Using Laplace Regression to Model and Predict Percentiles of Age at Death When Age Is the Primary Time Scale

Andrea Bellavia*, Andrea Discacciati, Matteo Bottai, Alicja Wolk, and Nicola Orsini

* Correspondence to Andrea Bellavia, Unit of Nutritional Epidemiology and Unit of Biostatistics, Institute of Environmental Medicine, Karolinska Institutet, Box 210, 171 77 Stockholm, Sweden (e-mail: andrea.bellavia@ki.se).

Initially submitted August 15, 2014; accepted for publication January 27, 2015.

Increasingly often in epidemiologic research, associations between survival time and predictors of interest are measured by differences between distribution functions rather than hazard functions. For example, differences in percentiles of survival time, expressed in absolute time units (e.g., weeks), may complement the popular risk ratios, which are unitless measures. When analyzing time to an event of interest (e.g., death) in prospective cohort studies, the time scale can be set to start at birth or at study entry. The advantages of one time origin over the other have been thoroughly explored for the estimation of risks but not for the estimation of survival percentiles. In this paper, we analyze the use of different time scales in the estimation of survival percentiles with Laplace regression. Using this regression method, investigators can estimate percentiles of survival time over levels of an exposure of interest while adjusting for potential confounders. Our findings may help to improve modeling strategies and ease interpretation in the estimation of survival percentiles in prospective cohort studies.

age; Laplace regression; survival analysis; survival percentiles; time scale

Abbreviation: CI, confidence interval.

In today’s epidemiologic research, results from time-to-event analysis are commonly reported in terms of increased/decreased risk of the event of interest in one group of individuals over another. To facilitate interpretation of the results, the estimation of risks may be complemented by time-based measures of association (1–3).

A possible way to combine information on risk and time is focusing on the percentiles of survival time (4). Given a follow-up period, the \( p \)th percentile of survival is the time \( t \) by which \( p \) percent of the study participants have experienced the event of interest. The percentiles of survival can be estimated at a univariable level with the nonparametric Kaplan-Meier estimator (5). When multivariable adjustment is required, other methods for estimating survival percentiles have been proposed, primarily consisting of postestimation calculation after fitting of Cox regression or other parametric survival models (6–9). Laplace regression was recently introduced in the epidemiologic literature as an intuitive and flexible method with which to estimate multivariable-adjusted survival percentiles (10).

Various reports have suggested that in observational prospective studies, when data are analyzed with Cox proportional hazards regression, attained age at the time of the event should be used as the underlying time scale instead of follow-up time (11–15). Given the extreme popularity of the Cox model, using attained age at the event as the primary time scale is becoming the standard way of analyzing time-to-event data.

However, the use of attained age has extensive and important consequences for the interpretation and estimation of the survival curve (16, 17). The presence of delayed entries introduces left-truncation, complicating interpretation of the survival curve, and changes the location of censoring, influencing the number of survival percentiles it is possible to estimate (13, 15–17).

In this paper, we investigate the implications of changing time scales in the estimation of survival percentiles with Laplace regression. When attained age is chosen as the primary time scale, the method models the percentiles of attained age at the event of interest (such as the median age at event conditionally on the predictors).
TIME SCALE, CENSORING LOCATION, AND PERCENTILES OF THE SURVIVAL DISTRIBUTION

A common option in survival analysis is assuming “time on study” (i.e., follow-up time or time since baseline) as the underlying time scale. In this situation, age-confounding is usually handled by adjusting for baseline age as a covariate in the model. An alternative option is to directly use age as the primary time scale (attained age, age at risk, age at onset), with subjects entering the study at their baseline age, causing the so-called delayed entries, and exiting at their event age or censoring age. When survival data are being analyzed with Cox regression, using age as the primary time scale has been widely recommended (11–14, 18, 19). This choice implies that adjustment for age is flexibly handled in a nonparametric form (13).

While the implications of choosing a specific time scale in the conduct of Cox regression have been discussed (18), the use of different time scales in the estimation of survival percentiles has never been investigated. The choice of attained age as the primary time scale has many relevant consequences for estimation and interpretation of the survival curve, and the derivation of a meaningful time-based measure of association may not be straightforward.

