood function is

$$\prod_{j=1}^{N}[\lambda_T(t_j|x_j;\ \underset{\sim}{z}_j)]^{\delta_j}S_T(t_j|x_j;\ \underset{\sim}{z}_j), \qquad (3.14)$$

where $\underset{\sim}{z}_j' = (z_{1j}, \ldots, z_{sj})$, and $\lambda_T(\cdot)$ and $S_T(\cdot)$ are defined in (3.12) and (3.13), respectively.

3.3. Relation to Logistic Linear Model

There is also an approximate relationship between the expanded Gompertz model and a logistic linear model. For the time period τ, we have

$$\log \frac{q(\tau|x;\, \underset{\sim}{z}(x))}{p(\tau|x;\, \underset{\sim}{z}(x))} = \log \frac{1 - \exp\{-\frac{R_0}{\alpha}\exp[\alpha_0 x + \underset{\sim}{\beta}'\underset{\sim}{z}(x)](e^{\alpha t} - 1)\}}{\exp\{-\frac{R_0}{\alpha}\exp[\alpha_0 x + \underset{\sim}{\beta}'\underset{\sim}{z}(x)](e^{\alpha\tau} - 1)\}}$$

$$\doteq \gamma + \alpha^* x + \beta_1^* z_1(x) + \ldots + \beta_S^* z_S(x). \qquad (3.15)$$

Using the same techniques as in Section 2, we obtain

$$\gamma \doteq \log[\frac{R_0}{\alpha}(e^{\alpha\tau} - 1)] \quad \text{(approximation 1);}$$

or

$$\gamma \doteq \alpha\tau + \log(R_0\tau) \quad \text{(approximation 2);} \qquad (3.16)$$

and

$$\alpha_0^* \doteq \alpha, \quad \beta_s^* \doteq \beta_s, \quad s=1,2,\ldots,S.$$

Of course, care should be taken, when using these approximations (R_0 and τ should be very small). Perhaps, it would be safer to fit multiple regression of observed values of $\log[q(\tau|x)/p(\tau|x)]$ on x and z's.

3.4. Fitting 'Piecewise' Gompertz SDF

As mentioned at the end of Section 2, we can obtain a better fit if, for example, two 'pieces', each representing a Gompertz distribution with different parameters, are used. This corresponds to assuming that the slope(s) of the linear function of z's on x has (have) been changed at a certain age, which may be associated with different phases of living processes, as exemplified below.

EXAMPLE 3. For simplicity, consider only a single concomitant, age

dependent covariable $z(x)$. Suppose that we fit two Gompertz distributions: the first 'piece' for ages $\xi_0 \le x < \xi_1$, and the second 'piece' for $x \ge \xi_1$.

(i) For $\xi_0 \le x < \xi_1$, we have $x = \xi_0 + (x-\xi_0)$, and

$$z(x) = z(\xi_0) + m_1(x-\xi_0), \tag{3.17}$$

so that

$$\begin{aligned}\mu_X(x|\, x>\xi_0) &= R_0\exp[\beta z(x) + \alpha_0 x] \\ &= R_0\exp\{[\beta z(\xi_0) + m_1(x-\xi_0)] + \alpha_0[\xi_0 + (x-\xi_0)]\} \\ &= R_0\exp[\beta z(\xi_0) + \alpha_0\xi_0]\exp[(\beta m_1 + \alpha_0)(x-\xi_0)] \\ &= R_1\exp[\alpha_1(x-\xi_0)], \end{aligned} \tag{3.18}$$

where

$$R_1 = R_0\exp[\beta z(\xi_0) + \alpha_0\xi_0] \quad \text{and} \quad \alpha_1 = \beta m_1 + \alpha_0. \tag{3.19}$$

(ii) Suppose that for $x \ge \xi_1$, we have

$$z(x) = z(\xi_1) + m_2(x-\xi_1), \; m_2 \ne m_1. \tag{3.20}$$

By the same argument as in (i), we obtain

$$\begin{aligned}\mu_X(x|\, x>\xi_1) &= R_0\exp[\beta z(\xi_1) + \alpha_0\xi_1]\exp[(\beta m_2 + \alpha_0)(x-\xi_1)] \\ &= R_2\exp[\alpha_2(x-\xi_1)], \end{aligned} \tag{3.21}$$

where

$$R_2 = R_0\exp[\beta z(\xi_1) + \alpha_0\xi_1] \quad \text{and} \quad \alpha_2 = \beta m_2 + \alpha_0 . \tag{3.22}$$

There is no difficulty in incorporating more than one covariables in the hazard rate model, allowing also for linear change of some z's with x. Also, fitting more than two Gompertz distributions is straightforward (for the general method, see Elandt-Johnson & Johnson (1980), Chapter 7).

EXAMPLE 4. *(Age dependent disease)*

Suppose that a chronic disease (e.g. cancer, diabetes) develops when an age dependent factor, z, reaches a threshold value at age x and changes the slope of the linear relation with age. The onset of such disease is age specific for a given individual. This is a situation analogous to that in Example 3, replacing the fixed value ξ_1 by x, and x by x+t, say. Therefore, for an individual who is free of the disease at age x, the hazard rate function is

$$\mu_X^{(1)}(x+t|X>x;\ z(x)) = \lambda_T^{(1)}(t|x;\ z(x))$$

$$= R_0\exp[\beta z(x) + \alpha_0 x]\exp[(\beta m_1 + \alpha_0)t]$$

$$R_0\exp[\beta z(x) + \alpha_0 x]e^{\alpha_1 t}, \qquad (3.23)$$

where

$$\alpha_1 = \beta m_1 + \alpha_0 ,$$

while for an individual who experienced onset t years ago at age x, the hazard rate function is

