Standardized martingale residuals applied to grouped left truncated observations of dementia cases.

Daniel Commenges, Virginie Rondeau

Lifetime Data Analysis, 6, 229–235, 2000

Standardized Martingale Residuals Applied to Grouped Left Truncated Observations of Dementia Cases

DANIEL COMMENGES daniel.commenges@bordeaux.inserm.fr
INSERM U330, 146 rue Leo Saignat, 33076 Bordeaux, France

VIRGINIE RONDEAU
INSERM U330, 146 rue Leo Saignat, 33076 Bordeaux, France

Received April 13, 1999; Revised September 28, 1999; Accepted December 22, 1999

Abstract. The use of martingale residuals have been proposed for model checking and also to get a non-parametric estimate of the effect of an explanatory variable. We apply this approach to an epidemiological problem which presents two characteristics: the data are left truncated due to delayed entry in the cohort; the data are grouped into geographical units (parishes). This grouping suggests a natural way of smoothing the graph of residuals which is to compute the sum of the residuals for each parish. It is also natural to present a graph with standardized residuals. We derive the variances of the estimated residuals for left truncated data which allows computing the standardized residuals. This method is applied to the study of dementia in a cohort of old people, and to the possible effect of the concentration of aluminum and silica in drinking water on the risk of developing dementia.

Keywords: martingale residuals, survival data, left-truncation, dementia, aluminum

1. Introduction

Martingale residuals are the most natural residuals in the context of survival analysis. Classically, with right censored data, we are interested in modelling the distribution of a variable T_i and we observe $(X_i, D_i), i = 1, \ldots, n$, where $X_i = min(T_i, C_i)$, C_i is a censoring variable and $D_i = 1$ if $X_i = T_i$, 0 otherwise. The martingale residuals are defined as $\tilde{M}_i = D_i - \hat{A}_i(X_i)$, where $\hat{A}_i$ is the Breslow estimator of the cumalitve hazard function of T_i. Score tests in the proportional hazard model are function of these residuals; this is the case of conventional score tests of regression coefficients (Fleming and Harrington, 1991) but also of tests of homogeneity based on random effect models (Commenges and Jacqmin-Gadda, 1997). Also, a graphical use of martingale residuals has been proposed (Barlow and Prentice, 1988; Fleming and Harrington, 1991) for non-parametrically estimating the effect of an explanatory variable on the risk of an event. This is particularly interesting when a non-linear effect is suspected. In epidemiology the interpretation of martingale residuals is easy because they can be interpreted as a number of cases observed minus a number expected.

We apply this approach to an epidemiological study presenting two characteristics: the data are left truncated because of a delayed entry in the cohort; the data are grouped in geographical units, the parishes. This gouping suggests a natural "smoothing" which

consists in computing the sum of the residuals in each parish. It is also natural to present a graph with standardized residuals. We derive the variances of the estimated martingale residuals (in the case of left truncated data) and we give an estimator for these variances; this allows to compute standardized residuals for the parishes.

This technique is applied to the study of the influence of aluminum and silica present in drinking water, on the risk of dementia, based on data of a large cohort study, Paquid. An analysis of prevalent cases of cognitive deficit in the same study has suggested a possible influence of these minerals on cerebral ageing (Jacqmin-Gadda et al., 1996).

2. The Estimated Martingale and its Compensator

We consider a model for survival data; the duration variable is T_i, and the cumulative intensity function is $r_i A_0(t)$, where r_i is the relative risk for subject i; generally we take $r_i = \exp(\beta z_i)$, where z_i is a vector of explanatory variables. This process is not completely observed because observations can be right censored and left truncated. We observe $(X_i, D_i), i = 1, \ldots, n$, where $X_i = min(T_i, C_i)$, where C_i is a censoring variable and $D_i = 1$ if $X_i = T_i$, 0 otherwise. The observed point process ${}_{V_i}N_i(t)$ for each subject i, $i = 1, \ldots, n$ is modelled by

$$d\,{}_{V_i}N_i(t) = dM_i(t) + {}_{V_i}Y_i(t) r_i\, dA_0(t)$$

where ${}_{V_i}Y_i(t) = I_{\{t \le X_i\}} I_{\{t \ge V_i\}}$ where V_i is the time at which the observation starts: this is called a started process (Andersen et al., 1993); V_i is also called a left-truncation variable; we shall treat V_i as fixed.

