PLISKA
STUDIA MATHEMATICA BULGARICA
ПЛИСКА
БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

For further information on
Pliska Studia Mathematica Bulgarica
visit the website of the journal http://www.math.bas.bg/~pliska/
or contact: Editorial Office
Pliska Studia Mathematica Bulgarica
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Telephone: (+359-2)9792818, FAX:(+359-2)971-36-49
e-mail: pliska@math.bas.bg

Pliska Stud. Math. Bulgar. **17** (2005), 5–12

PLISKA
STUDIA MATHEMATICA
BULGARICA

STUDY ON ROBUSTNESS OF CORRELATED FRAILTY MODEL

Dimitar Atanasov

This study considers a robust properties of correlated frailty models. The dependence between related individuals must be considered in order to be studied the difference between the gene information and the environment as causes of death. To do that, one can introduce the frailty parameter Z, which can be decomposed as $Z = Z_g + Z_e$, where Z_g represents the frailty, due to the gene information, and Z_e represents the influence of the environment. Using the $WLTE(k)$ one can obtain a robust maximum likelihood estimation of the unknown parameters of the model.

1. Introduction

This study considers a robust modification of correlated frailty models. Correlated frailty is a generalization of the Shared Frailty models which are very useful in studying the mortality of related individuals (ex. twins).

There are a lot of reasons for development of models like Frailty Models. Obviously, the assumption that the individuals are identical is not true. Furthermore, the difference, the heterogeneity and the dissimilarities between the individuals are some of the most important laws of nature.

Let us suppose that the mortality of the individuals in the population depends on an unknown variable Z: $\mu(x \mid Z) = Z\mu_0(x)$. It is reasonable to suppose that Z is Gamma distributed with $EZ = 1$. On the one hand, this allows us to estimate the survival function, on the other hand, by changing the parameters of the Gamma distribution we can handle quite a large class of probability distributions.

2000 *Mathematics Subject Classification:* 62J12
Key words: frailty models, robust maximum likelihood estimation
Partialy supported by Pro-ENBIS GTC1 -2001-43031

The dependence between related individuals must be considered in order to be studied the difference between the gene information and the environment as causes of death. To do that, we can decompose the frailty $Z = Z_g + Z_e$, where Z_g represents the frailty due to the gene information and Z_e represents the influence of the environment. As Z is Gamma distributed we can assume that Z_g and Z_e are also Gamma distributed.

We can extend this concept using the idea that not all, but almost all of the individuals have the same distribution of the frailty parameter. Therefore, using the $WLTE(k)$ we can obtain a robust maximum likelihood estimation of the unknown parameters of the model in the case when there are some outliers. Moreover, we can use the trimming parameter k to study if there are any subpopulations in the data.

2. The Correlated Frailty Model

The Shared Frailty Models, as their name tells, state that both individuals share the same frailty Z. Obviously this assumption leads to restriction in interpretation of the dependence between the individuals in the couple. To overcome this the Correlated Frailty Models were introduced (Yashin, Vapuel and Iachine, 1995c). The model can be represented in the following way.

Let us have the times of death of two related individuals T_1, T_2, and the frailties of these two individuals X_1, X_2. We can decompose them as:

$$X_1 = \frac{\lambda_0}{\lambda_1} Z + Z_1$$

$$X_2 = \frac{\lambda_0}{\lambda_2} Z + Z_2,$$

where $Z \in \Gamma(k_0, \lambda_0), Z_1 \in \Gamma(k_1, \lambda_1), Z_2 \in \Gamma(k_2, \lambda_2)$ and $EX_1 = EX_2 = 1$, $DX_1 = \frac{1}{\lambda_1} = \sigma_1^2$, $DX_1 = \frac{1}{\lambda_1} = \sigma_1^2$. If $Z_i = 0, i = 1, 2$ we obtain the Shared Frailty Model. We can assume that all the information about common genes and common environment is represented by Z, and $Z_i, i = 1, 2$ represent only the difference between the individuals and between their environments. Therefore, if there is any correlation between the individuals, it will be represented by Z. So, we can assume that $Z, Z_i, i = 1, 2$ are independent.

According to Weinke (2001) the unconditional survival function for this model is

$$S(x_1, x_2) = ES(x_1, x_2 \mid Z_1, Z_2) = ES(x_1 \mid Z_1) S(x_2 \mid Z_2) =$$

$$= (1 + \sigma_1^2 H(x_1) + \sigma_2^2 H(x_2))^{-\frac{\rho}{\sigma_1 \sigma_2}} (1 + \sigma_1^2 H(x_1))^{-\frac{1 - \frac{\sigma_1}{\sigma_2}\rho}{\sigma_1^2}} (1 + \sigma_2^2 H(x_2))^{-\frac{1 - \frac{\sigma_2}{\sigma_1}\rho}{\sigma_2^2}},$$

where $\rho = \frac{k_0}{\lambda_1 \lambda_2}$ and $H(x) = \int_0^x \mu_0(u)du$ is the cumulative hazard function. We assume that the hazard function follows the Gompertz law $\mu_0(x) = ae^{bx}$, but it is not difficult to extend the model for $\mu_0(x) = ae^{bx} + c$.

