GRAPHICAL GENERALIZED RESIDUALS IN FITTING DISTRIBUTIONS:
SOME PITFALLS AND ADVANTAGES

by

Regina C. Elandt-Johnson

and

Fraser B. Smith

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1811

December 1986
GRAPHICAL GENERALIZED RESIDUALS IN FITTING DISTRIBUTIONS: SOME PITFALLS AND ADVANTAGES

Regina C. Elandt-Johnson and Fraser B. Smith
Department of Biostatistics, University of North Carolina, Chapel Hill, N.C. 27514, USA

SUMMARY

Three kinds of generalized residuals based on probability integral transformation are defined, and their roles in graphical goodness of fit testing are discussed. They are useful in fitting parametric distributions but seem to be questionable in fitting Cox's proportional hazard rate (PHR) models. They almost always appear to give a good fit because of: (a) their nonparametric nature and (b) right-hand censoring. Thus the apparent overall fit has little inferential meaning. However, residuals are useful in exploratory analysis. A simple method based on nonparametric stratified residual plots in conjunction with log relative risk functions (LRRFs), in selecting risk factors and building survival models, is presented. Several examples illustrating the use and misuse of residual plots are given. Some theoretical and practical problems which still await solutions are suggested.

Key words: Goodness of fit; Log relative risk functions; Probability integral transformation; Proportional hazard rate model; Residual plot; Risk function; Survival analysis.

This work was supported by the U.S. National Heart, Lung, and Blood Institute contract NIH-NHLB-712273 from the National Institutes of Health.
INTRODUCTION

Graphical representation of data provides powerful methods in exploratory analysis, and is also helpful in the interpretation of the results. It requires, however, deeper analysis than might be thought, because it is easy to exaggerate or hide some of the information included in the data. This is so, in particular, when applying graphical methods based on generalized residuals in fitting semiparametric survival models with concomitant variables.

The purposes of this article are:

(i) To give a brief critical overview of the use of generalized residuals as graphical tools in assessing goodness of fit.

(ii) To present a simple method based on nonparametric stratified residuals in selection of risk factors, which together with analysis of log relative risk functions (LRRFs) is useful in constructing survival models.

(iii) To discuss some probabilistic and statistical problems which might still be of interest to theoretical and practical statisticians.

The paper consists of three parts.

In Part I, we define three kinds of generalized residuals, based on the probability integral transformation, and discuss theoretical aspects of their use in goodness of fit testing for parametric distributions. In Examples 1 and 2 the effects of scale transformations on visual interpretation of the results are emphasized.

Generalized residuals have recently been used in testing goodness of fit of semiparametric Cox's (1972) proportional hazard rate (PHR) models (e.g., Kalbfleisch and Prentice, Ch. 4, (1980), Aitken et al. (1983)). It has been shown (e.g., Smith (1985)) that such graphical tests are often not very helpful, because the PHR models almost always appear to fit, no matter how
many covariables are in the models, except, perhaps, for a few outliers. Some reasons for this phenomenon are discussed in Part II. Example 3 illustrates these points.

On the other hand, generalized residuals might be useful in the initial stages of analysis of survival data: in selection of covariables and incorporating them into the model. Some simple graphical techniques using stratified nonparametric residual plots and log relative risk functions (LRRFs) in model building are described in Part III and illustrated in Example 4.

PART I. PARAMETRIC MODELS

2. DEFINITIONS OF u-, e- AND w- RESIDUALS

We define three kinds of generalized residuals based on probability integral transformation.

Let \( F_X(x; \theta) = F_X(x) \) be the cumulative distribution function (CDF) of a continuous random variable \( X \) with parameter vector \( \theta \), and let \( S_X(x) = 1 - F_X(x) \). For convenience we will call \( S_X(x) \) the "survival function," even though \( X \) might not represent lifetime. We first consider complete data (i.e., no censoring).

2.1. Let \( x_1 < x_2 < \ldots < x_k \) represent \( k \) distinct (ordered) values of a continuous random variable \( X \) observed in a random sample of size \( N \) \((N \geq k)\), and let \( m_j \) denote multiplicity of \( x_j \). (Normally, we should have \( m_j = 1 \) for all \( j; \ m_j > 1 \) might arise from inaccuracy of measurement.)

The empirical distribution function (EDF) is

\[
F_X^0(x) = 1 - S_X^0(x) = 1 - \sum_{j=1}^{k} (1 - m_j/N_j),
\]

(2.1)

where \( N_j \) is the sample size at \( x_j \) \(- 0 \) \((N_j = N - \sum_{k=1}^{j-1} m_k)\), and
\[ S^O_X(x) = \prod_{x_j \leq x} \left(1 - \frac{m_j}{N_j}\right) \] (2.2)

is the empirical survival function (ESF). If there is no multiplicity, we have

\[ S^O_X(x) = \prod_{x_j \leq x} \left(1 - \frac{1}{N_j+1}\right) = \frac{N-1}{N} \quad \text{and} \quad F^O_X(x) = \frac{x}{N}, \] (2.3)

where \( x_1 < x < x_{i+1} \) (taking \( x_0 = 0 \), and \( x_{N+1} = \infty \)).

If the data fit the theoretical distribution, then plotting \( F_X(x_i) \) against \( F^O_X(x_i) \) should give roughly a straight line through the origin with slope equal to 1. Such a graph might be used as a basis for an "eye" goodness of fit test.

2.2. The transformed random variable

\[ U = F_X(X) \] (2.4)

has the (standard) uniform \((0,1)\) distribution. Let

\[ u_i = F_X(x_i; \theta) = F_X(x_i), \quad i = 1, 2, \ldots, N. \] (2.5)

The order of the \( u_i \)'s is the same as the order of the \( x_i \)'s, so that

\[ F^O_U(u_i) = F^O_X(x_i) \quad \text{for all } i. \] (2.6)

Hence,

- plotting \( F^O_U(u_i) \) against \( u_i \) is identical to plotting \( F^O_X(x_i) \) against \( F_X(x_i) \).

This also implies that formal goodness of fit tests based on the EDF-statistics of \( x \) can be transformed to equivalent tests based on the EDF-
statistics of \( u \) (Stephens (1986)). We will call the transformed variables \( u_i \) the \( u \)-residuals.

As possible alternatives to the \( u \)-residuals, we also define the \( e \)-residuals, where

\[
e_i = -\log S_X(x_i) = \log(1-u_i),
\]

(2.7)

and the \( w \)-residuals, where

\[
-w_i = -\log F_X(x_i) = -\log u_i,
\]

(2.8)

for \( i = 1, 2, \ldots, N \). Note that each of \(-\log S_X(x)\) and \(-\log F_X(x)\) has a (standard) exponential distribution with parameter \( \lambda = 1 \). Also notice that \(-\log S_X(x)\) represents the cumulative hazard function (HRF), while \(-\log F_X(x)\) has no special interpretation. Also, since \(-\log S_X(x)\) and \(-\log F_X(x)\) are monotone functions of \( u \), their ordering is the same as the ordering of \( u \).

In summary, if the data fit the parametric model \( F_X(x; \theta) \), then the plots of \( F^0_X(u_i) \) against \( u_i \) or of \( \log S_X^0(e_i) \) against \( e_i \), or of \( \log F_X^0(w_i) \) against \( w_i \), should each give approximately a straight line through the origin with slope 1. We will call them briefly the \( u \)-, or \( e \)-, or \( w \)-residual plots.

2.3. These results are also valid for censored data, which usually occur in survival analysis. Let \( t \) denote survival time, and let

\[ t_1 < t_2 < \ldots < t_k \]

be the distinct times at death, with multiplicity of deaths \( d_j \) "at \( t_j \)." Further, let \( R_j \) be the number of individuals alive at \( t_j - 0 \) (risk set \( R_j \)).