$M_i(t)$ is a martingale that we can express as a function of ${}_{V_i}N_i(t)$ and of ${}_{V_i}Y_i(t) r_i\, dA_0(t)$. Replacing $A_0(t)$ by its Breslow estimator, we obtain the estimated martingale of this process

$$\hat{M}_i(t) = {}_{V_i}N_i(t) - \int_0^t {}_{V_i}Y_i(s) r_i\, d\hat{A}_0(s),$$

where $\hat{A}_0(t) = \int_0^t \frac{d_V N(s)}{{}_V S^0(s)}$, with $d_V N(s) = \sum_{k=1}^n d_{V_k} N_k$ and ${}_V S^0(s) = \sum_{k=1}^n {}_{V_k} Y_k(s) r_k$; we assume the r_k known.

It is interesting to represent this estimated martingale as a stochastic integral relatively to the original martingales of a predictible process

$$\hat{M}_i(t) = \sum_{j=1}^n \int_0^t H_{ij}(s)\, dM_j(s),$$

with $H_{ij}(s) = \delta_{ij}\, {}_{V_i}Y_i(s) - p_i(s)\, {}_{V_j}Y_j(s)$, $p_i(s) = \frac{{}_{V_i}Y_i(s) r_i}{{}_V S^0(s)}$ and δ_{ij} is the kronecker symbol. This expression shows that $\hat{M}_i(t)$ is a martingale for each i and allows to show easily that

$\sum_{i=1}^{n} \hat{M}_i(t) = 0$. The predictable variation of the estimated martingale is

$$\begin{aligned}\langle \hat{M}_i \rangle(t) &= \sum_j \int_0^t H_{ij}^2(s) d\langle M_j \rangle(s) \\ &= \sum_j \int_0^t H_{ij}^2(s)\, {}_{V_j}Y_j(s) r_j \, dA_0(s) \\ &= \int_0^t (1 - p_i(s))\, {}_{V_i}Y_i(s) r_i \, dA_0(s). \end{aligned} \tag{1}$$

3. Standardized Martingale Residuals

The variance of $\hat{M}_i(t)$ is

$$E(\langle \hat{M}_i \rangle(t)) = \int_0^t E[(1 - p_i(s))\, {}_{V_i}Y_i(s)] r_i \, dA_0(s).$$

We remark that

$$\begin{aligned} E[(1 - p_i(s))\, {}_{V_i}Y_i(s)] &= E[1 - p_i(s) \mid {}_{V_i}Y_i(s) = 1] \Pr[{}_{V_i}Y_i(s) = 1] \\ &= \left\{ 1 - r_i E\left[\frac{1}{{}_{V}S^0(s)} \mid {}_{V_i}Y_i(s) = 1 \right] \right\} E[{}_{V_i}Y_i(s)]. \end{aligned}$$

We have $E[{}_{V_i}Y_i(s)] = E[I_{\{s \le T_i\}} I_{\{s \le C_i\}}] I_{\{s \ge V_i\}} = \pi_i(s) S_i(s) I_{\{s \ge V_i\}}$, where $\pi_i(s)$ is the survival function of the censoring variable and $S_i(s)$ is the survival function of T_i. Since ${}_{V}S^0(s)$ is a sum of many terms we have approximately $E[\frac{1}{{}_{V}S^0(s)} \mid {}_{V_i}Y_i(s) = 1] \approx \frac{1}{E[{}_{V}S^0(s) | {}_{V_i}Y_i(s) = 1]} = \frac{1}{r_i + \sum_{j \ne i} r_j E[{}_{V_j}Y_j(s)]}$

Finally this variance is

$$var \hat{M}_i(t) \approx \int_0^t E[{}_{V_i}Y_i(s)] \left[1 - \frac{r_i}{r_i + \sum_{j \ne i} r_j E[{}_{V_j}Y_j(s)]} \right] r_i \, dA_0(s).$$