The parameters of the model can be estimated using the maximum likelihood model. As the density function

$$f(x_1, x_2) = \frac{\partial^2}{\partial x_1 \partial x_2} S(x_1, x_2)$$

is a very complicated one, it was obtained by symbolic calculations with MATLAB.

One important point, that must be considered when applying to a real data is cencoring of the data. It occurs when some of the observations are not times of death, but moments when the individuals leave the population. Another restriction which comes with the real data is truncation. It means that we observe the times of death of these individuals that are alive at the moment when the study begins. Let us suppose that there is a right censoring and a left truncation. We have to calculate the density function in this case. Let the observations consist of $(T_1, T_2, \Delta_1, \Delta_2)$, where $T_i = min\{D_i, Y_i\}$ and $\Delta_i = \delta(Y_i - D_i)$, δ is the Heviside function. Here D_i are the times of death and Y_i are the moments of censoring. Therefore, there are four cases of cumulative densities functions

$$P(T_1 > t_1, T_2 > t_2, d_1 = 1, d_2 = 1) = \int_0^{t_1} \int_0^{t_2} \int_{t_1}^{\infty} \int_{t_2}^{\infty} f(x_1, x_2) g_1(y_1) g_2(y_2) dy_2 dy_1 dx_2 dx_1$$

$$= \int_0^{t_1} \int_0^{t_2} f(x_1, x_2)(1 - G_1(x_1))(1 - G_2(x_2)) dx_2 dx_1,$$

where $g_i(x)$ and $G_i(x)$ are the density and the cumulative density of the censoring times.

$$P(T_1{>}t_1, T_2{>}t_2, d_1{=}1, d_2{=}0){=} \int_0^{t_1} \int_{t_1}^{\infty} \int_0^{t_2} \int_{y_2}^{\infty} f(x_1, x_2) g_1(y_1) g_2(y_2) dx_2 dy_2 dy_1 dx_1$$

$$= \int_0^{t_1} \int_0^{t_2} \frac{\partial}{\partial x_1}(1 - F(x_1, y_2))(1 - G_1(x_1)) g_2(y_2) dy_2 dx_1.$$

By analogy

$$P(T_1{>}t_1, T_2{>}t_2, d_1{=}0, d_2{=}1){=}\int_0^{t_1}\int_0^{t_2}\frac{\partial}{\partial x_2}(1-F(y_1,x_2))(1-G_2(x_1))g_1(y_2)dy_1dx_2.$$

In the last case

$$P(T_1{>}t_1, T_2{>}t_2, d_1{=}0, d_2{=}0){=}\int_0^{t_1}\int_0^{t_2}\int_{t_1}^{\infty}\int_{t_2}^{\infty} f(x_1,x_2)g_1(y_1)g_2(y_2)dx_2dx_1dy_2dy_1$$

$$=\int_0^{t_1}\int_0^{t_2} S(x_1,x_2)g_1(y_1)g_2(y_2)dy_1dy_2$$

Therefore, the density function will be a mixture of the densities

$$f(x_1,x_2,d_1,d_2)=(f(x_1,x_2)(1-G_1(x_1))(1-G_2(x_2)))^{d_1d_2}\times$$

$$\times(\frac{\partial}{\partial x_1}(1-F(x_1,x_2))(1-G_1(x_1))g_2(x_2))^{d_1(1-d_2)}\times$$

$$\times(\frac{\partial}{\partial x_2}(1-F(x_1,x_2))(1-G_2(x_1))g_1(x_2))^{(1-d_1)d_2}(S(x_1,x_2))^{(1-d_1)(1-d_2)}.$$

As the censoring is not informative, in the likelihood function only the terms which give information will be included. So, the likelihood curve will be

$$l(x_1,x_2,d_1,d_2)=f(x_1,x_2)^{d_1d_2}\frac{\partial}{\partial x_1}(1-F(x_1,x_2))^{d_1(1-d_2)}\times$$

$$\times\frac{\partial}{\partial x_2}(1-F(x_1,x_2))S(x_1,x_2)^{(1-d_1)(1-d_2)}.$$

If there is a left truncation we can use the same density function, but only if the individuals have survived the truncation moment T^*.