The ESF can be represented by the product-limit estimator of SDF,

\[
S^0_T(t) = \prod_{t_j \leq t} (1 - d_j/R_j).
\]

(2.9)
2.4. Another approach to estimation of SDF (used in some package programs is by approximating the hazard rate function by a step-function

\[ \lambda_I(t) = \lambda_i \text{ for } t_{i-1} \leq t < t_i, \ i = 1, 2, \ldots, k, \] \hfill (2.10)

with \( t_0 = 0 \) and \( t_{k+1} = \infty \). Assuming that an individual censored in \( [t_{i-1}, t_i) \), has been censored "at \( t_{i-1} \)," we obtain

\[ \hat{\lambda}_i = \frac{d_i}{R_i(t_i-t_{i-1})}, \ i = 1, 2, \ldots, k. \] \hfill (2.11)

The cumulative hazard function (CHR) is

\[ \hat{\Lambda}_I(t) = \int_0^t \hat{\lambda}_I(y) dy = \sum_{j=1}^{i-1} (t_j-t_{j-1})\hat{\lambda}_j + (t-t_i)\hat{\lambda}_{i+1} \text{ for } t_i \leq t < t_{i+1}. \] \hfill (2.12)

The estimated SDF is

\[ \hat{S}_I(t) = \exp[-\hat{\Lambda}_I(t)]. \] \hfill (2.13)

Note that for small \( d_j/R_j \) (\( \leq 0.1 \)), we have

\[ \hat{S}_I(t_i) = \prod_{j=1}^{i-1} \left[ 1 - \frac{d_j}{R_j} \right], \quad \hat{S}_I(t_{i+1}) \approx \prod_{j=1}^{i} \left[ 1 - \frac{d_j}{R_j} \right] = S^0_I(t_i), \] \hfill (2.14)

where \( S^0_I(t_i) \) is defined by (2.9).

Thus, for sufficiently large \( R_i \) relative to \( d_i \), the difference between (2.9) and (2.13) might be negligible.

3. IS IT LEGITIMATE TO USE RESIDUALS WHEN THE PARAMETERS OF THE DISTRIBUTION ARE ESTIMATED?

We have seen how the \( u-, e- \) and \( w- \) residuals plots can be used in graphical assessment of goodness of fit of theoretical distributions. Problems arise, however, when the parameters \( \theta \) are estimated from the sample, which is usually the case. Thus the (unordered) \( U_i \)'s are:
(a) not uniformly distributed;
(b) not independent.

A general treatment of these problems was given by David and Johnson (1948) who gave, in general terms, the joint distribution function of the (unordered) \( U_i \)'s and, under certain conditions, the distributions of individual \( U_i \)'s. Explicit solutions were given for normal and exponential distributions. In both of these cases, the joint distributions of the \( U_i \)'s are independent of the parameters. Also, the distribution of any individual \( U_i \) tends to be uniform as \( N \) increases, and is quite close to uniform for \( N \) as small as 21. Thus for sufficiently large \( N \), condition (a) does not present a serious problem. This might be not true for other distributions, though from our (heuristic) experience, this seems to be approximately true for lifetime distributions.

(c) Even if the estimated \( U_i \)'s are independent of the true parameters of the distribution, they are themselves, not mutually independent, and it is difficult to assess the effect of lack of independence on the plots; it is possible that this tends to smooth the plots. This also affects the appropriateness of some EDF-tests (e.g., Kolmogorov-Smirnov test). Stephens (1986) gives an extensive overview on this topic. Construction of the EDF-statistics is, however, not of major concern in the present paper.

4. DIFFERENT ROLES OF \( u- \), \( e- \) AND \( w- \)RESIDUALS IN EXPLORATORY DATA ANALYSIS

In this section we will discuss graphical assessment of goodness of fit for parametric distributions fitted to non-survival data. We use an example based on a large data set, and show that the plots of \( u- \), \( e- \) or \( w- \)residuals, though mathematically equivalent, give different visual impressions, because of different transformation scales.
EXAMPLE 1. The data used in this example represent plasma cholesterol \( x \) in mg/dl of 4477 white males aged 30 and over, from the Follow-up Study of the Lipid Research Clinics (LRC) Program, measured at entry to the study. There was an excess of individuals with high values of cholesterol in these data, and the sample was not randomly chosen from the U.S. population. The basic descriptive statistics for these data are: mean \( \mu \) = 222.9, standard deviation \( \sigma = 43.7 \), minimum value = 94.0, and maximum value = 535.0 (all in mg/dl). For more details see LRC (1974). Here, the data are used for illustrative purposes.

A lognormal distribution was fitted to these data. Figures 1.1a, b and c give the \( u \)-, \( e \)-, and \( w \)-residual plots, respectively.

Figures 1.1a, 1.1b, 1.1c

Note that in each graph a large number of observations is hidden, because they are close to other observations. The plot of \( u \)-residuals suggests that the fit is quite good, while that of \( e \)-residuals gives the impression that there might be some doubt about this. The exaggerated visual effect in the upper tail of the distribution in the \( e \)-residual plot is caused by distortion of the scale due to the logarithmic transformation, \( e_i = -\log S(X_i) \). A few of the largest values (with \( e > 7.0 \)) seem to exhibit considerable deviation, but \( e > 7.0 \) corresponds to \( u > 0.9991 \), which means that less than 0.1% of all observations in the upper tail of the distribution are responsible for this effect. Even if it is considered that the fit is not good for \( e > 3.0 \), this corresponds to less than 5% of all observations.

Similarly, the plot of \( w \)-residuals exhibits deviations in the lower tail of the distribution. For example, corresponding to \( w = -5.0 \), we have \( u = 0.0067 \), that is, less than 1% of the observations contribute to the apparent lack of fit.
Fig. 1.1a

Fig. 1.1b

Fig. 1.1c

Plasma cholesterol: Plots of u-, e-, and w-residuals (Example 1)

Legend: A=1 obs, B=2 obs, C=3 obs, etc.
These results suggest that it would be most reasonable to use u-residuals, unless we are especially interested in goodness of fit of tails of the distribution. Fortunately, in this example, we were able to use other forms of graphical representation. Figure 1.2 exhibits the histogram and fitted lognormal density.

Figure 1.2

The data were grouped in intervals of length 10 units (mg/dl) with first interval $x < 120$ mg/dl and the last interval $x > 370$ mg/dl. The Chi-square statistic for 24 d.f. was 50.24 and is significant at significance level $\alpha = 0.05$ ($\chi^2_{0.95;24} = 36.42$). The major contribution to this statistic comes, however, from the first interval ($-\infty, 120$), and has the value 13.30. This is in agreement with the u-residual plot.

We also calculated the EDF using the 4477 individual observations. The graph (not given here) shows a good fit with a slight deviation in the upper tail, but the Kolmogorov-Smirnov statistic is $D = 1.41$; this also is significant at $\alpha = 0.05$ (critical value is $D = 1.36$).

As a matter of curiosity, a very good fit with a lognormal distribution was obtained for the data when each observation $x$ was replaced by $(x+100)$, giving the observed $X^2 = 23.98$ which is not significant at $\alpha = 0.05$. This transformation markedly reduces the contribution of the class $(x+100) \leq 220$, yielding only 0.2294 instead of 13.30. But also the central part of the distribution fits better. We cannot offer any biological interpretation for this result.