$$\mu_X^{(2)}(x+t|X>x;\ z(x)) = R_0\exp[\beta z(x) + \alpha_0 x]e^{\alpha_2 t}, \qquad (3.24)$$

where

$$\alpha_2 = \beta m_2 + \alpha_0 .$$

4. SEVERAL CAUSES OF DEATH: COMPETING RISK ANALYSIS

4.1. Exponential 'Crude' Hazard Rates

In epidemiological studies and clinical trials, the main concern is often whether a particular characteristic represents a risk factor associated with a particular cause of death such as, for example, cancer or heart disease. We again consider a vector of characteristics, $\underset{\sim}{z}(x)$, and assume

$$z_s(x) = z_s(0) + m_s x, \tag{4.1}$$

$s=1,2,\ldots,S$ (with, perhaps, some $m_s = 0$).

Consider K causes, $C_1, C_2, \ldots, C_K$, of death. Suppose that over a specified age range, we may assume an *exponential* 'crude' hazard rate, for cause C_k

$$\mu_k(x; \underset{\sim}{z}(x)) = R_{0k}\exp[\underset{\sim}{\beta}_k'\underset{\sim}{z}(x)]e^{\alpha_0 x}$$

$$= R_{0k}\exp[\underset{\sim}{\beta}_k'\underset{\sim}{z}(0)]e^{\alpha_k x} = R_k e^{\alpha_k x} \tag{4.2}$$

where

$$\alpha_k = \alpha_0 + \beta_{k1}m_1 + \beta_{k2}m_2 + \ldots + \beta_{kS}m_S \quad \text{and} \quad R_k = R_{0k}\exp[\underset{\sim}{\beta}_k'\underset{\sim}{z}(0)] \tag{4.3}$$

for $k=1,2,\ldots,K$.

Note that the relation (4.1) does not depend on the cause; however the contribution, β_{sk} for characteristic z_s might be cause specific. In particular, if z_s is a high risk factor for cause C_k, then β_{sk} should be rather large and, perhaps, not so large for other causes.

The future lifetime hazard rate function for cause C_k is

$$\lambda_k(t|x; \underset{\sim}{z}(x)) = \mu_k(x+t; \underset{\sim}{z}(x+t)) = R_{0k}\exp[\underset{\sim}{\beta}_k'\underset{\sim}{z}(x) + \alpha_0 x]e^{\alpha_k t} \tag{4.4}$$

for $k=1,2,\ldots,K$.

The overall (for all causes) force of mortality is

$$\mu_X(x|\underset{\sim}{z}(x)) = \sum_{k=1}^{K}\mu_k(x|\underset{\sim}{z}(x)) = \sum_{k=1}^{K} R_{0k}\exp[\underset{\sim}{\beta}'_k\underset{\sim}{z}(0)]\, e^{\alpha_k x}, \tag{4.5}$$

and the overall survival distribution function is

$$S_X(x; \underset{\sim}{z}(x)) = \prod_{k=1}^{K}\exp\{-\frac{R_{0k}}{\alpha_k}\exp[\underset{\sim}{\beta}_k'\underset{\sim}{z}(0)](e^{\alpha_k x} - 1)\}. \tag{4.6}$$

The corresponding future lifetime overall hazard rate function and SDF are

$$\lambda_T(t|x; \underset{\sim}{z}(x)) = \sum_{k=1}^{K}\lambda_k(t|x; \underset{\sim}{z}(x)) = \sum_{k=1}^{K} R_{0k}\exp[\underset{\sim}{\beta}_k'\underset{\sim}{z}(x) + \alpha_0 x]e^{\alpha_k t}, \tag{4.7}$$

and

$$S_T(t|x;\ \underset{\sim}{z}(x)) = \prod_{k=1}^{K} \exp\{-\frac{R_{0k}}{\alpha_k} \exp[\underset{\sim}{\beta}_k'\underset{\sim}{z}(x) + \alpha_0 x](e^{\alpha_k t} - 1)\}, \tag{4.8}$$

respectively.

Likelihood Function

The parameters of the distribution can be estimated from the *future lifetime* joint likelihood function (from all causes).

Let N_k denote the number of deaths from cause C_k, k=1,2,...,K, and N_{K+1} - the number of survivors over a specified period of investigation, with

$$N_1 + N_2 + \ldots + N_K + N_{K+1} = N.$$

Further, let t_j denote the time at which individual (j) was last seen; x_j - age at entry; $\underset{\sim}{z}_j' = (z_{1j},\ldots,z_{Sj})$ be the vector of observed values of z's at age x_j, and

$$\delta_{kj} = \begin{cases} 1 \text{ if individual (j) died at time } t_j \text{ from} \\ \quad \text{cause } C_k \\ 0 \text{ otherwise.} \end{cases}$$

The likelihood function is

$$\prod_{j=1}^{N} \prod_{k=1}^{K} [\lambda_k(t_j|x_j;\ \underset{\sim}{z}_j)]^{\delta_{kj}} S_T(t|x_j;\ \underset{\sim}{z}_j). \tag{4.9}$$

4.2. Relation to the Logistic Function

There is also an approximate relation between the multi-cause Gompertz SDF and a multi-response logistic function.