The martingale residuals are $\tilde{M}_i = \hat{M}_i(\infty) = D_i - r_i[\hat{A}_0(X_i) - \hat{A}_0(V_i)]$; an estimator of their variance is

$$\widehat{var} \hat{M}_i(\infty)) = \int_0^\infty \hat{\pi}_i(s) \hat{S}_i(s) I_{\{s \ge V_i\}} \left[1 - \frac{r_i}{{}_{V}S^0(s) + r_i(1 - {}_{V_i}Y_i(s))} \right] \frac{r_i}{{}_{V}S^0(s)} \, d_V N(s),$$

where $\hat{S}_i(s) = \exp[-r_i \hat{A}_0(s)]$. Hence we deduce the standardized residuals. More generally a group residual can be defined by $\tilde{M}_C = \sum_{i=1}^{n} I_{\{i \in C\}} \tilde{M}_i$ where C is a subset of $\{1, \ldots, n\}$. Its variance can be computed by the same argument as above, noting that $\tilde{M}_C = \hat{M}_C(\infty)$ where

$$\hat{M}_C(t) = \sum_{j=1}^{n} \int_0^t \sum_{i=1}^{n} I_{\{i \in C\}} H_{ij}(s) dM_j(s).$$

The predictable variation of this martingale can easily be found and an estimator given. However if C is a small subset of $\{1, \ldots, n\}$ the correlations between the $\tilde{M}_i$ can be neglected so that the variance can be approximately computed by the sum of the variances of the $\tilde{M}_i$ leading to the standardized group residual:

$$\tilde{M}_C^S \approx \frac{\sum_{i=1}^n I_{\{i\in C\}} \tilde{M}_i}{\left[\sum_{i=1}^n I_{\{i\in C\}} \widehat{var} \tilde{M}_i\right]^{1/2}}.$$

Note that if (and only if) the number of subjects in a parish $\sum_{i=1}^n I_{\{i\in C\}}$ was large we could obtain a simpler formula by estimating the variance by the observed compensator of $\sum_{i=1}^n I_{\{i\in C\}} \hat{M}_i(t)$, which leads, after replacing $A_0(s)$ in (1) by its Breslow estimator, to

$$\sum_i I_{\{i\in C\}} \int_0^\infty (1 - p_i)\, p_i\, d_V N(s).$$

Heuristically the reason why such an estimator would be good in the case of $\sum_{i=1}^n I_{\{i\in C\}}$ large, stems from the law of large numbers: the sum of a large number of (nearly) independent terms is, relatively, close to its expectation.

The standardized martingale residuals can be used in graphical examination of the fit of the model to the data. The advantage of standardized residuals over ordinary ones is that the different groups are given equal weights which can be safer because of possible intragroup correlation. One use of these residuals is to examine the possible non-linear effect of a variable: the residuals are computed in a model not including the variable of interest; then they are plotted against this variable. A smoothing can be performed in addition to facilitate the interpretation. Kernel smoothing method are simple and allow to compute confidence bands.

4. Application on Dementia in Parishes

The Paquid program on cerebral ageing is based on a large cohort randomly selected in a population of subjects aged 65 years or more, living at home in two departments of southwest France (Gironde and Dordogne). Dementia is one of the major public health problem in this context and Alzheimer's disease represents about two thirds of the cases. The ALMA study, which is a branch of the Paquid program, aims to analyse the role of components of drinking water on the risk of developing Alzheimer's disease and dementia. The etiology of Alzheimer's disease is still unknown, except in a small percentage of cases where a clear genetic factor is present. Thus the research of environmental factors is still ongoing. The hypothesis that aluminum plays a role in Alzheimer's disease has been put forward because aluminum is neurotoxic and it has been shown to cause dementia in dialysed patients (Alfrey et al., 1976); some epidemiological studies have concluded to a relation between aluminum in drinking water and risk of Alzheimer's disease (Martyn, 1989; Jacqmin-Gadda et al., 1996). In addition Birchall and Chappell (1989) have proposed the idea that silicium could protect against aluminum toxicity. The whole issue is very controversial, due in particular to recent negative epidemiological results (Martyn, 1997).