5. SURVIVAL DISTRIBUTIONS WITH CONCOMITANT VARIABLES

Let $T$ denote survival time, $z$ be a vector of covariates and $F_T(t;z,\theta)$ the parametric CDF of $T$ given $z$. For simplicity, assume that the $z$'s do not
Fig. 1.2 Plasma cholesterol: Frequency distribution and fitted lognormal density
depend on time and are measured at entry (with t=0). Let $z_1$ be the set of covariates and $t_i$ the time at death for the $i$th individual who died. It is not appropriate to plot EDF $F_T(t_i)$ against $F_T(t_i; z_i; \hat{\theta})$ since the latter depends on the $z_i$. However, an equivalent goodness of fit test can be obtained, using empirical distributions of the residuals.

Survival data are usually censored. Consider a random sample of $N$ individuals. Let $t_{k'}$ be the time of last observation (death or censoring) for the $k'$th individual, and let the corresponding $u$-residual be

$$u_{k'} = u(z_{k'}) \div F_T(t_{k'}; z_{k'}, \hat{\theta}), \quad (5.1)$$

where $\hat{\theta}$ is an estimator of $\theta$. If $\hat{\theta} \neq \theta$, and the data fit the model, the $u_{k'}'$s would approximately represent a random sample from the standard uniform distribution.

Suppose that a death occurred in this sample. Let $u_1' < u_2' < \cdots < u_k'$ denote the ordered (distinct) $u$-residuals corresponding to $k'$ times (not necessarily distinct) at death; let $d_j'$ denote the number of multiple deaths associated with $u_j'$, and let $R_j'$ be the number of individuals with values of $u$ not less than $u_j'$. Note that, in general, the order of the $(u_j')'$s is not the same as the order of the $t_j'$s.

The product-limit estimate of CDF of the $u$'s is

$$F_U^O(u) = 1 - \prod_{u_j' \leq u} (1 - d_j'/R_j) \quad (5.2)$$

Plotting $F_U^O(u_1')$ against $u_1'$ should give approximately a straight line, if the data fit the model. (In the remaining text, we will drop the 'prime', if no ambiguity arises.)

Construction of EDF's for $e$- and $w$-residuals is straightforward.
EXAMPLE 2. The data are from Weindruch and Walford (1982) and represent complete cancer mortality of 1 year old mice on normal \((N_1 = d_1 = 67)\) and restricted \((N_2 = d_2 = 68)\) diets. (Two incidental deaths were excluded to simplify the analysis, which is presented here for illustrative purposes.)

A Weibull model

\[
S_T(t; \alpha, \beta, \gamma, z) = \exp(-\alpha \gamma e^{\beta z}),
\]

where

\[
z = \begin{cases} 
0 & \text{if normal diet} \\
1 & \text{if restricted diet,}
\end{cases}
\]

was fitted by maximum likelihood, yielding the estimates \(\hat{\alpha} = 7.003 \times 10^{-13}\), \(\hat{\gamma} = 7.846\), and \(\hat{\beta} = -0.1015\).

The u-residual plot in Figure 2.1a suggests that the overall fit is quite good; on the other hand, the e-residual plot (Figure 2.1b) and the w-residual plot (Figure 2.1c) indicate that there might be some lack of fit in both tails.

Figures 2.1a, b, c

For further checking, we plotted the ESF for normal diet \((z = 0)\) and its fitted SDF, \(\hat{S}_1(\cdot) = \exp(-\hat{\alpha} \hat{\gamma})\), and the ESF for restricted diet \((z = 1)\) and its fitted SDF, \(\hat{S}_2(\cdot) = \exp(-\hat{\alpha} \hat{\beta} \hat{\gamma})\) (Figure 2.2).

Figure 2.2

It appears that the group on restricted diet does not exhibit a good fit, especially in the left-hand tail. However, the values of Kolmogorov-Smirnov D statistics are \(D = 0.595\) for \(z = 0\), and \(D = 0.665\) for \(z = 1\); neither is significant.

Another check would be to construct separate u-residual plots for each group of mice on the same graph; if they superpose on each other, then the fit is good. This was the case in our example (the graph is not shown here.)
Fitting Weibull distribution to mice survival data.
Plots of u-, e-, and w-residuals.
Fig. 2.2. Empirical and fitted Weibull distributions for mice data
Another example of fitting a Weibull distribution with concomitant variables and using e-residuals in testing the overall fit is given in a paper by Kay (1977).

PART II. SEMIPARAMETRIC PHR MODELS

5. WHY DO THE PHR MODELS ALMOST ALWAYS APPEAR TO FIT?

6.1. Nonparametric models. So far we have discussed parametric models with relatively few parameters. Consider now survival data with no covariables and suppose that the survival function is estimated by ESF, \( S_T^O(t) \), using the product-limit formula (2.9) or the hazard step-function method (Section 2.4). Of course, we have

\[
u_i = F_i^O(t_i) = 1 - S_T^O(t_i)
\]  
(6.1)

and

\[
F_{u_i}^O = F_T^O(t_i),
\]  
(6.2)

so that plotting \( F_{u_i}^O \) against \( u_i \) gives a perfect straight line.

Both approaches—product-limit and hazard step-function—are called nonparametric. In fact, both are parametric with a large number of parameters— as many as the number of distinct failures. These remarks give some basis for the discussion in the next sections.

6.2. Fitting PHR models. The method of generalized residuals for testing goodness of fit of Cox's (1972) models with concomitant variables has recently been used in the analysis of survival data (e.g., Crowley and Hu (1977), Kalbfleisch and Prentice (1980), Aitkin et al. (1983), Smith (1985)).

The PHR model is defined by the equation

\[
\lambda_T(t; z, \beta) = \lambda_0(t) \exp(\beta'z),
\]  
(6.3)
where \( \lambda_T(t; z, \beta) = -d \log S_T(t; z, \beta) / dt \) is the hazard rate function of the distribution, and \( \lambda_0(t) \) is often an arbitrary function of \( t \) (sometimes called the 'underlying' hazard). In a general situation, the multiplier \( \exp(\beta' z) \) can be replaced by any suitable (non-negative) function \( g(z, \beta) \).

The survival function is

\[ S_T(t) = S_T(t; z, \beta) = [S_0(t)]^{\exp(\beta' z)}, \tag{6.4} \]

where

\[ S_0(t) = \exp[- \int_0^t \lambda_0(y) dy] = \exp[- \Lambda_0(t)] \tag{6.5} \]

is the 'underlying' SDF.

Since the form of \( \lambda_0(t) \) is usually not specified, it is estimated in piecewise fashion. Two commonly used methods are: (i) discrete probability approach (an analogue of product-limit) and (ii) hazard step-function. We note that the maximum likelihood estimates of the \( \beta \)'s can be obtained independently of \( \lambda_0(t) \) (e.g., using marginal likelihood).

(i) Using a discrete approach, we define \( q_i \) as the probability of death "at \( t_i \)." The maximum likelihood estimate of \( q_i \) can be obtained by solving numerically the equation

\[ \sum_{j \in D_i} \frac{\exp(\beta' z_j)}{1 - (1 - q_i) \exp(\beta' z_j)} = \sum_{l \in R_i} \exp(\hat{\beta}' z_l), \tag{6.6} \]

where \( D_i \) is the set of multiple deaths "at \( t_i \)," and \( R_i \) is the risk set at \( t_i - 0 \). Thus

\[ \hat{S}_0(t) = \prod_{j=1}^{i} (1 - \hat{q}_j) \text{ for } t_i \leq t < t_{i+1}. \tag{6.7} \]

(Kalbfleisch and Prentice (1980), Section 4.3.)
(ii) Assuming constant hazard over the interval \([t_{i-1}, t_i]\) (cf. Section 2.4), \(\lambda_{ii}\) is estimated from the formula

\[
\hat{\lambda}_{ii} = \frac{d_i}{(t_i - t_{i-1}) \sum_{\ell \in R_i} \exp(\hat{\beta}'z_{i\ell})}
\]

where \(d_i\) is the number of multiple deaths "at \(t_i.\" Thus

\[
\hat{\lambda}_0(t) = \sum_{j=1}^{i} (t_j - t_{j-1})\hat{\lambda}_{jj} + (t - t_i)\hat{\lambda}_{i+1},
\]

and

\[
\hat{S}_0(t) = \exp[-\hat{\lambda}_0(t)],
\]

for \(t_i \leq t < t_{i+1}\). (Breslow (1974), Elandt-Johnson and Johnson (1980), Chapter 13.)