Let $q_k^*(\tau|x;\ z(x))$ denote the conditional probability of death from cause C_k over the period x to x+τ given alive at age x. We have (for small τ)

$$q_k^*(\tau|x;\ \underset{\sim}{z}(x)) = \int_0^{\tau} \lambda_k(t|x;\ \underset{\sim}{z}(x)) S_T(t|x;\ \underset{\sim}{z}(x))dt \tag{4.10}$$

$$\doteq \tfrac{1}{2}\tau\lambda_k(\tau|x;\ \underset{\sim}{z}(x)) S_T(\tau|x;\ \underset{\sim}{z}(x)) \quad , \tag{4.10a}$$

and the conditional probability of surviving the period x to $x+\tau$ given alive at age x is

$$p(\tau|x;\ \underset{\sim}{z}(x)) = S_T(\tau|x;\ \underset{\sim}{z}(x)). \tag{4.11}$$

Hence for the exponential hazard rate model, we have

$$\frac{q_k^*(\tau|x;\ \underset{\sim}{z}(x))}{p(\tau|x;\ \underset{\sim}{z}(x))} \doteq \tfrac{1}{2}\tau\lambda_k(\tau|x;\ \underset{\sim}{z}(x)) = \tfrac{1}{2}\tau R_{0k}\exp[\underset{\sim}{\beta}_k'\underset{\sim}{z}(x)]e^{\alpha_k\tau}$$

$$= \tfrac{1}{2}\tau R_{0k}\exp[\alpha_0 x + \beta_{k1}z_1(x) + \ldots + \beta_{ks}z_s(x)]e^{\alpha_k\tau}. \tag{4.12}$$

The multi-response logistic linear relation is defined by

$$\log\frac{q_k^*(\tau|x;\ \underset{\sim}{z}(x))}{p(\tau|x;\ \underset{\sim}{z}(x))} \doteq \gamma_k + \alpha_0^* x + \beta_{k1}^* z_1(x) + \ldots + \beta_{ks}^* z_s(x). \tag{4.13}$$

From (4.11) and (4.12), we obtain

$$[\alpha_k\tau + \log(\tfrac{1}{2}\tau R_{0k})] + \alpha_0 x + \beta_{k1}z_1(x) + \ldots + \beta_{ks}z_s(x)$$

$$\doteq \gamma_k + \alpha_0^* x + \beta_{k1}^* z_1(x) + \ldots + \beta_{ks}^* z_s(x). \tag{4.14}$$

Hence

$$\gamma_k \doteq \alpha_k\tau + \log(\tfrac{1}{2}\tau R_{0k}),\quad \alpha_0^* \doteq \alpha \text{ and } \beta_{ks}^* \doteq \beta_{ks} \tag{4.15}$$

for $k=1,2,\ldots,K$ and $s=1,2,\ldots,S$.

(compare (3.16)).

Note that the approximations are fair when τ is small (≤ 1, say) and R_{0k} very small (of order 10^{-4} or 10^{-3}).

EXAMPLE 5. Life table functions presented in Table 2, were calculated for U.S. Census population, 1970, White Males. We distinguish three causes of death: (1) - Malignant neoplasms; (2) - Diseases of circulatory system; (3) - Others. For $x=30, 35, \ldots, 85$, we use the approximation $\mu_k(x+\frac{1}{2}) \doteq {}_5m_{kx}$, where ${}_5m_{kx}$ is the central quinquennial death rate for

TABLE 2

MULTIPLE CAUSES. U.S. CENSUS POPULATION, 1970, WHITE MALES

Age Group x to x+5	Total Population				Malignant Neoplasms (C_1)			Circulatory System (C_2)			Others (C_3)		
	${}_5m_x$	${}_5q_x$	${}_5p_x$	$\log({}_5q_x/{}_5p_x)$	${}_5m_{1x}$	${}_5q^*_{1x}$	$\log({}_5q^*_{1x}/{}_5p_x)$	${}_5m_{2x}$	${}_5q^*_{2x}$	$\log({}_5q^*_{2x}/{}_5p_x)$	${}_5m_{3x}$	${}_5q^*_{3x}$	$\log({}_5q^*_{3x}/{}_5p_x)$
30-35	.001854	.009228	.990772	-4.6763	.000191	.000951	-6.9487	.000244	.001215	-6.7038	.001419	.007062	-4.9438
35-40	.002604	.012936	.987064	-4.3347	.000366	.001670	-6.3822	.000649	.003223	-5.7245	.001619	.008044	-4.8098
40-45	.004200	.020800	.979200	-3.8518	.000653	.003234	-5.7132	.001511	.007481	-4.8743	.002036	.010085	-4.5757
45-50	.006846	.033692	.966308	-3.3562	.001229	.006050	-5.0734	.003008	.014806	-4.1784	.002608	.012836	-4.3212
50-55	.010986	.053564	.946436	-2.8718	.002254	.010990	-4.4558	.005348	.026072	-3.5918	.003384	.016501	-4.0493
55-60	.017746	.085171	.914829	-2.3741	.003974	.019074	-3.8704	.009013	.043256	-3.0516	.004759	.022841	-3.6902
60-65	.027084	.127186	.872814	-1.9261	.006170	.028976	-3.4052	.014357	.067423	-2.5607	.006556	.030787	-3.3446
65-70	.040461	.184230	.815770	-1.4880	.008793	.040037	-3.0143	.022557	.102708	-2.0722	.009111	.041484	-2.9788
70-75	.058280	.254862	.745138	-1.0728	.011538	.050456	-2.6925	.034247	.149763	-1.6045	.012496	.054644	-2.6127
75-80	.086934	.356558	.643442	-0.5903	.014933	.061247	-2.3519	.053947	.221263	-1.0675	.018054	.074048	-2.1621
80-85	.126068	.474509	.525491	-0.3878	.017702	.066628	-2.0652	.083500	.314288	-0.5140	.024866	.093593	-1.7254
85+	.185517	1.000000	.000000	-	.017722	.095529	-	.131424	.708420	-	.036371	.196050	-