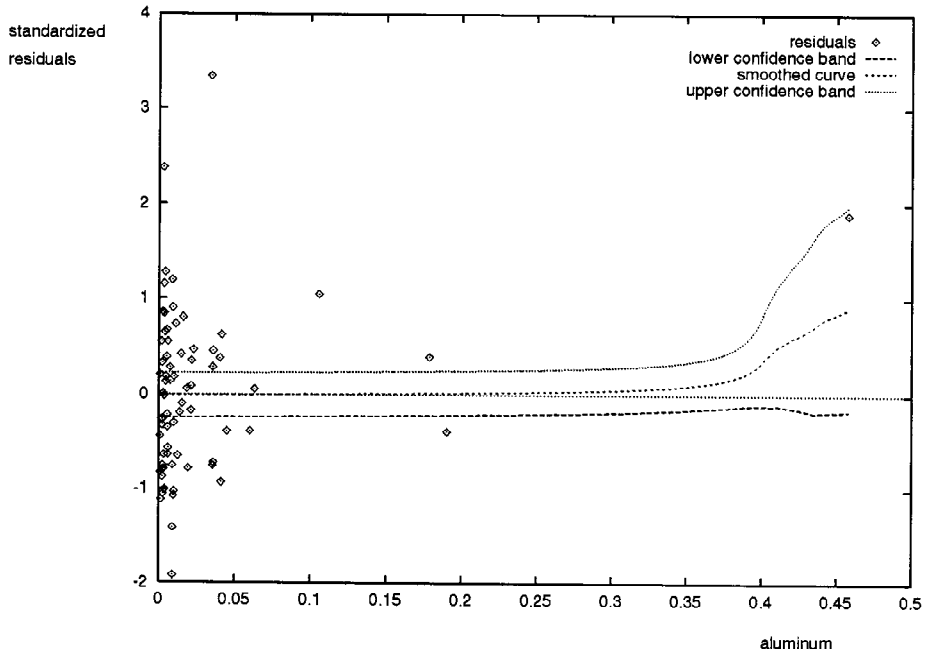

Figure 1.

We present here a preliminary analysis, as an illustration of the graphical methods based on residuals proposed in this paper; 3411 subjects, non-demented at entry in the cohort, scattered in 71 parishes (in Gironde and Dordogne) have been followed during five years and 176 incident cases of dementia have been observed. The concentrations of aluminum, silica and other minerals in the drinking waters of these parishes have been recorded.

The risk of developing dementia is modelled as a function of age, and since prevalent cases are excluded, the data are left truncated: the truncation variable is the age at entry in the cohort. A more complete analysis would involve a three-state model (such as presented in Andersen et al., 1993), but if we are only interested in incidence of dementia and if the age at onset of dementia is known with sufficient precision, the Nelson-Aalen estimator can be used as in an ordinary survival problem (Commenges et al., 1998); here death takes the role of a censoring because it removes the subjects from the set of subjects at risk of developing dementia. As the main occupation during life-time appeared to play a role in the risk of dementia (intellectual occupations having a lower risk than non-intellectual ones), we have computed the standardized residuals for a model including occupation; occupation was considered in three categories, "Intellectual", "Manual", "Farm workers". A Cox proportional hazard model yielded $r_i = 1$ for "Intellectual" occupation (the reference category), $r_i = 1.48$ for "Manual" and $r_i = 2.23$ for "Farm workers".

Only a coarse estimation of the survival function of the censoring variable could be obtained because there is another complication in the data set due to discrete observation

additional data is in preparation.

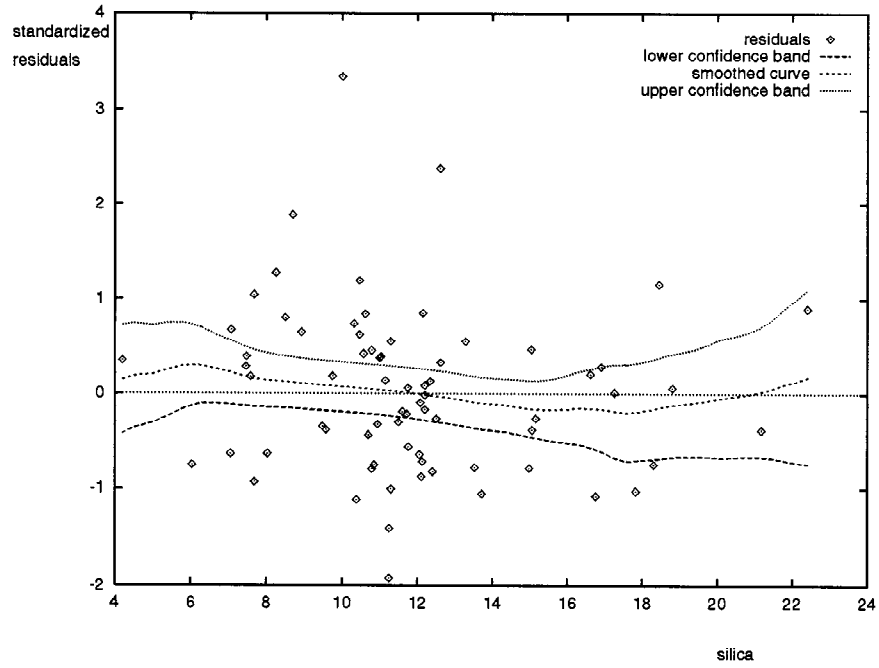

Figure 2.