Substituting \(\hat{S}_0(t)\) and \(\hat{\lambda}_0\) into (6.4), we obtain the estimated SDF for given \(z.\)

Let \(t_{i\ell} \) (\(t_i \leq t_{i\ell} < t_{i+1}\)) be the time last observed (death or survival) for an individual \(i\ell\) with covariates \(z_{i\ell}.\) Then the u-, e- and w-residuals are calculated from the formulae

\[
u_{i\ell} \triangleq \hat{F}_T(t_{i\ell}; z_{i\ell}, \hat{\beta}),
\]

\[
e_{i\ell} \triangleq \log \frac{\hat{S}_T(t_{i\ell}; z_{i\ell}, \hat{\beta})}{\hat{S}_0(t_{i\ell})},
\]

\[
w_{i\ell} \triangleq \log \frac{\hat{F}_T(t_{i\ell}; z_{i\ell}, \hat{\beta})}{\hat{F}_0(t_{i\ell}; z_{i\ell}, \hat{\beta})},
\]

respectively. In the remainder of this paper, we will confine ourselves to u-residuals.

All residuals (whether at death or censoring) are ordered and the EDF \(F_0^O(u)\) is plotted against \(u.\) If the data fit the model, the plotted points
should lie approximately on straight line through the origin with slope 1. Indeed, in most cases, they are very close to straight line, except for a few right-hand side outliers.

**EXAMPLE 3.** The data used in this example represent mortality from cardiovascular diseases (CVD) among N=4481 white males from the LRC Follow-up Study mentioned in Example 1. The follow-up (longest) period was slightly over 10 years. For details of this experiment see LRC (1974). There were d=200 CVD deaths during this period. The covariables in this example are:

- $z_1 = \text{age (in years)}$
- $z_2 = \text{cholesterol (in mg/dl)}$
- $z_3 = \text{logTriglyceride (logTrig)}$

(The range of triglycerides was 25-3717 mg/dl, and the distribution was very skew, so that it was decided to use logTrig, and not Trig.)

We first fitted three separate single covariable PHR models, with only age, or only cholesterol, or only logTrig in the model. In each case, the u-residual plots indicated a good fit. (To save space, these graphs are not given here.) Finally, we fitted the PHR model with all three covariables,

$$\lambda_T(t; z, \beta) = \lambda_0(t) \exp(\beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3). \tag{6.12}$$

The results of this last fitting are given below.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\hat{\beta}$</th>
<th>S.E.((\hat{\beta}))</th>
<th>$x^2$</th>
<th>Signif. level ($\alpha$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age ($z_1$)</td>
<td>0.09786</td>
<td>0.00599</td>
<td>266.83</td>
<td></td>
</tr>
<tr>
<td>Chol ($z_2$)</td>
<td>0.00487</td>
<td>0.00157</td>
<td>9.68</td>
<td>0.0019</td>
</tr>
<tr>
<td>LogTrig ($z_3$)</td>
<td>0.32149</td>
<td>0.11488</td>
<td>7.93</td>
<td>0.0051</td>
</tr>
</tbody>
</table>

All $\hat{\beta}$-coefficients are highly significant. The overall plot of u-residuals in Figure 3.1 also indicates a fairly good fit, except for a few odd points in the upper end.

**Figure 3.1**

(Note the points for u-plots are only marked for deaths.)
6.3. Why the "good fit" of the PHR models? The general answer is that
the fitted model includes such a large number of parameters that it becomes
almost nonparametric. (In our example, there were 200 λ's and 3 β's.) More
specifically:

(a) In clinical trials with patients suffering from a severe illness,
mortality is usually quite high, so that there is a large number of estimated
λ_i's. If there is also a small amount of censoring, then the model (6.3)
is almost nonparametric; this implies that the u-residual plot will represent a
good fit (cf. Section 6.1).

(b) In epidemiological follow-up studies with a large amount of
population data, mortality is often not too severe, but there is usually
heavy censoring. We are observing, in fact, the lower tail of the survival
distribution over a short period so that almost any form of distribution would
give a good fit (see Aitkin et al. (1983)). Also, in large data sets,
residuals are close to each other. The horizontal scale should then be
expanded. This is the situation in our Example 3. We plotted (the graphs
are not shown here) separate pieces of the graph for 0.00 ≤ u < 0.05,
0.05 ≤ u < 0.10, 0.10 ≤ u ≤ 0.15, and so on. It can be seen from Figure 3.1
that about 87% of the CVD deaths correspond to values of u between 0.00 and
0.16. We observed a fairly good fit in this range, but deviations from the
straight line were noticed for the remaining 13% of CVD deaths (for u > 0.16).

6.4. Stratification. In further checking, whether the mode of inclusion
of the covariables in the model is appropriate, stratification with regard
to each covariable in the model is used. Suppose that covariable z_i
is stratified into s strata. For each stratum, the EDF of u-residuals associated
with this stratum is calculated, and s different u-plots on the same graph
are obtained. If each u-plot fits approximately a straight line through the origin with slope 1, then inclusion of \( z_h \) in that form is deemed appropriate. If the fit is not good, some modification of the mode of inclusion of \( z_h \) might be desirable.

With the data of Example 3, we divided each covariable into 5 strata, using approximate sample quintiles, and graphical checking, as described above, was performed (the graphs are not given here). Because the plotted points were still concentrated in a narrow range, more than 60% of the points (corresponding to deaths) were hidden. It seems rather difficult to decide on the basis of such graphs how good is the fit; in our data the logTrig contribution appears to raise most doubts though we are also not certain that the linear exponent for cholesterol is sufficient to ensure a reasonable fit.

A similar method can be applied to a covariable \( z_g \) which is not in the model. Separation of residual plots for different strata would indicate that \( z_g \) should be included in the model after adjustment for covariables which are already in the model provided that the model is otherwise approximately correct.

Remark. In most published work, the e- rather than u-residuals have been used, because of their convenient interpretation in terms of cumulative hazard function. An example with several covariables and using e-residuals is given by Kalbfleisch and Prentice (1980), Section 4.5.

Lagakos (1980) argues that the e-residuals (formula (6.11b)) will not have standard unit exponential distributions, even if the true \( \beta \)'s are known, but are approximations to their conditional expectations given the ranks of the \( t_i \)'s. He suggested replacement of the \( e_i \)'s on the horizontal
axis by their ranks. Lagakos' arguments are not reviewed in the present paper, but might be of some interest for future work.

PART III. APPLICATIONS OF RESIDUALS IN EXPLORATORY ANALYSIS OF SURVIVAL DATA

7. SELECTION OF COVARIABLES FOR SURVIVAL MODELS

7.1. The use of residuals described in Section 6 was aimed at graphical testing of goodness of fit of a prespecified PHR model, which—as we have noticed—almost always appears to fit. Stratification of each factor included in the model and analysis of the EDF's of residuals in strata sometimes give some idea whether the model was or was not appropriately chosen, but it is not always clear from the graphs how to modify the model.

Of course, before the PHR model is chosen, some initial checking is often performed. Usually the assumption of proportionality is of major concern. A commonly used method is the following.