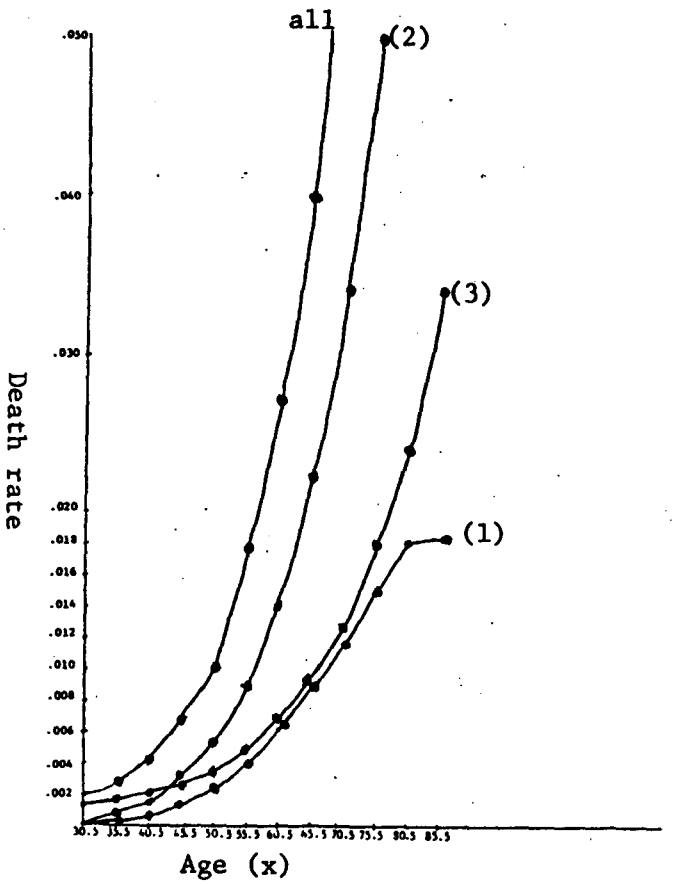

Fig. 3. Estimated hazard rates for different causes

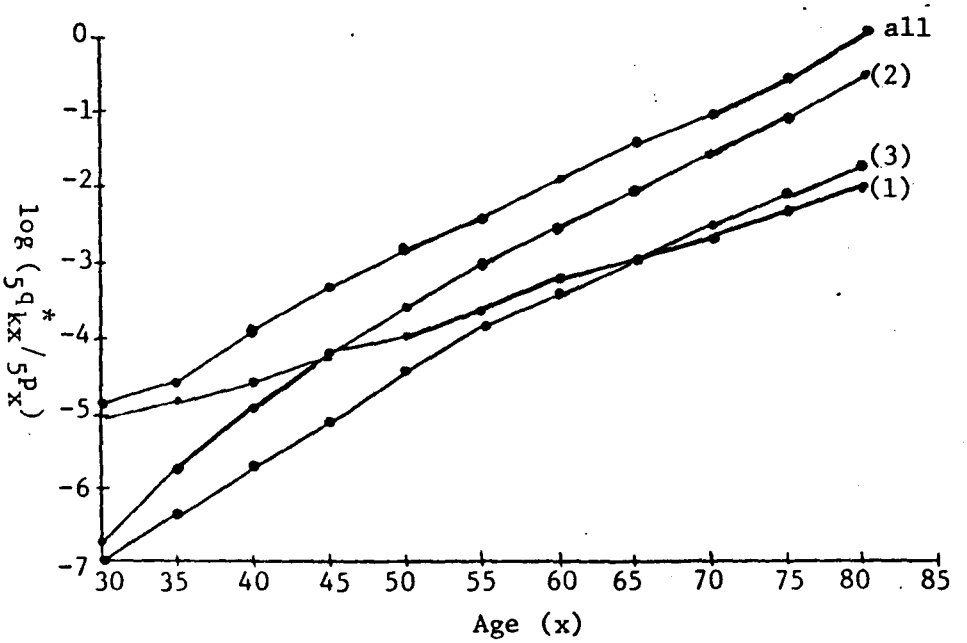

Fig. 4. Logistic relations for different causes

cause C_k. The $\mu_k(\cdot)$ for k=1,2,3, and the $\mu_x(\cdot)$ for all causes are represented in Fig. 3. The observed $\log({}_5q^*_{kx}/{}_5p_x)$ are shown in Fig. 4. As we can see, better fit, for each cause, can be obtained for ages $x > 50$.

It is clear from Figs. 3 and 4 that the majority of deaths over age 40 are due to diseases of circulatory system.

5. APPLICATION TO FOLLOW-UP STUDIES

One purpose of epidemiological prospective studies is to find out which characteristics ($\underline{z}$) represent *prognostic factors* for increased (decreased) mortality from a specific cause. Traditionally, epidemiologists use a logistic linear function for response, which is death from this cause. This method is inappropriate for various reasons; for example, in studies with fixed termination time, the recruitments sometimes takes a long period (1-3 years) and the participants do not represent a cohort, so that the logistic prediction probability function is underestimated. This fact is now well recognized. More appropriate approximation for this probability is obtained from Cox's (1972) approach. Often the partial likelihood function is used to estimate α_0 and β's, while $\lambda_0(t)$ is unspecified. A fair amount of information may be lost, however, with this approach. Another possibility would be to use a Gompertz future lifetime survival distribution with regard to specific cause, defined in (4.1), with $\lambda_k(\cdot)$ defined in (4.3), provided that the data are extensive enough to obtain reasonable standard errors. Technical details on this topic, and critical evaluation of various approximations to the prediction probability will be discussed elsewhere.

