Turning the Analysis of Obesity–Mortality Associations Upside Down: Modeling Years of Life Lost Through Conditional Distributions

Henry T. Robertson, Gustavo de los Campos and David B. Allison

Objective: We demonstrate the utility of parametric survival analysis. The analysis of longevity as a function of risk factors such as body mass index (BMI; kg/m²), activity levels, and dietary factors is a mainstay of obesity research. Modeling survival through hazard functions, relative risks, or odds of dying with methods such as Cox proportional hazards or logistic regression are the most common approaches and have many advantages. However, they also have disadvantages in terms of the ease of interpretability, especially for non-statisticians; the need for additional data to convert parameter estimates to estimates of years of life lost (YLL); debates about the appropriate time scale in the model; and an inability to estimate median survival time when the censoring rate is too high.

Design and Methods: We will conduct parametric survival analyses with multiple distributions, including distributions that are known to be poor fits (Gaussian), as well as a newly discovered “Compressed Gaussian” distribution.

Results: Parametric survival analysis models were able to accurately estimate median survival times in a population-based data set of 15,703 individuals, even for distributions that were not good fits and the censoring rate was high, due to the central limit theorem.

Conclusions: Parametric survival models are able to provide more direct answers, and in our analysis of an obesity-related data set, gave consistent YLL estimates regardless of the distribution used. We recommend increased consideration of parametric survival models in chronic disease and risk factor epidemiology.

INTRODUCTION

The associations or effects of chronic disease risk factors such as body mass index (BMI, kg/m²), serum cholesterol, or blood pressure on health and lifespan are of great interest and importance. Interested parties include litigators trying wrongful death cases and determining appropriate settlements, demographers estimating population trends and planning accordingly, insurers setting premium rates, public health officials advising the public, policy makers determining priorities, clinicians advising their patients, and the general public. The analysis of such data is made challenging by several factors, perhaps most notably that not all individuals will be observed until their time of death, leading to censoring in the survival times.

Cox proportional hazards regression is the most common way to accommodate censoring (1). The many advantages of this approach are well-documented and described elsewhere (2). However, there are at least three major disadvantages. The first involves the inability to estimate medians or other quantiles of survival time when the censoring rate exceeds the quantile of interest. The second concerns interpretability. The primary output of such an analysis is a hazard ratio, which requires understanding of calculus to interpret, is not easily understood by many nonstatisticians, and is not expressed in units such as years of life lost (YLL) that are part of everyday parlance and well understood. Third, the proportionality assumption may not hold.

In practice, large epidemiological data sets such as the National Health and Nutrition Examination Surveys (NHANES), the National Health Interview Surveys (NHIS), Atherosclerosis Risk in Communities (ARIC), and the Framingham Heart Study are often analyzed at follow-up times which have censoring rates well above 50%; hence, the median survival times for subjects may be estimable only for subjects at greatly increased risk.

Moreover, additional data beyond those necessary for the initial analysis are required in some approaches (3) to convert hazard ratio
estimates into expected survival times or YLL. YLL is defined here as the difference in conditional expectations of survival time between individuals who differ only in their level of risk factors.

Fontaine et al. (3) developed a method for converting hazard ratio estimates obtained with Cox proportional hazards regression to estimates of YLL that can be used by clinicians, the general public, and those interested in understanding the effects of factors such as high BMI on relevant aspects of lifespan. Such an approach, while useful, is cumbersome to implement and required three different data sets (one to estimate the hazard ratios, one to estimate distributions of the risk factor in the general population, and one to estimate overall survival distributions in the general population). Furthermore, there was no readily accessible solution for obtaining a confidence interval for YLL estimates. Hence, a method which more directly yielded estimates expressed in terms of years of survival time would be more desirable.

Fully parametric models offer an alternative to Cox regression that can provide direct estimates of YLL even in the presence of high censoring rates. However, new problems emerge: namely, which distribution should be used? Human longevity is characterized by (a) bimodality, including peaks at infancy and old age; (b) strong left skew. If investigators seek a good fit over all ages, then they may need to address the bimodality challenge by considering complex mixture models such as the five-parameter Siler model or the eight-parameter Heligman–Pollard and Mode–Busby models. However, models with fewer parameters are better for interpretability and reproducibility (4). In contrast, when the outcome of interest is YLL, life expectancy, or median survival times, then the fit of the tails may not matter greatly.

For instance, in a typical epidemiological study that estimates the effects of obesity or other metabolic risk factors on morbidity or mortality, the patients are adults, obviating the need for the distribution to accurately estimate mortality among the very young or very old. The central limit theorem guarantees that a normal parametric model (given a sufficient sample size) will accurately estimate the mean, even if the outcome is not normally distributed.