The covariable of interest, z_h, say, is stratified into s strata, and a separate model,

\[ \lambda_{Tr}(t; \mathbf{z}, \beta) = \lambda_{0r}(t) \exp(\hat{\beta}'\mathbf{z}), \quad \mathbf{z} = 1, 2, \ldots, s, \]  

(7.1)

with \( z_h \) not included in the model is fitted for each stratum. The maximum likelihood estimators of the common \( \beta \)-coefficients are obtained and the \( \hat{S}_{0r}(t) \)'s are calculated by methods described in 6.2(i) or 6.2(ii). Plotting \( \log[-\log \hat{S}_{0r}(t)] \) against \( t \) for \( \mathbf{z} = 1, 2, \ldots, s \) should give a set of approximately parallel lines if the proportionality assumption holds. Examples of application of this technique can be found in Kay (1977), and Kalbfleisch and Prentice (1980), Section 4.5. Since these plots usually form a set of curves (not straight lines), it is sometimes difficult to decide whether the curves can
be judged to be parallel. Using these plots (not given here) for the data in our Example 3, it appears that logTrig stratification indicates some doubt about the proportionality assumption.

7.2. We will now present a simple method, based on nonparametric residuals, which may serve several purposes:

(a) To decide whether a covariable \( z_h \) could be considered as a prognostic factor and be included in the model;

(b) Under certain conditions, to check proportionality assumptions;

(c) To investigate the mathematical form of the contribution of covariable \( z_h \) to the hazard function;

(d) Using the results from (a), (b) and (c), to construct survival models and examine their goodness of fit.

The preliminary steps in the method we propose are the following:

(i) First, we construct the EDF of survival time \( t \) with no covariables, \( F^O_T(t) \), using the product-limit formula (2.9) or the stepwise hazard function (2.13). The u-residuals for all individuals (dead or alive) are calculated and their EDF, \( F^O_U(u) \), is obtained. Since \( u = F^O_T(t) \equiv F^O_U(u) \), the plots of \( F^O_U(u) \) against \( u \) gives a perfect straight line (see Section 6.1).

(ii) Let \( z_h \) be a covariable of interest, stratified into \( s \) strata (levels). The EDF's of u-residuals associated with each stratum are calculated and \( s \) separate u-plots are obtained on the same graph. If the plots overlap, the covariable \( z_h \) should not be included in the model. If they give separate lines, \( z_h \) is a predicting factor and should be included in the model.

If there is heavy censoring and/or the mortality is light (no more than approximately 10% deaths over the specified period), then for the \( r \)th stratum, we have

\[
F^O_{U_r}(u) = \bar{S}^O_{U_r}(u) = 1 - S^O_{U_r}(u) = 1 - \exp\left[-\Lambda^O_{U_r}(u)\right] \approx \Lambda^O_{U_r}(u). \tag{7.2}
\]
Thus, if the plots of \( F^o_u(r) \)'s against \( u \) form straight lines which do not intersect, the hazard rates for the strata are approximately proportional. Otherwise, the proportionality assumption could be subject to some doubt. Of course, the picture might not be "clear-cut" as a result of random variation. Also, a careful examination of these plots may suggest some suitable mathematical form for inclusion of covariable \( z_h \).

(iii) Some covariables may exhibit interaction. Of special interest is the interaction of \( z_h \) with age \( (z_1, \text{say}) \). Suppose that a model with age,

\[
\lambda(t; z) = \lambda_0(t) \exp(\beta_1 z_1),
\]

is first fitted to the data. Thus, after adjustment for age, we may calculate u-residuals and apply stratification analysis for the remaining covariables. If, for example, the u-plots for the strata of covariable \( z_h \), now cross but did not cross when no covariables were in the model, it would suggest that there might be interaction between \( z_h \) and age. A better method of studying interaction would be by response surface techniques (but this is not a topic of this paper).

(iv) We may also use double stratification with respect to combinations of any two covariables, \( z_h \) and \( z_{h'} \). Such plots would sometimes suggest the degree of importance of each covariable as a predicting factor of mortality (cf. Figure 4.3). Interpretation of such plots, however, requires a deeper knowledge about the relations among covariables.

The methods we suggest in this Section are useful in preliminary analysis in construction of survival models. Some of these ideas are illustrated in Example 4.
EXAMPLE 4. The data in this example are the same as in Example 3. We divided each covariable into 5 strata, using (approximately) sample quintiles. Note that since logTrig is a monotonic function of Trig, stratification with respect to logTrig or Trig are equivalent, and so the u-plots for logTrig (or Trig) strata are identical.

The stratum limits for the three covariables are given below.

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Age ($z_1$)</th>
<th>Cholesterol ($z_2$)</th>
<th>Triglyceride</th>
<th>logTrig ($z_3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30.000-36.267</td>
<td>94-185</td>
<td>25-90</td>
<td>3.2189-4.5108</td>
</tr>
<tr>
<td>2</td>
<td>36.268-42.058</td>
<td>186-208</td>
<td>91-122</td>
<td>4.5109-4.8121</td>
</tr>
<tr>
<td>3</td>
<td>42.059-48.999</td>
<td>209-231</td>
<td>123-165</td>
<td>4.8122-5.1119</td>
</tr>
<tr>
<td>4</td>
<td>49.000-56.795</td>
<td>232-258</td>
<td>266-239</td>
<td>5.1120-5.4805</td>
</tr>
<tr>
<td>5</td>
<td>56.796-91.507</td>
<td>259-535</td>
<td>240-3717</td>
<td>5.4806-8.2207</td>
</tr>
</tbody>
</table>

(i) A nonparametric model with no covariables (using step-wise constant hazard) was first fitted. As expected, a perfect fit was obtained (the u-plots are not given here).

(ii) Each of the Figures 4.1a, b and c represent u-plots for 5 strata of age ($z_1$), cholesterol ($z_2$) and logTrig ($z_3$) (or equivalently, for triglyceride), respectively, as described in Section 7.1(i).

Figures 4.1a, b, c

Although about 25% of observations (deaths) are hidden, the graphs give a fair idea about the model. It appears that for age ($z_1$), model $\lambda_T(t,z_1) = \lambda_0(t)\exp(\beta_1z_1)$, is adequate, while cholesterol ($z_2$) and logTrig ($z_3$) might need some extra terms besides linear in the exponent of the hazard function. (For further discussion, see Section 8.2.)

(iii) We also fitted model (7.3) with age as a covariable, and constructed u-plots for strata of cholesterol and logTrig. Figures 4.2a and b represent these plots.
Fig. 4.1a. LRC data: Plots of nonparametric u-residuals stratified by age

Fig. 4.1b. LRC data: Plots of nonparametric u-residuals stratified by cholesterol

Fig. 4.1c. LRC data: plots of nonparametric u-residuals stratified by logTriglyceride (or triglyceride)

21a
4.2(a) Cholesterol

4.2(b) Triglyceride

LRC data: Plots of u-residuals adjusted by age in five strata of each of two covariables
Figures 4.2a and 4.2b

For these data, age is the most important risk factor. After adjustment for age, the separation of the strata is not so evident. It is also difficult to decide whether there is an interaction between age and remaining covariables. If so, it is probably not very strong, and in our preliminary analysis, we neglect this possibility. The plots for logTrig indicate that the proportionality assumption might be questioned. But since age is a dominating factor and the HRF is approximately proportional for age and for cholesterol, we do not raise this question in our preliminary analysis.

(iv) Figure 4.1 indicates that there is a substantial difference between strata (1+2) and strata (3+4+5), that is, between Chol <210 vs. Chol ≥210 (rounding to the nearest 10). Similarly, for logTrig, we combine strata (1+2+3) and strata (4+5). (In our analysis, we took logTrig = 5.00 as a division point.) We then consider four strata:

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Chol</th>
<th>logTrig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt;210</td>
<td>&lt;5.00</td>
</tr>
<tr>
<td>2</td>
<td>&lt;210</td>
<td>≥5.00</td>
</tr>
<tr>
<td>3</td>
<td>≥210</td>
<td>&lt;5.00</td>
</tr>
<tr>
<td>4</td>
<td>≥210</td>
<td>≥5.00</td>
</tr>
</tbody>
</table>

We fitted the model with no covariables, and then constructed u-plots for these four double-stratified groups in Figure 4.3.

