

Local Likelihood Methods to Smooth Hazard Rates from Grouped Data¹

*Stima non parametrica del rischio per dati aggregati
mediante verosimiglianza locale*

Francesco Lagona²
DIPSS - Università Roma Tre
lagona@uniroma3.it

Riassunto: Si propone un approccio basato sulla verosimiglianza locale per la stima non parametrica della funzione di rischio di dati aggregati. La procedura è illustrata attraverso due applicazioni riguardanti la mortalità umana e la sopravvivenza di cellule di lievito.

Keywords: hazard rate, local likelihood, grouped survival data.

1. Introduction

When the data are the result of a follow-up study and individual survival times are available, statistical models based on counting processes provide a natural framework to estimate the hazard function. For this type of data, the literature extends from purely non parametric to strictly parametric methods and suggests how to adjust for censored and truncated observations. These methods, however, do not cover the area of those biodemographic survival data that are constructed by drawing samples from a cohort at discretely spaced times and are hence gathered in the form of proportions or absolute counts. Examples include life-table data (counts of deaths occurred in an observation period within samples of contemporary survivors) and environmental sampling (counts of survivors in samples drawn from the environment where the individuals live).

When using grouped data to estimate the mortality hazard rate, models based either on the Binomial or the Poisson distribution may be useful. Hazard estimation may then be implemented by parametric or nonparametric methods.

When a parametric approach is pursued, the hazard function is usually assumed to be known up to a number of parameters to be estimated and plugged into the likelihood function of a Generalized Linear or Nonlinear model (GLM or GNLM), depending on the functional form of both the hazard and the link function.

Obviously, the parametric approach is useful when a mathematical theory of aging is available and needs to be tested on empirical data. But even when such a theory is not available, removing sampling errors is still important to capture systematic mortality patterns and nonparametric procedures can be helpful.

The principal idea of the present work is that a nonparametric estimate of the mortality hazard function can be derived from a smoothed estimate of the (link-transformed) regression function. In the case of life-table data, for example, the regression function models the expected number of deaths in the sample and the hazard function can be easily

¹Il lavoro è stato svolto con il contributo del Progetto 2002-2004 "Statistica nel supporto alla decisione ambientale: identificazione, monitoraggio e valutazione di intervento", Cofin - MURST 2002.

²Indirizzo per corrispondenza: Università Roma Tre, DIPSS, Via C. Segre 4, 00100 Roma

derived by transforming the estimated death probabilities via the negative complementary log transformation (Muller, *et al.*, 1997). Under the environmental-sample setting, the regression function models the expected number of survivors in the sample, and the hazard function is simply the slope of the (log-transformed) regression function. In both the applications, hence, the problem of nonparametric hazard estimation reduces to nonparametric estimation of the regression function.

Nonparametric regression has been extensively studied in the literature. The two most popular procedures are based on roughness penalty methods and local modelling. Each of these estimation schemes can be viewed as a "natural" approach to smoothing; each has reasonable properties in the sense of closeness to the "true" regression function (including the boundary region of the predictor's values) and each can be easily adapted to arbitrary likelihood functions. The basic difference between the two approaches is that roughness penalty estimators put the smoothing problem in the framework of optimizing a penalized version of the likelihood function, where penalty relates to the roughness of the regression function, resulting in elegant theoretical developments leading to smoothing splines theory (Hastie and Tibshirani, 1987). Local modelling (Fan and Gijbels, 1996), instead, fits a polynomial model locally within a smoothing window and relies on the idea that a smooth regression function can be approximated locally by a polynomial; in other words, models based on this approach are the local counterpart of GLMs. Since there are several results showing the (asymptotic) equivalence of smoothers based either on splines or varying bandwidth kernels (e.g., Silverman, 1984), the choice between the two methods is generally driven by the problem under study which may suggest the most "natural" approach to pursue, in order to interpret the results in an intuitively appealing manner.

2. A local likelihood smoother of the mortality hazard

The hazard smoother suggested in this paper is based on a simple two-steps procedure: first, a local likelihood approach is used to smooth the regression function of the deaths (in the life-table case) and the survivors (in the environmental sample case); second, the regression fitted values are used to estimate the mortality hazard. Regarding this second step, in the life-table case, we use the transformation approach (Muller *et al.*, 1997); for environmental sample data, we show how the mortality hazard is equal to the slope of the (log-transformed) Poisson regression function.

Local likelihood methods allow to smooth the regression function (and its derivatives) of random variables with an arbitrary distribution. To illustrate, let us consider a vector $(y_1 \dots y_n)$ of independent responses (e.g., deaths or survivors) paired to the n values $t_1 \dots t_n$ of one predictor (e.g., years or days). Let us assume that the distribution of the i th response belongs to the one-parameter exponential family

$$f(y_i) = \exp\{(\theta_i y_i - b(\theta_i))/a(\phi) + c(y_i, \phi)\} \quad (1)$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known function and θ_i is the canonical parameter. In parametric GLMs, the regression function μ_i is modelled linearly via a specified link function g and set of predictors, say $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$, where $\mathbf{x}_i^T$ contains the values of k predictors for the i th observation. When a parametric specification of the hazard rate is available and needs to be tested on the data, using GLMs leads to a *parametric* smoothing of the hazard rate. If a parametric model is not available, we could still try to model the

link of the regression function through a low-order polynomial time, for example

$$g(\mu_{t_i}) = \beta_0 + \beta_1 t_i + \beta_2 t_i^2. \quad (2)$$

Instead, we pursue a nonparametric approach, by observing that, while (2) is likely to give a poor *global* fit, it could fit better the data if used *locally*. More precisely, if the link of the regression function μ is smooth enough (i.e., it satisfies the regularity conditions of the Taylor's power series approximation theorem), then it can be approximated at each point t in a neighborhood of t_i by a low-order polynomial, say

$$g(\mu(t)) \approx \beta_{0,t} + \beta_{1,t}(t - t_i) + \beta_{2,t}(t - t_i)^2 \quad \beta_{k,t} = \frac{1}{k!} \frac{\partial^k g(\mu)}{\partial t^k} \quad k = 1, 2.$$