6. NON-GOMPERTZIAN POPULATIONS WITH EXPONENTIAL HAZARD RATE FUNCTION

So far we have been concerned with exponential hazard rate models

with time dependent concomitant variables and no interaction among the z's, and between the z's and x. In these cases, the original Gompertz distribution, $S_X(\cdot)$, as well as the future lifetime distribution, $S_T(\cdot)$, have the same α.

We now consider situations in which hazard rate function is still of *exponential* form, but the relation between z and x is *non linear* or there is an *interaction* term between z and x. To show the consequences of such assumptions, we use two examples. For simplicity, we shall consider only a single covariable z.

EXAMPLE 6. *(Quadratic relation of z with x)*

(a) Let

$$z(x) = z(0) + k_1 x + k_2 x^2, \tag{6.1}$$

so that

$$\begin{aligned} z(x+t) &= z(0) + k_1(x+t) + k_2(x+t)^2 \\ &= [z(0) + k_1 x + k_2 x^2] + (k_1 + 2k_2 x)t + k_2 t^2 \\ &= z(x) + (k_1 + 2k_2 x)t + k_2 t^2. \end{aligned} \tag{6.2}$$

The exponential force of mortality is

$$\begin{aligned} \mu_X(x;\, z(x)) &= R_0 \exp[\beta z(x) + \alpha_0 x] \\ &= R_0 \exp[\beta z(0)] \exp[(\beta k_1 + \alpha_0)x + \beta k_2 x^2] \\ &= R_0 \exp[\beta z(0)] \exp(\alpha_1 x + \alpha_2 x^2) \end{aligned} \tag{6.3}$$

or

$$\mu_X(\cdot) = R \exp(\alpha_1 x + \alpha_2 x^2), \tag{6.3a}$$

where

$$\alpha_1 = \beta k_1 + \alpha_0, \quad \alpha_2 = \beta k_2 \quad \text{and} \quad R = R_0 \exp[\beta z(0)]. \tag{6.4}$$

The hazard rate for future lifetime is

$$\lambda_T(t|x;\ z(x)) = \mu(x+t;\ z(x+t)|X>x)$$

$$= R_0\exp[\beta\cdot z(x+t) + \alpha_0(x+t)]$$

$$= R_0\exp\{\beta[z(x) + (k_1+2k_2x)t + k_2t^2] + (\alpha_0x + \alpha_0t)\}$$

$$= R_0\exp[\beta z(x) + \alpha_0x]\exp[(\beta k_1+\alpha_0)t + 2\beta k_2xt + \beta k_2t^2]$$

$$= R_0\exp[\beta z(x) + \alpha_0x]\exp(\alpha_1t + 2\alpha_2xt + \alpha_2t^2), \qquad (6.5)$$

or

$$\lambda_T(\cdot) = R_0\exp[\beta z(0) + \alpha_1x + \alpha_2x^2)]\exp(\alpha_1t + 2\alpha_2xt + \alpha_2t^2). \qquad (6.5a)$$

Note that the preservation property no longer holds; the second factor in (6.5) includes the *interaction* term, $2\alpha_2xt$. Also, (6.5) does not represent a Cox's proportional hazard rate model.

(b) Suppose now that the accelerated change of z with x, expressed by the quadratic relation, is associated with contracting a certain age dependent disease. Suppose, then that up to age of onset, x, the relation is linear

$$z(x) = z(0) + mx, \qquad (6.6)$$

and after onset, it becomes quadratic

$$z(x+t) = z(x) + k_1t + k_2t^2$$

$$= [z(0) + mx] + k_1t + k_2t^2. \qquad (6.7)$$

In this case, the future lifetime hazard rate function is

$$\lambda_T(t|x;\ \underset{\sim}{z}(x)) = \mu_X(x+t;\ z(x+t)|X>x) = R_0\exp[\beta\cdot z(x+t) + \alpha_0(x+t)]$$

$$= R_0\exp[\beta z(x) + \alpha_0x]\exp(\alpha_1t + \alpha_2t^2), \qquad (6.8)$$

or

$$\lambda_T(\cdot) = R_0\exp[\beta z(0) + \alpha_3x]\exp(\alpha_1t + \alpha_2t^2), \qquad (6.8a)$$

where

$$\alpha_1 = \beta k_1 + \alpha_0, \quad \alpha_2 = \beta k_2 \quad \text{and} \quad \alpha_3 = \beta m + \alpha_0. \tag{6.9}$$

Now, there is no interaction term; the hazard rate (6.8) belongs to the class of Cox's models with specified $\lambda_0(t) = R_0 \exp(\alpha_1 t + \alpha_2 t^2)$.

EXAMPLE 7. *(Linear relation of z with x, including interaction between z and x)*

We now consider an exponential hazard rate model when z is a *linear* function of x, but there is an *interaction* term of z with x, expressed as product δzx.

(a) We have

$$z(x) = z(0) + mx, \tag{6.10}$$

and

$$z(x+t) = z(0) + m(x+t) = z(x) + mt. \tag{6.11}$$

Suppose that

$$\begin{aligned} \mu_x(x;\, z(x)) &= R_0 \exp[\beta z(x) + \alpha_0 x + \delta z(x) \cdot x] \\ &= R_0 \exp\{\beta[z(0) + mx] + \alpha_0 x + \delta[z(0) + mx]x\} \\ &= R_0 \exp[\beta z(0) + (\beta m + \alpha_0)x + \delta z(0)x + \delta m x^2] \\ &= R_0 \exp[\beta z(0)] \exp\{[\alpha_1 + \delta z(0)]x + \alpha_2 x^2\}, \end{aligned} \tag{6.12}$$

where

$$\alpha_1 = \beta m + \alpha_0, \quad \alpha_2 = \delta m. \tag{6.13}$$

Formula (6.12) resembles (6.3), except for the interaction term, $\delta z(0)x$.