Unfortunately, the normal distribution, while easy to interpret, does not very effectively address the strong left skew challenge. The mean may differ significantly from the median, and other quantities will be inaccurately predicted. Closer approximations can be obtained through extreme value distributions such as Weibull or Gompertz. By convention, the Gompertz distribution is typically used to model all-cause mortality, whereas the Weibull distribution is used for specific causes of death (4), but a more effective solution would be desirable.

Recently, Robertson and Allison (11) introduced the compressed normal distribution, which was especially designed to accommodate features of the observed distribution of human lifespan after the period of high mortality rate in early childhood. This distribution expanded upon the findings of Kannisto (12), who observed that the distribution of longevity conditioned on survival to the modal age closely resembled the behavior of a normal distribution. Kannisto also noted that the standard deviation of remaining lifespan conditional upon having survived until age \( x \) seemed to decrease more rapidly as a function of \( x \) than would occur were total lifespan were normally distributed.

The normal distribution is characterized by the location-scale transformation:

\[
g(x) = \frac{x - \mu}{\sigma}
\]

Robertson and Allison (11) derived a distribution where a compression of the standard deviation occurs with advancing age, by modifying the location-scale transformation:

\[
g(x) = \frac{x - \mu}{\sigma(1 - x/\lambda)}
\]

TABLE 1 Descriptive statistics for ARIC study

| Group        | N   | Variable | Mean | SD  | Min | Max |
|--------------|-----|----------|------|-----|-----|-----|
| White males  | 5,420| Smoking  | 24.7%|      |     |     |
|              |     | Baseline age | 54.8 | 5.7 | 45  | 64  |
|              |     | Age at exit  | 68.2 | 5.9 | 46  | 81  |
|              |     | BMI        | 27.4 | 4   | 16.1| 56.3|
|              |     | Dead       | 16.8%|      |     |     |
| White females| 6,043| Smoking  | 25.0%|      |     |     |
|              |     | Baseline age | 54  | 5.7 | 45  | 64  |
|              |     | Age at exit  | 67.8 | 5.9 | 45  | 80  |
|              |     | BMI        | 26.6 | 5.5 | 14.4| 56.3|
|              |     | Dead       | 9.7% |      |     |     |
| Black males  | 1,620| Smoking  | 38.1%|      |     |     |
|              |     | Baseline age | 53.9 | 5.9 | 45  | 64  |
|              |     | Age at exit  | 66.5 | 6.2 | 46  | 80  |
|              |     | BMI        | 27.6 | 4.9 | 15.4| 54.4|
|              |     | Dead       | 26.4%|      |     |     |
| Black females| 2,620| Smoking  | 24.7%|      |     |     |
|              |     | Baseline age | 53.3 | 5.7 | 45  | 64  |
|              |     | Age at exit  | 66.5 | 5.9 | 45  | 80  |
|              |     | BMI        | 30.8 | 6.5 | 14.2| 65.9|
|              |     | Dead       | 16.6%|      |     |     |
| TOTAL        | 15,703| Smoking | 26.2%|      |     |     |
|              |     | Baseline age | 54.2 | 5.7 | 45  | 64  |
|              |     | Age at exit  | 67.6 | 6   | 45  | 81  |
|              |     | BMI        | 27.7 | 5.4 | 14.2| 65.9|
|              |     | Dead       | 15.0%|      |     |     |
where $\lambda$ is an upper bound of longevity and $(1 - x/\lambda)$ is the unspent portion of longevity at age $x$. The denominator decreases as $x$ increases. In effect, it conditions the scale on attained age, and models the increasing homogeneity of survivors as they age. The compressed normal distribution was found to model life table data more accurately than other three-parameter distributions, including the Makeham–Gompertz, generalized extreme value, generalized gamma, and the Azzalini skew-normal distributions.

To demonstrate the advantages of parametric survival analysis, we fit models of different distributions to a large epidemiologic data set with a high censoring rate. We also demonstrate the uses of multi-parameter optimization, which is not currently a common practice in survival analysis.

**Materials and Methods**

**Statistical analysis**

When age is the outcome of interest, and a study enrolls participant $i$ who was alive as of age $a_i$, the model should incorporate left truncation to reflect the conditional probability of surviving to age $a_i$ at the beginning of the study (6). Doing so makes the proper adjustments for older participants who have higher life expectancies (Figure 1). The likelihood equation is then:

$$L = \prod_{\text{observed}} f(X = a_i | \theta) \cdot \prod_{\text{censored}} \frac{P(X > a_i | \theta)}{P(X > c_i | \theta)}$$

where $c_i$ is the age of participant $i$ upon exiting the study, whether alive or dead; $f$ is the density function; $\theta$ is the vector of distribution parameters; and $\Pi$ is the product of a sequence.