Figure 4.3

It is clear from this graph that the major division is between strata (1+2) vs. (3+4). But strata 3 and 4 are those which include high level of cholesterol. This, of course, confirms the known fact that cholesterol (not triglyceride) is the major risk factor in CVD mortality.
Fig. 4.3. LRC data: Plots of nonparametric u-residuals using double stratification of Chol x logTrig
8. BUILDING THE PHR MODELS: HAZARD RATIOS

8.1. Consider a PHR model

\[ \lambda_T(t; z, z_h) = \lambda_0(t) \exp[\phi(z, \beta) + \psi(z_h, \beta_h)], \]  

(8.1)

where \( z_h \) is a covariable of interest. The functions \( \phi(\cdot) \) or \( \psi(\cdot) \) are some functions (usually polynomials) of covariables \( z \) and \( z_h \), respectively. Let \( z_{h0} \) and \( z_{hr} \) be two values of the covariable \( z_h \). Then

\[ v = v(z_{hr}) = \log \frac{\lambda_T(t, z, z_{hr})}{\lambda_T(t, z, z_{h0})} = \psi(z_{hr}, \beta_h) - \psi(z_{h0}, \beta_h) \]  

(8.2)

is the logarithm of the ratio of the hazard at \( z_h = z_{hr} \) to the "reference" hazard at \( z_h = z_{h0} \), when the covariables, \( z \), are kept the same. It is also sometimes referred to as a logarithm of the (instantaneous) relative risk (LRR). In particular, for the (log linear) model,

\[ \lambda_T(t, z, z_h) = \lambda_0(t) \exp(\beta' z + \beta_h z_h), \]  

(8.3)

we have

\[ v(z_{hr}) = \beta_h (z_{hr} - z_{h0}). \]  

(8.4)

Using different values of \( z_h \), we obtain the logarithm of the relative risk function (LRRF), \( v = v(z_h) \), which will be useful in further analysis of hazard rate models.

8.2. We now analyze our covariables by the LRRF techniques.

(a) For covariable \( z_h \), we introduce indicator variables, \( y_{hr} \), using stratum 1 as a "reference" level. For example, with 5 strata, we have
<table>
<thead>
<tr>
<th>Stratum</th>
<th>$y_{h2}$</th>
<th>$y_{h3}$</th>
<th>$y_{h4}$</th>
<th>$y_{h5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

and then fit the PHR model

$$
\lambda_T(t;y_{hr}) = \lambda_0(t)\exp\left(\sum_{r=2}^{5} \beta_r y_{hr}\right).
$$

The LRRF is represented by a step-function,

$$
v(y_{hr}) = \beta_r \text{ for } z_{hr} \leq z_h < z_{hr+1}, \ r = 2, 3, \ldots, s.
$$

For each stratum, we calculate the mean value, $\bar{z}_{hr}$, and join the points $(\bar{z}_{hr}, v(y_{hr}))$ by straight lines; such a graph gives us some idea what kind of continuous LRRF might be suitable.

(b) Next, we fit a PHR model with $z_h$ continuous. We usually start with a linear contribution,

$$
\lambda_T(t,z_h) = \lambda_0(t)\exp(\beta z_h),
$$

even though the graph discussed in (a) may suggest a higher order polynomial, for example, a cubic

$$
\lambda_T(t,z_h) = \lambda_0(t)\exp(\beta_1 z_h + \beta_2 z_h^2 + \beta_3 z_h^3).
$$

The evaluation of LRRF's for each of the (8.6) - (8.8) models at the mean values, $\bar{z}_{hr}$, using $\bar{z}_{h1}$ as a "reference" level, is straightforward.

Introducing a superscript (j) to denote the degree of the polynomial, we have
\[ v^{(0)}(\tilde{z}_{hr}) = \beta_r \quad \text{with} \quad v^{(0)}(\tilde{z}_{hl}) = 0, \quad (8.9a) \]

\[ v^{(1)}(\tilde{z}_{hr}) = \beta(\tilde{z}_{hr} - \tilde{z}_{hl}) \quad , \quad (8.9b) \]

\[ v^{(3)}(\tilde{z}_{hr}) = \beta_1(\tilde{z}_{hr} - \tilde{z}_{hl}) + \beta_2(\tilde{z}_{hr} - \tilde{z}_{hl})^2 + \beta_3(\tilde{z}_{hr} - \tilde{z}_{hl})^3 \]  

for \( r = 2, 3, 4, 5. \)

As mentioned before, the LRRF is a discrete function, while \( v^{(1)} \) and \( v^{(2)} \) are continuous functions over the whole range of \( z_h \).

We now illustrate this method, using the data from Example 4.

**EXAMPLE 4 (cont.).** We compare the LRRF for the discrete model (8.5) with the LRRF's for continuous PHR models for each of three covariables: age, cholesterol and logTrig.

1. **AGE.** We fitted model (8.5) and model (8.6). We will call the latter briefly, a "linear" model. The estimate of \( \beta \) in (8.6) is \( \hat{\beta} = 0.09503548 \); it is, of course, highly significant (\( \chi^2 = 266.37 \) with 1 d.f.). The estimated LRR functions, \( \hat{\nu}^{(0)} = \hat{\nu}^{(0)}(\tilde{z}_{lr}) \) and \( \hat{\nu}^{(1)} = \hat{\nu}^{(1)}(\tilde{z}_{lr}) \), are given in Table 4.1. To obtain some idea about mortality, we also include, for each stratum, the number of individuals at risk (\( N_r \)) and the number of CVD deaths (\( d_r \)) over about a 10-year period of investigation.

<table>
<thead>
<tr>
<th>Table 4.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 4.4 represents the results.</td>
</tr>
<tr>
<td><strong>Figure 4.4</strong></td>
</tr>
</tbody>
</table>

It suggests that the "linear" model (8.6) is approximately satisfactory. This is in agreement with a Gompertz law. Since, for these data, \( t \) represents future lifetime of this cohort, the arbitrary \( \lambda_0(t) \) could be, perhaps, replaced by the parametric function \( \lambda_0(t) = Ae^{\beta t} \).
### Table 4.1

**AGE: LOG RELATIVE RISK FUNCTIONS FOR STRATIFIED AND LINEAR COX'S MODELS**

<table>
<thead>
<tr>
<th>Stratum (r)</th>
<th>(N_r)</th>
<th>(d_r)</th>
<th>Mean (\bar{z}_{1r})</th>
<th>Log Relative Risks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>(\hat{v}(0)) (\bar{z}_{1r})</td>
<td>(\hat{v}(1)) (\bar{z}_{1r})</td>
</tr>
<tr>
<td>1</td>
<td>896</td>
<td>5</td>
<td>33.096</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>894</td>
<td>8</td>
<td>39.091</td>
<td>0.4611</td>
</tr>
<tr>
<td>3</td>
<td>899</td>
<td>21</td>
<td>45.510</td>
<td>1.4236</td>
</tr>
<tr>
<td>4</td>
<td>896</td>
<td>42</td>
<td>52.776</td>
<td>2.1450</td>
</tr>
<tr>
<td>5</td>
<td>896</td>
<td>124</td>
<td>64.885</td>
<td>3.3023</td>
</tr>
</tbody>
</table>

---

![Graph of \(v(z_1)\) vs. \(z_1\)](image)

**Fig. 4.4.** Log relative risk plots for age
2. **CHOLESTEROL.** Since cholesterol is an important risk factor, we present more detailed analysis of its contribution to the survival model.