For future lifetime, we have

$$\begin{aligned} \lambda_T(t \,|\, x;\, z(x)) &= R_0 \exp[\beta \cdot z(x+t) + \alpha_0(x+t) + \delta \cdot z(x+t) \cdot (x+t)] \\ &= R_0 \exp\{\beta[z(x) + mt] + [\alpha_0 + \delta(z(x) + mt)](x+t)\} \end{aligned}$$

$$= R_0\exp[\beta z(x) + \alpha_0 x + \delta z(x)x]\exp\{(\beta m + \alpha_0)t + \delta[mx + z(x)]t + \delta mt^2\}$$

$$= R_0\exp\{\beta z(x) + [\alpha_0 + \delta z(x)]x\}\exp\{\alpha_1 t + [\alpha_2 x + \delta z(x)]t + \alpha_2 t^2\}, \tag{6.14}$$

or

$$\lambda_T(\cdot) = R_0\exp[(\beta + \delta x)z(0) + \alpha_1 x + \alpha_2 x^2]\exp[(\delta z(0) + \alpha_1 + 2\alpha_2 x)t + \alpha_2 t^2]. \tag{6.14a}$$

(b) Now, suppose that interaction occurs only *after* age x. Then

$$\lambda_T(t|x;\ z(x)) = \mu_X(x+t;\ z(x+t)) = R_0\exp[\beta\cdot z(x+t) + \alpha_0(x+t) + \delta\cdot z(x+t)t]$$

$$= R_0\exp\{\beta[z(x) + mt] + \alpha_0 x + \alpha_0 t + \delta[z(x) + mt]t\}$$

$$= R_0\exp[\beta z(x) + \alpha_0 x]\exp[(\beta m + \alpha_0)t + \delta z(x)t + \delta mt^2]$$

$$= R_0\exp[\beta z(x) + \alpha_0 x]\exp[\alpha_1 t + \delta z(x)t + \alpha_2 t^2], \tag{6.15}$$

or

$$\lambda_T(\cdot) = R_0\exp\{\beta[z(0) + mx] + \alpha_0 x\}\exp\{\alpha_1 t + \delta[z(0) + mx]t + \alpha_2 t^2\}$$

$$= R_0\exp[\beta z(0) + \alpha_1 x]\exp[(\delta z(0) + \alpha_1)t + \alpha_2 xt + \alpha_2 t^2]. \tag{6.15a}$$

In this case, the interaction term is present only in the second factor of (6.15).

Comparing model (6.5) (Example 6) and model (6.14) (Example 7), we notice that both include quadratic terms in t; both may even fit well the same data. To distinguish between them, one may study the function

$$u(t|x) = z(x+t) - z(x). \tag{6.16}$$

(a) For situation (a), the quadratic relation, (Example 6(a)) leads to

$$u(t|x) = (k_1 + 2k_2 x)t + k_2 t^2 \tag{6.17}$$

which depends on age x. For different given ages (or age groups)

the coefficient of t in (6.17) will be different.

On the other hand, for the linear model with interaction (Example 7(a)), we have

$$u(t|x) = mt, \tag{6.18}$$

which does not depend on x.

(b) Comparing these models for situations discussed in (b), we have for the quadratic model

$$u(t|x) = k_1 t + k_2 t^2, \tag{6.19}$$

while for the linear model with interaction,

$$u(t|x) = mt.$$

Repeated Measurements

In order to study these relationships empirically, it is essential to obtain *repeated measurements* on the covariables z's. For each characteristic, z_s, we try to estimate the function $u_s(t|x)$, s=1,2,...,S and then use these functions in constructing an appropriate hazard function, $\lambda_T(t|x; \underset{\sim}{z}(x))$.

7. COMPARATIVE TRIALS

There is a large number of experimental designs for comparisons of two or more populations. Here we will consider only a few examples, to illustrate some basic methodological techniques.

7.1. Two Demographic Populations

Suppose that we wish to compare survivorship in two populations, (1) and (2), say. For example, the two populations may represent males and females or, in general, two *demographic* populations.

The simplest way would be to evaluate a survival function separately for each population. Mathematically more elegant results might be obtained by introducing concomitant variable(s) z.

EXAMPLE 8. *(Indicator variable)*

Let

$$z = \begin{cases} 1 \text{ for population (1)} \\ 0 \text{ for population (2).} \end{cases} \tag{7.1}$$

In general, there might be an *interaction* between z and x, so that the force of mortality can be written in the form

$$\mu_X(x;\ z) = R_0 \exp(\beta z + \alpha_0 x + \delta z x) = R_0 e^{\beta z} e^{(\alpha_0+\delta z)x}, \tag{7.2}$$

and the future lifetime hazard rate is

$$\begin{aligned} \lambda_T(t|x;\ z) = \mu_X(x+t|x;\ z) &= R_0 e^{\beta z} \exp[(\alpha_0 + \delta z)(x+t)] \\ &= R_0 \exp[\beta z + (\alpha_0 + \delta z)x] e^{(\alpha_0+\delta z)t}. \end{aligned} \tag{7.3}$$

Noticing that z takes only the values 0 or 1, we may write (7.2) in the form of two hazard rates, one for each population.