This suggests the following smoothing procedure. Fix a smoothing point t and define a window centered at t with bandwidth h containing all the observation points t_i such that $|t_i - t| \leq h$. Moreover, attach a weight $w_i(t, h)$ to all the points within the window to adjust for the performance of the polynomial approximation at the boundaries of the window. Define $l_i(\beta_t) = \log f(y_i; \beta_t) = (\theta_i(\beta_t)y_i - b(\theta_i(\beta_t))) / a(\phi) + c(y_i, \phi)$ as the contribution of the i th observation to the log-likelihood centered at t , where $\theta_i(\beta_t) = \beta_{0,t} + \beta_{1,t}(t - t_i) + \beta_{2,t}(t - t_i)^2$. Maximize the local, weighted log-likelihood function

$$\sum_{i=1}^n l_i(\beta_t) w_i(t)$$

with respect to the parameters β_t and, finally, estimate $\frac{\partial^k g(\mu)}{\partial t^k}$ at t by $k! \hat{\beta}_{k,t}$. Repeating these steps for a grid of point in the predictor space gives smoothed estimates of the regression functions and its derivatives.

In this paper, the degree of the polynomial and the bandwidth have been selected by visual examination of the residuals and a local version of the AIC criterion adapted from Loader (1999).

3. Data and results

Human mortality life-table data can be used to compare local likelihood and alternative parametric and nonparametric procedures. We also discuss results on yeast environmental samples which have motivated this work.

Oldest-old human mortality. Recent findings have suggested that, in contrast to what it might be expected, the force of mortality of different species decelerates at the oldest ages. One possible explanation of such a deceleration is heterogeneity of the population and can be tested by fitting life-table data via parametric frailty models. Since a parametric model is already available for this situation, we use this example to compare the results yielded by estimating the mortality hazard function via the frailty model and a local likelihood procedure. Fig. 1 shows the results after fitting the death rates for the cohort of Sweden females born in 1861 via a local binomial regression and the resulting hazard estimation compared with the estimates given by a parametric frailty model and a weighted kernel smoother. The principal result is that the binomial smoother captures better the deceleration estimated by the frailty model than alternative kernel smoothers which do not take into account the binomial character of noise.

Figure 1: left: death rates and local Binomial fit for the cohort of Sweden females born in 1861 (Human Mortality Database - www.mortality.org); right: hazard estimation by the Local Binomial smoother compared to the frailty model (dashed) and a weighted kernel smoother (dots).

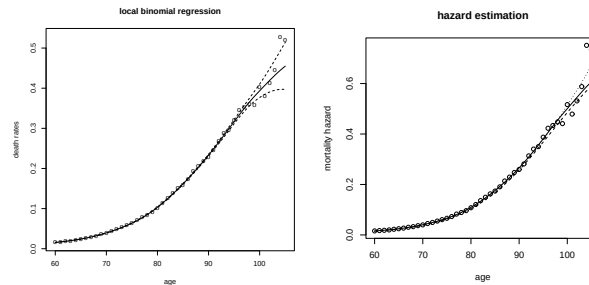
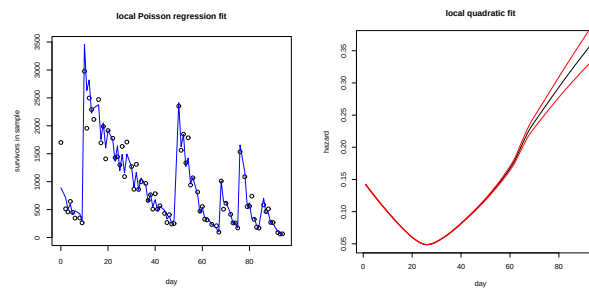


Figure 2: left: yeast cells able to divide in daily samples and local Poisson regression fit; right: hazard estimation with 99% confidence intervals; MPIDR Yeast Lab Database.



Yeast cells mortality. Yeast cells can exist in a form of suspended animation, known as stationary phase, in which they remain alive in the absence of food. They are not able to reproduce in such conditions, but some are able to reproduce again once food is provided. This has led biologists to define a cell as alive if it is able to divide in the presence of food. Notice that, in this case, the mortality rate relates to the ability to divide and the only way to measure the mortality hazard (according to the definition above) is to sample cells from the population, provide them with food and count those able to divide. These data are an example of environmental sampling. Fig. 2 shows the local Poisson fit of the number of cells able to divide in presence of food in one of the experiments considered in this study (the break points in the survival pattern are due to periodic changes in the dilution factor used in the experiments) and the resulting hazard estimation.

References

- Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and its Applications*, Chapman & Hall, London.
- Hastie, T.J. and Tibshirani, R.J. (1987) Local Likelihood Estimation, *JASA*, 82, 559-567.
- Loader, C. (1999) *Local Regression and Likelihood*, Springer, New York.
- Muller, H.-G., Wang, J.-L. and Capra, W.B. (1997) From Lifetables to hazard Rates: the Transformation Approach, *Biometrika*, 84, 881-892.
- Silverman, B.W. (1984) Spline smoothing: the equivalent variable kernel method, *Ann. Statist.*, 12, 898-916