$$P(T_1>t_1,T_2>t_2,d_1,d_2\mid T_1>T^*,T_2>T^*)=$$

$$P(T_1>t_1,T_2>t_2,d_1,d_2,T_1>T^*,T_2>T^*)=\frac{f(x_1,x_2,d_1,d_2)}{S(y_1,y_2)},$$

where $y_i=T^*-b_i$. Here with b_i denotes the time of birth. Therefore, for the truncated model, the likelihood function becomes

$$l(x_1,x_2,d_1,d_2\mid T_1>T^*,T_2>T^*)=\frac{l(x_1,x_2,d_1,d_2)}{S(y_1,y_2)}. \tag{1}$$

3. Statistical model and robustness

Applying this model to a real data it turns out that the optimization algorithm strongly depends from the starting point of the optimization procedure. For some values of the estimated parameters for a certain observations the probability density function becomes negative and the log - likelihood goes to the complex plane.

So we came to the idea to use a robustified maximum likelihood estimator in order to filter out these observations, which for a given value of the estimated parameters give a negative value for the density function.

A robust extension of the maximum likelihood estimators (MLE) that possesses a high breakdown point was introduced by Vandev and Neykov (1993). This modification considers the likelihood of individual observations as residuals and applies the basic idea of the LTS estimators of Rousseeuw (1984) using appropriate weights. In this way Vandev and Neykov (1993) and put in a general framework many kinds of statistical estimators and in particular the $LME(k)$ and $LTE(k)$ estimators, previously proposed by Neykov and Neytchev (1990), and studied by Vandev (1993) and Vandev and Neykov (1993).

Generally speaking, Vandev and Neykov (1998) defined the $WLTE(k)$ estimators, $\widehat{\theta}$, for the unknown parameter $\theta \in \Theta^p$ as

$$\widehat{\theta} = \underset{\theta \in \Theta}{argmin} \sum_{i=1}^{k} w_i f_{\nu(i)}(\theta),$$

where $f_{\nu(1)}(\theta) \leq f_{\nu(2)}(\theta) \leq \cdots \leq f_{\nu(n)}(\theta)$ are the ordered values of $f_i = -\log \varphi(x_i, \theta)$ at θ, $\varphi(x_i, \theta)$ is a probability density, θ is an unknown parameter and $\nu = (\nu(1), \cdots, \nu(n))$ is the corresponding permutation of the indices, which may depend on θ. The weights $w_i \geq 0$, $i = 1, \cdots, k$, are such that an index $k = \max\{i : w_i > 0\}$ exists.

Vandev and Neykov (1998) proved that the finite sample breakdown point of the $WLTE(k)$ estimators is not less than $(n-k)/n$ if $n \geq 3d$, $(n+d)/2 \leq k \leq n-d$, when Θ^p is a topological space and the set $F = \{f_i(\theta), i = 1, \cdots, n\}$ is d-full. We remind the reader that a finite set F of n functions is called d-full, according to Vandev (1993), if for each subset of cardinality d of F, the supremum of this subset is a subcompact function. A real valued function $g(\theta)$ is called subcompact, if its Lesbegue sets $L_g(C) = \{\theta : g(\theta) \leq C\}$ are compact for any constant C (see Vandev and Neykov, 1993).

For the sake of completeness, we draw the attention to the fact that the finite sample breakdown point of an estimator T, at the finite sample $X = \{x_i; i =$

$1, \cdots, n\}$, is defined as the largest fraction m/n for which the

$$\sup_{\tilde{X}} \left\| T(X) - T\left(\tilde{X}\right) \right\|$$

is finite, where $\tilde{X}$ is a sample obtained from X by replacing any m of the points in X by arbitrary values (see Hampel et al. 1986, Rousseeuw and Leroy, 1987).

Thus, if one wants to study the breakdown point of the $WLTE(k)$ estimators for a particular distribution, one has to find out the index d of fullness of the corresponding set of log-density functions.

According to Atanasov and Neykov (2001) if D is an open subset of R^n, θ_0 belongs to the boundary of D and $g(\theta)$ is a real valued continuous function defined on D, then we have the following theorem.

Theorem 1. *The function $g(\theta)$ is subcompact if and only if for any sequence $\theta_i \to \theta_0$ $g(\theta_i) \to \infty$ when $i \to \infty$.*

Remark: If D is a compact set, then any continuous function defined on D is subcompact.

We will apply this for the likelihood function described above. This function depends on 5 unknown parameters: σ_1, σ_2, ρ are parameters describing the Gamma distributed frailty and a, b represent the parameters of the Gompertz hazard function.