In the process of writing this article, we identified a shortage of available software that is able to fit parametric survival models for left-truncated, right-censored data. As of this writing, parametric survival analysis in SAS is performed via *PROC LIFEREG*, but does not allow specification of age at entry (7). SPSS is not able to fit parametric survival models (8). In R, procedure *phreg* in package “eha” is theoretically able to do the above, but did not return plausible results (9). In STATA, procedure *streg* is able to do the above, but is not able to fit the normal or logistic distributions (10). Additionally, there appeared to be inconsistencies in the way log-likelihood scores are tabulated: some software dropped constant terms from the equations (such as $1/\sqrt{2\pi}$ for the normal distribution), whereas others kept them.

As equation (3) is a straightforward optimization problem, we decided to write our own software to maximize the likelihood and solve for the parameters. The software was written in R, and made
use of procedure *optim*. We specified the conjugate gradients method with gradient functions. The programs were short (a few dozen lines per distribution), and the model calculations only took a few seconds on modern desktop computers. All constant terms were preserved. We fit the Gompertz, Weibull, logistic, normal, and compressed normal distributions. We verified the consistency of estimated parameters with STATA software for the Gompertz and Weibull distributions.

**Study data**

For the purposes of illustrating parametric models and gauging their real-world utility, we selected a recent population-based study with a simple data structure that did not involve complex sampling, as with NHANES (13) or the National Health Interview Survey (14). A large data set including measured BMI values, smoking status, and age at follow-up was obtained from the Atherosclerosis Risk in Communities (ARIC) study (15), begun in 1987. These data are characterized by a high censoring rate (85.0%), such that the Cox model could not estimate median survival times. All ARIC participants were African-American or European-American, male, or female.

The variables fitted were smoking and BMI. Smoking was coded with indicator variables for current and former smokers. BMI was fitted as a cubic polynomial in keeping with conventions (3,16). For ease of interpretation, BMI variables were also centered and scaled as (BMI-25)/10, such that the “intercept” terms corresponded to a BMI of 25. We also tested for interactions between BMI and smoking, sex, and race. We checked that the interaction terms yielded results consistent with stratified analyses by race, sex, or smoking status. Also, we validated our findings with Cox models.

A total of 15,703 participants had known values for BMI, smoking status, and age at follow-up (Table 1). Fifteen percent of the participants had deaths observed over the course of the study, whereas 85% of the observations were censored. The participants in the study were from a relatively narrow age range of 45-61 years at baseline; the subjects were no older than 81 years at the end of the study. ARIC exemplifies the characteristics of many population-based studies, which have limited age ranges and high censoring rates. Nevertheless, the large sample sizes yielded estimates that were consistent with the previous population-based studies.

**Results**

**Model comparisons**

All five distributions yielded similar log-likelihood scores and gave similar estimates of longevity (Figure 2). This phenomenon occurred owing to the limited age range of the patients in the study, which limited the information on the tails of the distribution. All models yielded
FIGURE 4 YLL owing to BMI.

FIGURE 5 Hazard ratios inferred by Cox model.
similar results for the effects of predictor variables; smoking was associated with reduced longevity while BMI exhibited a J-curve pattern. The compressed normal distribution (in the solid black line) gave slightly lower estimates owing to its thicker left tail, which was found elsewhere to follow the distribution of life table longevities more closely than other distributions. The J-curve was more pronounced for smokers, consistent with some previous studies.

We found that YLL estimates were similar whether we defined them in terms of means or medians (Figure 4); again, this was consistent with the limited age range of patients in the study. We also verified our findings by fitting the Cox model (Figure 5); the estimated hazards are a mirror image of Figure 4.

As some past research has found, African-Americans had higher optimal BMIs than did whites and the difference was statistically significant (Table 2). White American nonsmokers had an optimal BMI of approximately 20, whereas African-American nonsmokers had an optimal BMI of 25. Smokers fared relatively better in the overweight (25-30) range of BMI. The shifting of the peak may reflect the greater prevalence of chronic diseases among smokers and African-Americans, such that a lower BMI was more likely to be a result of disease rather than good lifestyle; we will explore this topic further in future articles.