$$\mu_X^{(1)}(x;\ z=1) = R_1 e^{\alpha_1 x}, \tag{7.4.i}$$

where

$$R_1 = R_0 e^{\beta} \quad \text{and} \quad \alpha_1 = \alpha_0 + \delta,$$

and

$$\mu^{(2)}(x;\ z=0) = R_0 e^{\alpha_0 x}. \tag{7.4.ii}$$

Similarly, for the future lifetime hazard rates, we have

$$\lambda_T^{(1)}(t|x;\ z=1) = R_1 e^{\alpha_1 x} e^{\alpha_1 t}, \tag{7.5.i}$$

and

$$\lambda_T^{(2)}(t|x;\ z=0) = R_0 e^{\alpha_0 x} e^{\alpha_0 t}. \tag{7.5.ii}$$

EXAMPLE 9. *(Linear relation of z with x, with a different slope for each population)*

In Example 8, two different hazard rates $\mu_X^{(1)}(\cdot)$ and $\mu_2^{(2)}(\cdot)$ for populations (1) and (2), respectively, were obtained by introducing a "population effect" (βz), and an interaction term ($\delta z x$), without interpreting their meaning. The same general result can be obtained

by introducing a *continuous* variable z, and assuming that z is a *linear* function of x, but with a different slope in each population.

$$\text{for population (1):} \quad z^{(1)}(x) = z^{(1)}(0) + m_1 x, \tag{8.6.i}$$

$$\text{for population (2):} \quad z^{(2)}(x) = z^{(2)}(0) + m_2 x, \tag{8.6.ii}$$

Then

$$\begin{aligned} \mu_X^{(1)}(x) &= R_0 \exp[\beta z^{(1)}(x) + \alpha_0 x] \\ &= R_0 \exp\{\beta[z^{(1)}(0) + m_1 x] + \alpha_0 x\} \\ &= R_0 \exp[\beta z^{(1)}(0) + (\beta m_1 + \alpha_0)x] \\ &= R_0 \exp[\beta z^{(1)}(0) + \alpha_1 x] = R_1 e^{\alpha_1 x}, \end{aligned} \tag{7.6.i}$$

where

$$\alpha_1 = \beta m_1 + \alpha_0 \quad \text{and} \quad R_1 = R_0 \exp(\beta z^{(1)}(0)).$$

(Compare (7.6.i) with (7.4.i).)

In a similar way, we obtain

$$\mu_X^{(2)}(x) = R_2 e^{\alpha_2 x}, \tag{7.6.ii}$$

where

$$\alpha_2 = \beta m_2 + \alpha_0 \quad \text{and} \quad R_2 = R_0 \exp(\beta z^{(2)}(0)).$$

(Compare (7.6.ii) with 7.4.ii).)

The corresponding $\lambda_T^{(1)}(\cdot)$ and $\lambda_T^{(2)}(\cdot)$ can be obtained straightforwardly.

The question, whether the different lifetime distributions for males and females can be adequately represented by models of the types introduced in Examples 8 and 9, still remains open. Of course, neither may be correct. The decision does not belong to statisticians alone.

7.2. Clinical Trials

Most clinical trials are designed to investigate whether some

treatments or drugs are better than others, for a particular disease. Preventive trials are usually aimed at treating a specific characteristic considered as a risk factor so as to prevent or, at least, delay the occurrence of an intervening event which accelerates, or leads to sudden death. For example, reducing blood pressure and blood sugar in diabetic patients is considered as likely to extend a patient's life. At present, the question whether the reduction of the level of plasma cholesterol prevents heart failure is among the controversial topics in many current clinical trials.

There can be a large number of experimental designs, models and kinds of analysis. Here we just use two examples, again to illustrate techniques in constructing hazard rate functions.

EXAMPLE 10. *(Treatment and control populations)*

Suppose that we consider a controlled clinical trial with populations (1) and (2) representing *treatment* and *control* populations, respectively. We assume exponential hazard.

Let

$$z = \begin{cases} 1 \text{ for treatment population (1)} \\ 0 \text{ for control population (2).} \end{cases} \tag{7.7}$$

Similarly, as in Example 8, for the *control* population, we have

$$\mu_X^{(2)}(x;\ z=0) = R_0 e^{\alpha_0 x}, \tag{7.8}$$

and

$$\lambda_T^{(2)}(t|x;\ z=0) = R_0 e^{\alpha_0 x} e^{\alpha_0 t} \tag{7.9.i}$$

(see (7.5.ii)).

Suppose that for an individual to whom treatment is applied at age x, the hazard rate changed (allowing for interaction with age). Then

up to age x, the hazard rate, $\mu_x^{(1)}(\cdot)$, is the same as $\mu_x^{(2)}(\cdot)$ defined in (7.8). While the future lifetime hazard rate is

$$\lambda_T^{(1)}(t|x;\ z=1) = R_0 \exp[\beta z + \alpha_0(x+t) + \delta z t]$$

$$= R_0 e^{\beta} \exp[\alpha_0 x + (\alpha_0 + \delta)t]$$

$$= R_1 e^{\alpha_0 x} e^{\alpha_1 t}\ , \qquad (7.9.ii)$$

where

$$R_1 = R_0 e^{\beta}\ ,\quad \alpha_1 = \alpha_0 + \delta\ . \qquad (7.10)$$

(Notice the difference between (7.9.ii) and (7.5.ii)).