Using the symbolic calculations of MATLAB we can find out that if we study all five unknown parameters there is no set of likelihood curves that satisfis the conditions of d-fullness theory. But if one considers the parameters of the Gamma distributed frailty variable (the parameters of the hazard function are fixed to a constant) one can find out that the index of fullness of the set $F = \{f_i(\sigma_1, \sigma_2, \rho), i = 1, \cdots, n\}$ is 3. Here $\{f_i\}_{i=1}^n$ are logarithms of the likelihood functions defined with (1).This result was expected, according to Atanasov and Neykov (1999). They show that the index of fullness of the set of log-likelihood curves for Gamma distributed observations is equal to 2.

This allows us to study the correlation coefficient between the frailty variables in the case when there are some outliers in the data. Also it is possible, using weights, to find out if there are subpopulations in the observed population (Atanasov,2002). If the weights are calculated during the minimization algorithm we will obtain different weights for the observations of the different subpopulations.

The likelihood curves are not subcompact on the mortality parameters. So there is no subset of any cardinality which satisfies the theory of d-fullness in this

case. Therefore, in order to obtain a robust estimator on them we have to restrict their values in compacts. Then, according to the remark above, the likelihood function will be subcompact as it is continuous.

The maximum likelihood estimator can be defined as

$$\widehat{\theta} = \underset{\theta\in\Theta}{argmin} \sum_{i=1}^{k} w_i(-logl_{\nu(i)}(x^1_{\nu(i)}, x^2_{\nu(i)}, d^1_{\nu(i)}, d^2\nu(i), \theta \mid x^1_{\nu(i)} > T^*, x^2_{\nu(i)} > T^*)),$$

where $\theta = (\sigma_1, \sigma_2, \rho, a, b)$ in the case when all parameters are studied or $\theta = (\sigma_1, \sigma_2, \rho)$ in the case when only frailty parameters are studied.

According to the theory of d-fullness the breakdown point properties of this estimator are not less than $(n-k)/n$ if $n \geq 3d$, $(n+d)/2 \leq k \leq n-d$, where $d = 5$ in the case when all parameters studied and $d = 3$ in the other case.

REFERENCES

[1] Atanasov, D. V. About the Concept of Weights of $WLTE(k)$ Estimators. In: Seminar on Statistical Data Analysis 2002.

[2] Atanasov, D. V., N. M. Neykov About the Finite Sample Breakdown Point of the WLTE(k) Estimators. In: Proceedings of the XXV Summer School Sozopol'99, (eds. Cheshankov, B.I. Todorov, M.D.), Heron Press, Sofia, (2000), 105–106.

[3] Atanasov, D. V., N. M. Neykov On the Finite Sample Breakdown Point of the $WLTE(k)$ and d-fullness of a Set of Continuous Functions. In: Proceedings of the VI International Conference "Computer Data Analysis And Modeling", Minsk, Belarus, (2001).

[4] Hampel F. R., E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel Robust Statistics: The Approach Based on Influence Functions, John Wiley and Sons, New York, 1986.

[5] Neykov, N. M., P. N. Neytchev A Robust Alternative of the Maximum Likelihood Estimators. COMPSTAT 1990, Short Communications, 99–100.

[6] Vandev, D. L. A Note on Breakdown Point of the Least Median of Squares and Least Trimmed Estimators. *Statistics and Probability Letters.* **16** (1993), 117–119.

[7] Vandev, D. L., N. M. Neykov Robust Maximum Likelihood in the Gaussian Case. In: New Directions in Statistical Data Analysis and Robustness. (eds. S. Morgenthaler, E. Ronchetti, and W.A. Stahel). Birkhauser Verlag. Basel., 1993.

[8] Vandev, D. L., N. M. Neykov About Regression Estimators with High Breakdown Point. *Statistics* **32** (1998), 111–129.

[9] Wienke, A. Frailty Models in Survival Analysis. Max Planck Institute for Demographic Research, 2001.

[10] Wienke, A., Christensen K., Skytthe A. and Yashin A. Genetic analysis of cause of death in a mixture model of bivariate lifetime data. (2002).

[11] Yashin, A. I., I. A. Iachine Genetic Analysis of Durations: Correlated Frailty Model Applied to Survival of Danish Twins. *Genetic Epidemiology* **12** (1995a), 529–538.

[12] Yashin, A. I., I. A. Iachine Survival of Related Individuals: An Extension of some Fundamental Results of Heterogeneity Analysis. *Mathematical Population Studies* **5** (1995b), 321–339.

[13] Yashin, A. I., J. W. Vaupel, I. A. Iachine Correlated Individual Frailty: An Advantageous Approach to Survival Analysis of Bivariate data. *Mathematical Population Studies* **5** (1995c), 145–159.

Dimitar Atanasov
Faculty of Mathematics and Informatics
Sofia University, 5 J. Boucher Str., 1407 Sofia, Bulgaria
e-mail: datanasov@fmi.uni-sofia.bg