times leading to interval censoring. This problem is not treated here so that we pretend that observations are in continuous time. We note that the probability of censoring depended essentially of the time of follow-up. We first restricted to subjects seen at least at one follow-up visit because other subjects carry no information. Among these subjects the proportion seen at the visit done around 36 months (resp. 60) was 0.88 (resp. 0.71); the last visit was done after 73 months; hence the estimated distribution of the censoring was $\hat{\pi}_i(s) = \phi(s - V_i)$ with $\phi(u) = 1$ for $u < 15$, $= 0.88$ for $15 \leq u < 50$, $= 0.71$ for $50 \leq u < 74$, $= 0$ for $u \geq 74$, where the time is expressed in months.

Examination of a plot of the residuals against the aluminum concentrations (Figure 1) makes it clear that one parish will be very influent on the result of a formal test on the effect of aluminum, because it has a high aluminum concentration and a relatively large positive residual (around 2). The other parishes do not seem to bring much evidence of a link beween the risk of the disease and aluminum concentration. We have added a smooth estimate of the residuals using a kernel smoothing technique based on Epanechnikov kernels, together with pointwise 95%-confidence bands.

A plot of the same residuals against concentration of silica (Figure 2) let appear a slight trend of positive residuals for concentrations below 11mg/l and negative residuals for values

larger than 11mg/l. This is in favor of a protective effect of silicium and it is apparent that this is a global trend not due to one particular parish. It happens that the tests of the regression coefficients in a Cox model are significant for both aluminum and silica treated as dichotomized variables with cut-off points equal to 0.1 mg/l for aluminum (a value aready used in the literature) and to 11.25 mg/l for silica (the median). The plots of residuals downgrades the impression that aluminum plays a role; however the result of the test on silica is rather supported by the examination of the residuals although there is no apparent dose-effect relationship. Finally we note that these findings go in the direction predicted by Birchall's hypothesis. However they are not conclusive; a more detailed study involving additional data is in preparation.

Acknowledgment

We thank Laurent Bordes for helpful discussion about the computation of the variance of the residuals.

References

A. C. Alfrey, G. R. Legendre, W. D. Kaehny, "The dialysis encephalopathy syndrome: possible aluminum intoxication," *N Engl J Med* vol. 294 pp. 184–188, 1976.

P. K. Andersen, Ø. Borgan, R. D. Gill and N. Keiding, *Statistical Models Based on Counting Processes*, Springer-Verlag: New York, 1993.

W. E. Barlow, R. L. Prentice, "Residuals for relative risk regression," *Biometrika* vol. 75 pp. 65–74, 1988.

J. D. Birchall, J. S. Chappell, "Aluminum, water chemistry and Alzheimer's disease," *Lancet* 1989; i pp. 953, 1989.

D. Commenges, L. Letenneur, P. Joly, A. Alioum and J. F. Dartigues, "Modelling age-specific risk: application to dementia," *Stat Med* vol. 17 pp. 1973–1988, 1998.

D. Commenges and H. Jacqmin-Gadda, "Generalized score test of homogeneity based on correlated random effects model," *J Roy Stat Soc B* vol. 59 pp. 157–171, 1997.

T. R. Fleming and D. P. Harrington, *Counting Processes and Survival Analysis*, Wiley: New-York, 1991.

H. Jacqmin–Gadda , D. Commenges, L. Letenneur, J. F. Dartigues, "Silica and aluminum in drinking water and cognitive impairment in the elderly," *Epidemiology* vol. 7 pp. 281–285, 1996.

C. N. Martyn, D. J. Barker, C. Osmond, E. C. Harris, J. A. Edwardson, R. F. Lacey, "Geographical relation between Alzheimer's disease and aluminum in drinking water," *Lancet* vol. 1; i pp. 59–62, 1989.

C. N. Martyn, D. N. Coggon, H. Inskip, R. F. Lacey, W. F. Young, "Aluminum concentrations in drinking water and risk of Alzheimer's disease," *Epidemiology* vol. 8 pp. 281–6, 1997.