First, using age as the time scale implies that participants’ entries into the study are spread across the age range, with left-truncation occurring at the age of inclusion. In this scenario, the survival curve and its estimators, such as the Kaplan-Meier estimator, represent survival conditioned on having survived up to the earliest truncation time (20).

The mechanism of censoring is also affected by the change in the time scale (13). Let us assume a simple situation in which we follow a closed cohort for 15 years, during which 20% of the participants die. When focusing on follow-up time, defined as time from entry into the study to the date of death or censoring, cases are distributed across the follow-up period, while all censored observations (80%) occur at the end of the observation period, assuming no losses to follow-up. In this situation, any estimation of survival percentiles above the 20th percentile (such as median survival) would require data extrapolation beyond the range of observed data (10). When changing the time scale to attained age, defined as age between entry into the study and age at event or censoring, the 80% of censored observations are distributed across the entire range of age. In this situation, one could potentially estimate all the percentiles of age, up to the 99th percentile if the oldest participant were a case.

USE OF LAPLACE REGRESSION TO MODEL AND PREDICT ATTAINED AGE

The classical linear regression analysis establishes a linear relationship between a set of continuous predictors and the conditional mean of a response variable. Modeling only the mean, however, may miss important aspects of a predictor-outcome association, especially when the outcome’s distribution is skewed or has a large variance, and summarizing the association with a single coefficient might lead to a substantial loss of information. Evaluating the conditional percentiles of the outcome distribution helps to overcome these limitations, providing a comprehensive view of the association between the predictors and the entire distribution of the outcome (21). Given a percentile \( p \), a simple quantile regression model establishes a linear relationship between the predictors \( x \) and the conditional percentile of the outcome:

\[
Y_i(p) = x_i^T \beta(p).
\]

In survival analysis, the outcome is represented by a time variable that is typically skewed and censored. In this situation, modeling the mean of the outcome would not only introduce the discussed limitation but also typically require data extrapolation beyond the range of the observed data. The estimation of mean survival is commonly computed by assuming a parametric distribution for the unobserved part of the survival function, although it is not easy to capture the shape of the survival function after the end of follow-up (22).

Laplace regression was introduced in the biostatistical literature as a method of evaluating conditional quantiles of a potentially censored outcome (23). The method has been presented in the epidemiologic literature as a percentile-based approach for the analysis of time-to-event data (10). A user-friendly command for conducting Laplace regression in Stata (StataCorp LP, College Station, Texas) is available (24), and the method has been recently used in various epidemiologic studies (3, 25–29).

Laplace regression directly models the time variable of interest, without requiring any outcome transformation. When the time outcome is defined as time between entry into the study and either the event of interest or censoring (the end of follow-up), the method estimates adjusted survival percentiles. Let \( T_i \) be the follow-up time and \( x_i \) a \( k \)-dimensional vector of observed covariates for individual \( i \). \( T_i \) is censored, and we observe \( Y_i = \min(T_i, C_i) \), where \( C_i \) is a censoring variable. Given a proportion of events \( p \), Laplace regression establishes a linear relationship between the \( p \)th survival percentile \( T_i(p) \) and the covariates (23).

\[
T_i(p) = x_i^T \beta(p). \tag{1}
\]

We now change the time variable from the follow-up time of participant \( i \), \( T_i \), to the attained age of participant \( i \), \( A_i \). When not all of the participants experience the event of interest during follow-up, \( A_i \) is censored, and we observe \( Z_i = \min(A_i, B_i) \), the smaller value between the attained age at the event, \( A_i \), and the attained age at the end of follow-up, \( B_i \).

By fitting a Laplace regression model to the percentiles of attained age-at-event with only the intercept \( \beta_0 \), we obtain an overall crude estimate of the percentiles of attained age. For example, when \( p \) is fixed at 50, we are estimating the median age-at-event. These estimates correspond to those obtained with the Kaplan-Meier estimator using age as the primary time scale, and because of left-truncation they share the same limitations. Without left-truncation (e.g., use of follow-up time as the time scale) the \( p \)th survival percentile represents the time \( t \) by which \( p \% \) of the study participants experience the event of interest. With left-truncation (e.g., age-at-event), entries are spread throughout follow-up, and the estimate \( a \) for the \( p \)th percentile of age cannot be interpreted as the age at which \( p \% \) of the population experience the event. For

example, it could happen that by age \( a \), less than \( p \% \) of participants enter the study. An intuitive way of taking into account the presence of delayed entries, simplifying interpretation of the model at the individual level, is to condition age at the event on individual age at baseline. This can be done by including a function of age at baseline in the model.