The discrete model (8.5) indicates that a "cubic" model (8.8) might be more suitable than the "linear" model (8.7) (Figure 4.5a). However, the $\beta$-coefficients for (8.8) are not significant (Table 4.2), and a formally inclined statistician might disregard them, and apply the analysis to trimmed data. We fitted models (8.7) and (8.8) to the data over the restricted range $125 \leq \text{chol} < 300$. This includes $N = 4258$ observations with $d = 186$ CVD deaths (the corresponding figures for deleted values for chol $< 125$ were $N = 18$, $d = 1$, and for chol $\geq 300$, $N = 201$ and $d = 13$). The fit of the "cubic" model is quite good (see Table 4.2, Table 4.3, and Figure 4.5b), but this model might not be satisfactory for epidemiologists—why delete so many valuable observations because a convenient mathematical model does not fit all the data?

As can be seen from Example 1, a lognormal distribution of cholesterol fits better than normal, though it is not so skew as for triglycerides. We then fitted models (8.7) and (8.8) to logChol. Of course, we now use different units of measurements for cholesterol and have reduced the variation—this is reflected in the completely different values of the $\hat{\beta}$-coefficients, as well as their significance (Table 4.2), and also in the values and graphical presentation of the LRRF's (Table 4.3 and Figure 4.5c).

Table 4.2

Table 4.3

Figures 4.5a, 4.5b and 4.5c

**Remark 1.** Our stratification by quintiles is, of course, arbitrary. To obtain a better picture in the tails, the data were stratified into deciles. Since there were not many deaths in each decile range, the step-LRR function was not so smooth as for quintiles (a common situation with grouping into
<table>
<thead>
<tr>
<th>Variable</th>
<th>$\hat{\beta}$</th>
<th>S.E.$(\hat{\beta})$</th>
<th>$x^2$</th>
<th>Signif. level ($\alpha$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chol (Stratum)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_2$</td>
<td>0.2881</td>
<td>0.2928</td>
<td>0.97</td>
<td>0.3251</td>
</tr>
<tr>
<td>$y_3$</td>
<td>0.7788</td>
<td>0.2688</td>
<td>8.40</td>
<td>0.0038</td>
</tr>
<tr>
<td>$y_4$</td>
<td>1.0190</td>
<td>0.2605</td>
<td>15.30</td>
<td>0.0001</td>
</tr>
<tr>
<td>$y_5$</td>
<td>0.9365</td>
<td>0.2638</td>
<td>12.60</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

**Discrete model (8.5)**

<table>
<thead>
<tr>
<th>Chol (all data). Linear model (8.7)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chol</td>
<td>0.006188</td>
<td>0.001444</td>
<td>18.37</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chol (all data). Cubic model (8.8)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chol</td>
<td>-0.05751016</td>
<td>0.07746806</td>
<td>0.55</td>
<td>0.4579</td>
</tr>
<tr>
<td>(Chol)$^2$</td>
<td>0.00032528</td>
<td>0.00032839</td>
<td>0.98</td>
<td>0.3219</td>
</tr>
<tr>
<td>(Chol)$^3$</td>
<td>-0.00000051</td>
<td>0.00000046</td>
<td>1.28</td>
<td>0.2584</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>125 Chol 300. Linear model (8.7)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chol</td>
<td>0.008822</td>
<td>0.002001</td>
<td>19.45</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>125 Chol 300. Cubic model (8.8)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chol</td>
<td>-0.26575268</td>
<td>0.16257152</td>
<td>2.67</td>
<td>0.1021</td>
</tr>
<tr>
<td>(Chol)$^2$</td>
<td>0.00133398</td>
<td>0.00074660</td>
<td>3.19</td>
<td>0.0740</td>
</tr>
<tr>
<td>(Chol)$^3$</td>
<td>-0.00000210</td>
<td>0.00000112</td>
<td>3.49</td>
<td>0.0617</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LogChol (all data). Linear model (8.7)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LogChol</td>
<td>1.5932</td>
<td>0.3657</td>
<td>18.98</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LogChol (all data). Cubic model (8.8)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LogChol</td>
<td>-612.7773</td>
<td>273.3446</td>
<td>5.03</td>
<td>0.0250</td>
</tr>
<tr>
<td>(LogChol)$^2$</td>
<td>115.6003</td>
<td>51.3447</td>
<td>5.06</td>
<td>0.0245</td>
</tr>
<tr>
<td>(LogChol)$^3$</td>
<td>-7.2380</td>
<td>3.2175</td>
<td>5.06</td>
<td>0.0245</td>
</tr>
<tr>
<td>Stratum (r)</td>
<td>N_r</td>
<td>d_r</td>
<td>Cholesterol Range</td>
<td>Cholesterol Mean</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>875</td>
<td>20</td>
<td>94-185</td>
<td>165.731</td>
</tr>
<tr>
<td>2</td>
<td>908</td>
<td>28</td>
<td>186-208</td>
<td>197.169</td>
</tr>
<tr>
<td>3</td>
<td>912</td>
<td>45</td>
<td>209-231</td>
<td>220.231</td>
</tr>
<tr>
<td>4</td>
<td>893</td>
<td>56</td>
<td>232-258</td>
<td>244.534</td>
</tr>
<tr>
<td>5</td>
<td>889</td>
<td>51</td>
<td>259-535</td>
<td>286.574</td>
</tr>
</tbody>
</table>
small intervals). Nevertheless, the graph (not given here) also suggests a third order polynomial.

**Remark 2.** Although cubic polynomials fitted to all data for cholesterol and logChol exhibit fair fits in the range 150 to 250, say, they are completely unreliable for very low and very high levels of cholesterol (see Figures 4.5a and 4.5c). This should be taken into account when calculating relative risks.

3. **Triglycerides.** Similar analyses were performed for triglycerides and logTrig. Stratification was done using quartiles, quintiles and deciles, and we fitted linear and cubic LRR functions. To save space, the full analysis is not given here. Figure 4.6 (using deciles) gives some idea of the goodness of fit. Remark 2 for cholesterol, about using extreme values, applies here also.

**Figure 4.6.**

8.3. "Final" models. We fitted several other models, and finally chose to consider three models with age \((z_1)\), cholesterol \((z_2)\) (or logChol \((z_2')\)) and logTrig \((z_3)\) as being, perhaps, suitable for use in the first stage of analysis, when including all data (subject to the Remark 2). The models are the following:

(i) Discrete (in chol and logTrig) and "linear" in age

\[
\lambda_T(\cdot) = \lambda_0(t) \exp(\beta_1 z_1 + \sum_{r=2}^{5} \beta_{2r} y_{2r} + \sum_{r=2}^{5} \beta_{3r} y_{3r}),
\]

where \(y_{2r}\) and \(y_{3r}\) are indicator variables corresponding to quintiles of cholesterol and logTrig, respectively.

(ii) A "cubic" (in cholesterol) and linear (in age and logTrig) model,

\[
\lambda_T(\cdot) = \lambda_0(t) \exp(\beta_1 z_1 + \beta_{21} z_2 + \beta_{22} z_2^2 + \beta_{23} z_2^3 + \beta_3 z_3).
\]
(iii) A model similar to (8.11), with cholesterol \( z_2 \) replaced by

\[ \log \text{Chol} \ (z^*_2), \]

\[ \lambda_1(t) = \lambda_0(t) \exp(\beta_1 z_1 + \beta_2 z^*_2 + \beta_3 z^*_2^2 + \beta_4 z^*_2^3 + \beta_5 z_3). \]  

(8.12)

The \( \hat{\beta} \)-coefficients and their standard errors are given in Table 4.4.