EXAMPLE 11. *(Treatment acting upon a concomitant variable)*

A drug is used for lowering blood pressure (z) in hypertensive patients. Suppose that for hypertensive patients

$$z(x) = z(0) + mx\ . \qquad (7.11)$$

Thus for the *control* group (2), we have

$$z^{(2)}(x+t) = z(x) + mt\ . \qquad (7.12)$$

Assuming Gompertzian model

$$\mu_x(x;\ z(x)) = R_0 \exp[\beta z(x)] e^{\alpha_0 x}\ , \qquad (7.13)$$

the future hazard rate for control group (2) takes the form

$$\lambda_T^{(2)}(t|x;\ z(x)) = R_0 \exp\{\beta[z(x) + mt]\} e^{\alpha_0(x+t)}$$

$$= R_0 \exp[\beta z(x) + \alpha_0 x] e^{\alpha_2 t}\ , \qquad (7.14)$$

where

$$\alpha_2 = mt + \alpha_0\ . \qquad (7.15)$$

We now speculate about the $\lambda_T^{(1)}(\cdot)$ for the *treatment* group (1). Suppose that we can distinguish *two phases* in the action of the drug. Up to time τ_1, say, the blood pressure *decreases linearly*, and after time τ_1 it is *stabilized* at a certain level.

We then have for the *first phase*, for the period $0 \le t < \tau_1$

$$z^{(1)}(x+t) = z(x) - \theta_1 t\ ,\quad \theta_1 > 0\ , \qquad (7.16)$$

so that

$$\lambda_T^{(1)}(t|x;\ z(x)) = R_0 \exp[\beta z^{(1)}(x+t) + \alpha_0(x+t)]$$

$$= R_0 \exp\{\beta[z(x) - \theta_1 t] + \alpha_0 x + \alpha_0 t\}$$

$$= R_0 \exp[\beta z(x) + \alpha_0 x]\exp[(\alpha_0 - \beta\theta_1)t]$$

$$= R_0 \exp[\beta z(x) + \alpha_0 x]e^{\alpha_1 t}\ , \qquad (7.17)$$

where

$$\alpha_1 = \alpha_0 - \beta\theta_1\ . \qquad (7.18)$$

(Note that $\alpha_1 \le \alpha_2$.)

For the second phase, covariable z is stabilized at the value $z^{(1)}(x+\tau_1)$ (except, perhaps, for random variation). Thus for $t \ge \tau_1$, we have

$$\lambda_T^{(1)}(t|x;\ z(x)) = R_0 \exp\{[\beta z^{(1)}(x+\tau_1) + \alpha_0[x+\tau_1 + (t-\tau_1)]\}$$

$$= R_0 \exp\{[\beta z(x) - \beta\theta_1\tau_1] + \alpha_0 x + \alpha_0\tau_1 + \alpha_0(t-\tau_1)\}$$

$$= R_0 \exp(-\beta\theta_1\tau_1)\exp[\beta z(x) + \alpha_0 x]e^{\alpha_0 t}$$

$$= R_0^* \exp[\beta z(x) + \alpha_0 x]e^{\alpha_0 t}\ , \qquad (7.19)$$

where

$$R_0^* = \exp(-\beta\theta_1\tau_1)\ . \qquad (7.20)$$

(Note $R_0^* \le R_0$.)

We then have, for the treatment group (1)

$$\lambda_T^{(1)}(t|x;\ z(x)) = \begin{cases} R_0 \exp[\beta z(x) + \alpha_0 x]e^{\alpha_1 t} & \text{for } 0 \le t < \tau_1 \\ R_0^* \exp[\beta z(x) + \alpha_0 x]e^{\alpha_0 t}\ , & \text{for } t \ge \tau_1\ , \end{cases} \qquad (7.21)$$

with $\alpha_1 \le \alpha_0 \le \alpha_2$ and $R_0^* \le R_0$.

To establish the changes in the blood pressure (z) under treatment, repeated measurements are needed. Provided that the model (7.21) is correct, it would still be difficult to establish the value of τ_1. It would be, perhaps, not unreasonable, to consider τ_1 as a parameter.

Further complications occur when more than one covariables are considered simultaneously; also, there might be an interaction among them, and they might be correlated. It is, however, useful to examine, first, models using one *covariable at a time*.

8. DISCUSSION

8.1. The basic idea of this article is that for epidemiological follow-up studies and clinical trials, the survival models are, in fact, *future lifetime distributions*, and have to be consistent with the original distribution of lifetime.

8.2. For modeling survival distributions or, equivalently, hazard rate functions, we have chosen an *exponential* function; the corresponding SDF is a Gompertz survival function. Among the reasons for choosing this form are the following: Gompertz distributions fit most human mortality data, at least for older ages; its parameters have some simple biological interpretations; it has convenient mathematical properties such as, for example, the 'preservation' property; the extension of the model by incorporating concomitant variables in straightforward, when introducing functional relationship of covariates with age.

8.3. We especially emphasize the importance of *repeated measurements* of covariables in time, on each individual. These measurements are needed in estimating the relations of z's with x, and, in turn, in constructing appropriate hazard rate models.

8.4. We also stressed the relationship of Gompertz models with Cox's

and logistic linear models. Several examples indicate circumstances in which the latter models are consistent with exponential hazard rate model, and in which they deviate from it.

REFERENCES

1. Cox, D.R. (1972). Regression models and life tables (with Discussion). *J. Roy. Statist. Soc. Ser. B, 33*, 187-220.

2. Elandt-Johnson, R.C. and Johnson, N.L. (1980). Survival Models and Data Analysis, (Chapter 7), Wiley, New York.

3. United States Life Tables: 1969-71, Vol. 1. No. 1 (1975). U.S. DHEW Publication No. (HRA) 75-1150. National Center for Health Statistics, Rockville, MD.

4. Vital Statistics of the United States, 1970. Vol. II - Mortality, Part B (1974). DHEW Publication No. (HRA) 75-1101, Rockville, MD.