\[
A_i(p) = \beta_0(p) + \beta_1(p) \times f(\text{age}_\text{baseline}_i).
\]

(2)

Age at baseline could be included in the model as a numerical predictor without any transformation. This would assume a linear relationship between age at baseline and the percentiles of age at death. To relax this assumption, the inclusion of a function of age at baseline, such as a categorized covariate or a mathematical transformation, should be preferred.

It is straightforward to extend model 2 by including exposures and potential confounders. For example, if we were interested in evaluating the impact of sex on the distribution of attained age, we could use the model

\[
A_i(p) = \beta_0(p) + \beta_2(p) \times \text{sex}_i + \beta_3(p) \times f(\text{age}_\text{baseline}_i).
\]

(3)

If sex were coded as 0 for female and 1 for male, the predicted \( p \)th percentile of age-at-event for a female participant would be \( \beta_0(p) + \beta_2(p) \times f(\text{age}_\text{baseline}_i) \), while the predicted value for a female would be \( \beta_0(p) + \beta_3(p) \times f(\text{age}_\text{baseline}_i) \). For a given \( p \), \( \beta_3(p) \) expresses the difference in the \( p \)th percentile of attained age between men and women, conditioned on baseline age. When \( p = 0.5 \), the model estimates sex differences in median age at the event. The inclusion of potential confounders in the model would produce, as in any regression method, multivariable-adjusted percentile differences. An implicit assumption of model 3 is that the effect of the main exposure on the percentiles of age at death is constant across levels of age at baseline. This assumption might be relaxed by including in the model a term for interaction between baseline age and the exposure of interest.

The model is easily extendable to include continuous exposures that can be modeled with flexible transformations such as splines or fractional polynomials. Interactions between covariates can also be assessed, and multiple percentiles can be estimated in simultaneous fits. The bootstrap method provides a simple estimator of the variance-covariance matrix for the multiple-quantile coefficient, and coefficients can be tested within and between different quantiles. Further mathematical and computational details of the Laplace estimator, together with the estimation procedure of the model, are extensively discussed in the original paper on Laplace regression by Bottai and Zhang (23).

**SIMULATION STUDY**

To examine the properties of Laplace regression in estimating multivariable-adjusted differences in age-at-event, we performed a simulation study. We simulated from model 3 a sample of 5,000 subjects with sex distributed as a Bernoulli variable with parameter 0.55. The true values of \( \beta_0(p) \), \( \beta_2(p) \), and \( \beta_3(p) \) were set to 77 years, 2 years, and 1 year, respectively. Baseline age was a uniform variable, and we used the square root as a function to allow for a possibly nonlinear relationship between age at baseline and age at death. An error term from a standard normal variable was added to the model. We specified 6 simulation scenarios according to an increasing fraction of censored observations (20%, 50%, and 80%) and an increasing range of age at baseline (45–55 years and 45–85 years). For each scenario, we estimated 3 percentiles of age at death \( p = 0.25, p = 0.50, \) and \( p = 0.75 \) with Laplace regression, modeling age at baseline using 4 different strategies (linear, square root, 5-year categories, and restricted cubic splines with 3 knots at fixed percentiles of its distribution). Web Tables 1 and 2 (available at http://aje.oxfordjournals.org/) show mean values for regression coefficients and standard errors estimated in 500 replications of the different scenarios, together with the empirical standard deviation of the estimates. The model showed good performance in estimating the parameters in all of the settings considered.

**ILLUSTRATIVE EXAMPLE**

To illustrate the use of Laplace regression in estimating attained age percentiles, we used data from 2 large prospective studies of Swedish men and women, the Swedish Mammography Cohort and the Cohort of Swedish Men (30). Information used herein was collected in 1997 with a self-administered questionnaire on lifestyle and dietary factors. After exclusion of participants with diabetes, cancer, or cardiovascular diseases at baseline, 71,238 participants aged 45–83 years were followed for 16 years (January 1, 1998–December 31, 2013). Through linkage with the Swedish Death Register, 16,346 deaths were documented during the follow-up period. We evaluated the impact of smoking on the percentiles of age at death.