Conditional expectations

Finally, we illustrate one more benefit of parametric survival analysis (Figure 6). Life expectancy changes conditioned on attained age, as \( E(Y) \neq E(Y|Y>y) \). By making use of conditional expectations, one can compute remaining life expectancy for a patient at a given age. As baseline mortality rises with advancing age, the effect of risk factors on life expectancy decreases; this is apparent in the converging lines. This is a natural consequence of mathematics: among young people whose baseline mortality is low, a small change in the hazard rate causes a large increase or decrease in life expectancy. But among older people whose baseline mortality is high, the same change makes little difference in life expectancy. This phenomenon is consistent with repeated observations in the literature where BMI was centered and scaled as \((BMI-25)/10\).

| Variable     | Estimate | SE   | P-value |
|--------------|----------|------|---------|
| Mu           | 89.726   | 0.725| 0.0001  |
| Sigma        | 25.644   | 2.600| 0.0001  |
| Lambda       | 140.014  | 14.848| 0.0001 |
| Male         | −4.114   | 0.414| 0.0001  |
| African-American | −6.038 | 0.524| 0.0001  |
| Former smoker| −2.010   | 0.464| 0.0001  |
| Current Smoker| −8.310 | 0.607| 0.0001  |
| BMI^2        | −2.756   | 0.822| 0.0008  |
| BMI^3        | −2.202   | 1.064| 0.0386  |
| BMI^4        | 0.480    | 0.337| 0.1548  |
| Smoker × BMI | 4.723    | 1.187| 0.0001  |
| Smoker × BMI^2| −7.533 | 1.860| 0.0001  |
| Smoker × BMI^3| 2.909 | 0.831| 0.0005  |
| African-American × BMI| 2.821| 0.725| 0.0001  |

*BMI was centered and scaled as \((BMI-25)/10\).
prepared using ARIC research materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the ARIC of NHLBI.

© 2012 The Obesity Society

References

1. van Dijk PC. The analysis of survival data in nephrology: basic concepts and methods of Cox regression. *Kidney Int* 2008;74:705–709.
2. Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. Springer, 2005. p 74.
3. Fontaine KR, Redden DT, Wang C, Westfall AO, Allison DB. Years of life lost due to obesity. *J Am Med Assoc* 2003;289:187–193.
4. Juckett DA, Rosenberg B. Comparison of the Gompertz and Weibull functions as descriptors for human mortality distributions and their intersections. *Mech Ageing Dev* 1993;69:1–31.
5. Gage TB, Mode CJ. Some laws of mortality: how well do they fit? *Hum Biol* 1993;65:445–461.
6. Prabhakar C, Chicken E, McGee D. Time scales in epidemiological analysis: an empirical comparison. *Stat Med* 2009;1–13.
7. SAS Institute Inc. SAS 9.2 Help and Documentation. Cary, NC, 2002–2010.
8. UCLA: Academic Technology Services, Statistical Consulting Group. Introduction to SPSS. Available from: http://www.ats.ucla.edu/stat/spss/examples/asa2/chap8.htm (accessed October 6, 2011).
9. R Foundation for Statistical Computing, Vienna, Austria (2011).
10. StataCorp. 2011. *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.
11. Robertson HT, Allison DB. A novel generalized normal distribution for modeling human longevity and other negatively skewed data. *PLoS One* 2012, in press.
12. Kannisto V. Mode and dispersion of the length of life. *Population: an English selection. Biodemographic Perspectives on Human Longevity* 2001;13:159–171.
13. National Health and Nutrition Examination Survey. Centers for Disease Control and Prevention, National Center for Health Statistics [Internet] [accessed 2011-08-02]. Available from: http://www.cdc.gov/nchs/.
14. National Center for Health Statistics. Atlanta, GA: Centers for Disease Control and Prevention; 2010. 2009 National Health Interview Survey (NHIS) Public Use Data Release: NHIS Survey Description URL: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2009/srvydesc [accessed 2011-08-02].
15. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am J Epidemiol* 1989;129:687–702.
16. Heo M, Faith MS, Mott J, Gorman BS, Redden DT, Allison DB. Development of natural growth curves for body mass index in obese adults: an illustration of hierarchical linear models. *Stat Med* 2003;22:1911–1942.
17. Gelber RP, Kurth T, Manson JE, Buring JE, Gaziano JM. Body mass index and mortality in men: evaluating the shape of the association. *Int J Obes* 2007;31:1240–1247.
18. Garrison RJ, Feinleib M, Castelli WP, McNamara PM. Cigarette smoking as a confounder of the relationship between relative weight and long-term mortality. *The Framingham Heart Study. J Am Med Assoc* 1983;249:2199–2203.