Table 4.4

We also calculated the conditional LRR for a given covariable, while the others are kept constant, for models (8.10) and (8.12). Their values evaluated at the stratum means are given in Table 4.5.

Table 4.5

From Table 4.4, model (8.12) seems to be most appealing, since using \( \log \text{Chol} \) and \( \log \text{Trig} \) greatly reduced the variation, and all the \( \hat{\beta} \)-coefficients are significant. However, the biological interpretation in terms of \( \log z \) rather than original \( z \) is not always straightforward.

From Table 4.5, the conditional LRRs for age (when cholesterol and \( \log \text{Trig} \) are kept constant) are practically the same as those where only age is in the model (cf. Table 4.1). This is not surprising, since age is a dominating risk factor in this study. However, the conditional LRRs for cholesterol (when age and \( \log \text{Trig} \) are kept constant) are reduced (cf. Table 4.3). For \( \log \text{Trig} \), these figures show some variations which do not seem to be of great importance (the comparative figures, when only \( \log \text{Trig} \) is in the model, were calculated but are not given here).

9. DISCUSSION AND SOME PROBLEMS FOR FURTHER INVESTIGATION

(1) This presentation is intended to be helpful in the preliminary analysis of data, using plots and graphs. The emphasis is on what is suggested by their appearance, less attention being paid to formal tests of
### Table 4.4

The \(-\)coefficients for different Cox's models with three covariables

<table>
<thead>
<tr>
<th>Model</th>
<th>Variable</th>
<th>$\hat{\beta}$</th>
<th>S.E.((\hat{\beta}))</th>
<th>$X^2$</th>
<th>Signif. level ($\alpha$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(8.10)</td>
<td>Age</td>
<td>0.096951</td>
<td>0.005949</td>
<td>265.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chol: $y_2$</td>
<td>0.1729</td>
<td>0.2930</td>
<td>0.35</td>
<td>0.5551</td>
</tr>
<tr>
<td></td>
<td>$y_3$</td>
<td>0.5745</td>
<td>0.2693</td>
<td>4.55</td>
<td>0.0329</td>
</tr>
<tr>
<td></td>
<td>$y_4$</td>
<td>0.8259</td>
<td>0.2635</td>
<td>9.83</td>
<td>0.0017</td>
</tr>
<tr>
<td></td>
<td>$y_5$</td>
<td>0.7033</td>
<td>0.2685</td>
<td>6.86</td>
<td>0.0088</td>
</tr>
<tr>
<td>LogTrig</td>
<td>$y_2$</td>
<td>0.3114</td>
<td>0.2484</td>
<td>1.57</td>
<td>0.2101</td>
</tr>
<tr>
<td>(or Trig)</td>
<td>$y_3$</td>
<td>0.1496</td>
<td>0.2575</td>
<td>0.34</td>
<td>0.5613</td>
</tr>
<tr>
<td></td>
<td>$y_4$</td>
<td>0.5275</td>
<td>0.2411</td>
<td>4.79</td>
<td>0.0287</td>
</tr>
<tr>
<td></td>
<td>$y_5$</td>
<td>0.5590</td>
<td>0.2448</td>
<td>5.21</td>
<td>0.0224</td>
</tr>
<tr>
<td>(8.11)</td>
<td>Age</td>
<td>0.09776405</td>
<td>0.00599898</td>
<td>265.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chol</td>
<td>-0.07190264</td>
<td>0.07804398</td>
<td>0.85</td>
<td>0.3569</td>
</tr>
<tr>
<td></td>
<td>(Chol)$^2$</td>
<td>0.00037635</td>
<td>0.00033322</td>
<td>1.28</td>
<td>0.2587</td>
</tr>
<tr>
<td></td>
<td>(Chol)$^3$</td>
<td>-0.00000058</td>
<td>0.00000046</td>
<td>1.56</td>
<td>0.2116</td>
</tr>
<tr>
<td>LogTrig</td>
<td>0.34825931</td>
<td>0.11820546</td>
<td>8.68</td>
<td>0.0032</td>
<td></td>
</tr>
<tr>
<td>(8.12)</td>
<td>Age</td>
<td>0.09777769</td>
<td>0.00599755</td>
<td>265.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LogChol</td>
<td>-645.9776</td>
<td>304.3960</td>
<td>4.50</td>
<td>0.0338</td>
</tr>
<tr>
<td></td>
<td>(LogChol)$^2$</td>
<td>121.5588</td>
<td>57.1086</td>
<td>4.53</td>
<td>0.0333</td>
</tr>
<tr>
<td></td>
<td>(LogChol)$^3$</td>
<td>-7.5981</td>
<td>3.5679</td>
<td>4.54</td>
<td>0.0332</td>
</tr>
<tr>
<td></td>
<td>LogTrig</td>
<td>0.34300453</td>
<td>0.11764730</td>
<td>8.50</td>
<td>0.0036</td>
</tr>
<tr>
<td>Stratum</td>
<td>Mean</td>
<td>Model (8.10)</td>
<td>Model (8.11)</td>
<td>Model (8.12)</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Age</td>
<td>Chol</td>
<td>LogChol</td>
<td>Trig</td>
</tr>
<tr>
<td>1</td>
<td>33.096</td>
<td>165.731</td>
<td>5.1056</td>
<td>71.159</td>
<td>4.2326</td>
</tr>
<tr>
<td>2</td>
<td>39.091</td>
<td>197.169</td>
<td>5.2835</td>
<td>106.437</td>
<td>4.6512</td>
</tr>
<tr>
<td>3</td>
<td>45.510</td>
<td>220.086</td>
<td>5.3936</td>
<td>143.085</td>
<td>4.9535</td>
</tr>
<tr>
<td>4</td>
<td>52.776</td>
<td>244.534</td>
<td>5.4989</td>
<td>197.636</td>
<td>5.2805</td>
</tr>
<tr>
<td>5</td>
<td>64.883</td>
<td>286.574</td>
<td>5.6538</td>
<td>412.340</td>
<td>5.9134</td>
</tr>
</tbody>
</table>
goodness of fit of theoretical distributions to data. It might be worth-while comparing the graphical "eye" results with those obtained analytically. Besides this, there are a few other problems worth investigating.

(2) In fitting parametric models, what is the effect of high correlations among the estimators of the parameters of the original distribution of the form of the distribution of the transformed variable \( U = F_X(X) \)? How much does it deviate from uniformity? Though general solutions were given by David and Johnson (1948), they discussed explicitly only normal and exponential distributions. Gompertz and Weibull distributions might be of interest.

(3) For Cox's PHR models with concomitant variables, the numerical ordering of generalized residuals is, in general, not the same as that of survival times. We investigated the rank correlations between the \( u_i \)'s and \( t_i \)'s. The coefficients were sometimes considerably less than 1, especially when heavy censoring was present. This problem requires more formal and detailed treatment, to study its effects on the form of the distribution of generalized residuals (cf. Lagakos (1980)).

(4) "Hidden" points in graphical presentation could lead to incorrect conclusions. How can available computer procedures be improved to give more reliable pictorial presentation?

(5) The crucial point, however, seems to be in the treatment of extreme values of a given risk factor. These may, or may not, be outliers and/or influential observations. If, for example, we apply the "cubic" model (8.8) for cholesterol, it will be found that very high values of cholesterol (e.g., 500 mg/dl) give substantially lower risks of dying from CVD than low values of cholesterol (e.g., 120 mg/dl), which is clearly nonsense. We
cannot call this "extrapolation" since some high levels of cholesterol were present in our data. Nevertheless, the data at these levels are very meagre. This problem does not arise if we use a "linear" model, but it exaggerates the results in the opposite direction, and also does not give a good fit for the bulk of data.

Acknowledgement

We would like to thank Larry Sink for computation of tables and performing some of the plots and graphs.
References