Figure 1 illustrates the discussed consequences of changing the primary time scale from follow-up time to attained age at death. Panels A and B show the individual observation time for a sample of 10 participants. Delayed entries (observable in panel B) cause censoring observations to be spread across the age range (panel D). This situation is illustrated, for instance, in individual 5, who enters the study when 4 other participants have already been censored and another one has experienced the event. For this reason, the Kaplan-Meier estimator of the survival curve using attained age might still be calculated (panel F), but it becomes hard to interpret.

Given that age at death is a censored variable, a crude estimate of the median age at death of the study population can be obtained by fitting a Laplace regression on the 50th percentile of age at death with only the intercept. By fitting this model to our data, we obtained a coefficient of 88 years, representing an estimate of the median age at death conditioned on having survived up to age 45 years, the lowest age at entry into the study. Because of the presence of delayed entries, this result does not represent the age at which 50% of the study participants had died. To facilitate interpretation of the estimates, the model can be adjusted for age at baseline. For example, after fitting a Laplace regression model to our data, adjusting for 5-year categories of age at baseline, we observed that median survival for persons aged 65–69 years at baseline was 86.6 years. This result shows that 50% of
participants who entered the study between 65 and 69 years of age died before reaching age 86.6 years.

To evaluate the impact of smoking on the median age at death, we fitted a Laplace regression model to the 50th percentile of age at death using smoking categories as the main exposures (current, former, or never smoker) and adjusting for the categorical variable of age at baseline. We then evaluated the same association in a multivariable model that
Table 1. Percentile Differences, in Years, in Age at Death According to Smoking Status in the Cohort of Swedish Men and the Swedish Mammography Cohort, 1998–2013

<table>
<thead>
<tr>
<th>Model and Percentile</th>
<th>Smoking Status</th>
<th>Current Smoker</th>
<th>Former Smoker</th>
<th>Never Smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PD 95% CI</td>
<td>PD 95% CI</td>
<td>PD 95% CI</td>
<td></td>
</tr>
<tr>
<td>Model 1*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25th</td>
<td>0 Refferent</td>
<td>2.44 2.19, 2.71</td>
<td>4.13 3.89, 4.38</td>
<td></td>
</tr>
<tr>
<td>50th</td>
<td>0 Refferent</td>
<td>2.20 1.96, 2.44</td>
<td>3.76 3.53, 3.98</td>
<td></td>
</tr>
<tr>
<td>75th</td>
<td>0 Refferent</td>
<td>1.76 1.52, 1.99</td>
<td>3.15 2.93, 3.38</td>
<td></td>
</tr>
<tr>
<td>Model 2b</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25th</td>
<td>0 Refferent</td>
<td>2.66 2.42, 2.91</td>
<td>3.82 3.58, 4.06</td>
<td></td>
</tr>
<tr>
<td>50th</td>
<td>0 Refferent</td>
<td>2.40 2.14, 2.65</td>
<td>3.43 3.19, 3.68</td>
<td></td>
</tr>
<tr>
<td>75th</td>
<td>0 Refferent</td>
<td>1.98 1.72, 2.25</td>
<td>2.91 2.65, 3.16</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: CI, confidence interval; PD, percentile difference.

* Estimates were obtained by conducting a Laplace regression on the 25th, 50th, and 75th percentiles of age at death, with smoking status used as the main exposure and adjustment for age at baseline (6-year categories).

b Results were further adjusted for sex, alcohol consumption, body mass index, and educational level (all categorical variables).

DISCUSSION

A primary objective of medical research is to present results in an understandable form in order to facilitate the translation of scientific findings and help people in making healthy choices (31). Epidemiologic studies play a key role in the process of translating scientific discoveries into a population health context (32). In prospective cohort studies, the standard choice for data analysis is Cox regression with results presented in terms of hazard ratios, despite the recognized shortcomings of these measures (4). To facilitate the interpretation of results, it is becoming more frequent to complement the standard analyses by providing a time-based measure of association, which allows for direct interpretation of the exposure-outcome association in terms of time gained or lost (1–3, 33–35).

A possible approach to acquiring information on survival time is to focus on the survival curve, evaluating survival percentiles (4, 10). At the univariate level, this can be accomplished with the nonparametric Kaplan-Meier method. When multivariable adjustment is required, estimation of the survival curve is not straightforward, and even if methods for calculating adjusted survival curves have been proposed (6–9, 36), this type of analysis remains unpopular. Often, the survival curve is derived after performing Cox proportional hazards regression (34), but the original model assumptions force the curves to be proportional, strongly affecting the estimation of differences in survival between levels of the exposure (35).

Other methods for calculating and presenting time-based measures of exposure impact have been proposed, such as the rate or risk advancement period (37) or years of potential life lost (38), and they are occasionally applied in epidemiologic...
Laplace regression was introduced in the epidemiologic literature as a flexible and intuitive method of evaluating multivariable-adjusted survival percentiles (10, 42). In recent years, different applications of this method have appeared in epidemiologic and medical journals (3, 10, 25, 26, 28, 43).

To the best of our knowledge, none of the previous methods for deriving time-based measures of association are directly extendable to the situation in which attained age at the event is chosen as the primary underlying time scale in prospective studies. Various authors have suggested that this should be the primary choice when observational data are being analyzed with Cox regression (11–15, 18, 19), and given the popularity of the Cox model, using age as the primary time scale is becoming the standard option for analyzing time-to-event data.

However, the change in time scale strongly influences estimation and interpretation of the survival curve (16, 17). The presence of delayed entries introduces left-truncation, changing the meaning and interpretation of common estimators such as the Kaplan-Meier estimator. Therefore, introducing a flexible and intuitive approach with which to easily estimate and present time-based results when age is the primary time scale might represent a great advantage in the epidemiologic context.

We have shown in this paper how Laplace regression can help fill this gap. The major strength of this method is that percentiles of age-at-event are directly modeled as a linear function of a set of predictors. Coefficients estimated with this method are interpreted as differences in age-at-event (i.e., years, months, days) according to levels of a given exposure. Moreover, Laplace regression shares all of the advantages of a quantile/percentile approach, offering a more complete and accurate picture of the association of interest, by modeling the effect of explanatory variables on the location, shape, and scale of the distribution of the response variable (21, 23, 44). For example, in our data we observed that the difference in survival percentiles between current smokers and former/never smokers was higher in the lowest percentiles, suggesting that the association between smoking and mortality was stronger for people who died earlier, probably representing participants with weaker health. Finally, this application shares all the strengths of Laplace regression that have been widely discussed in previous applications (10, 28), such as allowing inclusion of interaction terms, simultaneously fitting the model to different percentiles, and testing coefficients within and between different quantiles. It also does not require any assumption about the shape of the survival curve, and it provides interval estimates and significance tests of the estimated differences. In the presence of delayed entries, one could potentially estimate all percentiles of age-at-event. However, as in any method dealing with quantiles of censored variables, caution must be adopted when there is a large amount of censoring before the estimated quantile and a low fraction of cases after the estimated quantile. Under this scenario, the limited information might influence the validity of the results.

In conclusion, in this paper we presented Laplace regression as a percentile approach for evaluating age at the time of an event. This method, conditioning on age at baseline, models the percentiles of the outcome, providing multivariable-adjusted differences in age at the event according to levels of exposure. Despite the presence of left-truncation, coefficients obtained with Laplace regression can be used to calculate individual predictions of age-at-event for specific covariate patterns. This application extends the use of the method to the numerous situations in which researchers adopt age as the underlying time scale. Laplace regression is a flexible and intuitive percentile approach that can be used to derive and present time-based measures of association, thus providing a useful complement to current epidemiologic research and facilitating the translation of scientific results.

ACKNOWLEDGMENTS

Author affiliations: Unit of Nutritional Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden (Andrea Bellavia, Andrea Discacciati, Alicja Wolk, Nicola Orsini); and Unit of Biostatistics, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden (Andrea Bellavia, Andrea Discacciati, Matteo Bottai, Nicola Orsini).

This work was partly supported by a Young Scholar Award from the Karolinska Institutet’s Strategic Program in Epidemiology. The study was also supported by the Swedish Research Council.

The funders played no role in the study design, data collection, or analysis, the decision to publish, or the preparation of the manuscript.

Conflict of interest: none declared.

REFERENCES