行政院國家科學委員會專題研究計畫成果報告

用 Gamma 過程分析單調長期追蹤資料

Analyzing Monotone Longitudinal Data via Gamma Processes

計畫編號：NSC 90-2118-M-009-009-
執行期限：90 年 8 月 1 日至 91 年 7 月 31 日

主持人：陳志榮　　國立交通大學統計學研究所
E-mail：cchen@stat.nctu.edu.tw

一、中文摘要

在這篇文章中，我們提出一般變換路徑模型來建構長期追蹤資料。我們用修正冪變換族和修正指數變族換當例子來闡明我們的模型。最後，我們討論相對應的估計方法。

關鍵詞：長期追蹤資料、一般路徑模型、變換、不等變異數、修正冪變換、修正指數變換、故障前時間分布。

Abstract

In this paper, the general transformed path (GTP) model with heteroscedastic errors for modeling longitudinal data is proposed. The families of modified power transformations and modified exponential transformations are used as examples to illustrate our models. The corresponding estimation method is also discussed.

Keywords: longitudinal data, general path model, transformation, heteroscedasticity, modified power transformation, modified exponential transformation, time-to-failure distribution.

二、緣由與目的

In this paper, we are mainly concerned with modeling longitudinal data. In this section, we briefly introduce the general path (GP) model (Lu and Meeker, 1993) and the transformation model (e.g., see Atkinson and Cox, 1988 and Taylor, 1998) with heteroscedastic errors to proceed the discussion.

For modeling degradation data, Lu and Meeker (1993) proposed the following GP model: For each test unit i in a random sample of size n from the population of all units, assume that degradation measurements are available for prespecified times $t(1)$, $t(2)$, …, $t(m(i))$, where $t(j)$ is the time of the jth measurement or inspection on test unit i for j = 1, 2, …, $m(i)$ and $m(i)$ is the total number of inspections on test unit i. For each test unit i, let $x_{i,t} \geq 0$ be its actual path for $t \geq 0$. Then the sample path $y_{i,t(1)}, y_{i,t(2)}, \ldots, y_{i,t(m(i))}$ of test unit i for $t = t(1), t(2), \ldots, t(m(i))$ is given by

$$\begin{aligned} h(y_{i,t(j)}) &= h(x_{i,t(j)}) + \text{£}_{i,t(j)} \\ &= \text{«}(t(j);\ \ _0,\ \ _i) + \text{£}_{i,t(j)}, \end{aligned}$$

j = 1, 2, …, $m(i)$; i = 1, 2, …, n, where h is a known strictly monotone transformation, e.g., the log transformation, $\text{£}_{i,t(j)}$'s are *i.i.d.* $N(0,\pm^2)$ measurement errors with unknown variance $\pm^2 > 0$, « is a known regression function indexed by a fixed-effects parameter vector $\ _0$ and a random-effects parameter vector $\ _i$, $\ _i$'s are independent of $\text{£}_{i,t(j)}$'s and *i.i.d.* $N(0,p\ )$ with unknown covariance matrix p , and $\text{«}(t;\ \ _0,\ \ _i) = h(x_{i,t})$ for $t \geq 0$. Let h^{-1} be the inverse function of h and set $(h^{-1})^{(k)}(v) = \partial^k h^{-1}(v)/\partial v^k$ for k = 1, 2, …. Then, for each (i,j) pair, $x_{i,t(j)} = h^{-1}(\text{«}(t(j);\ \ _0,\ \ _i))$ is the conditional median of $y_{i,t(j)}$ given $\ _i$, but in general not the conditional mean of $y_{i,t(j)}$ given $\ _i$. Since all measurement errors are normally distributed, the range of the transformation is $\boldsymbol{R}$, where $\boldsymbol{R} = (-\infty,\infty)$. In the literature, the most commonly used transformation for positive data with this property is the log transformation. In such a situation, for each (i,j) pair, $x_{i,t(j)}$ = exp(«($t(j)$; $_{0}$, $_{i}$)) is the conditional median of $y_{i,t(j)}$ given $_{i}$, but less than exp(±^{2}/2)·$x_{i,t(j)}$, the conditional mean of $y_{i,t(j)}$ given $_{i}$.

For modeling independent and continuous data, the transformation model with heteroscedastic errors is as follows:

$$h(y_i; {}^{a}) = \text{«}(x_i;\ \ ) + g(\text{«}(x_i;\ \ ), z_i; \text{¥})\cdot\text{£}_i,$$

i = 1, 2, …, n, where y_i is the observation for subject i, h is a known strictly monotone transformation indexed by a transformation parameter vector ^{a}, both x_i and z_i are known covariates of subject i, « is a known regression function indexed by a regression parameter vector , g is a known positive weight function indexed by a variance parameter vector ¥, and £$_i$'s are *i.i.d.* N(0,1) standardized errors. Since all standardized errors are normally distributed, the range of the transformation is ***R***. For each fixed ^{a}, let $h^{-1}(\cdot; {}^{a})$ be the inverse function of $h(\cdot; {}^{a})$ and set $(h^{-1})^{(k)}(v; {}^{a}) = \partial^k h^{-1}(v; {}^{a})/\partial v^k$ for k = 1, 2, …. Then, for each i, h^{-1}(«(x_i;); ^{a}) is the median of y_i, but in general not the mean of y_i.

In the literature, the classical likelihood inference in transformation models for continuous data is usually based on the key assumption that transformed data are normally distributed, e.g., the transformation model with heteroscedastic errors. However, this assumption would fail for the case where the range of the transformation is not ***R***. As an example, the family of modified power transformations (Box and Cox, 1964),

$$h(u; {}^{a})$$

$$= 1_{\boldsymbol{R}\backslash\{0\}}({}^{a})\cdot(u^{a}-1)/{}^{a} + 1_{\{0\}}({}^{a})\cdot\log(u) \equiv u^{({}^{a})},$$

u > 0, is most commonly used to transform positive data, where $1_S({}^{a})$ = 1 if ^{a} ∈ S and 0 if ^{a} ∈ $\boldsymbol{R}\backslash S$ for any subset S of ***R***. By applying a modified power transformation to positive data, the range of the transformation is not ***R*** except for the log transformation. As another example, the family of modified exponential transformations,

$$h(u; {}^{a})$$

$$= 1_{\boldsymbol{R}\backslash\{0\}}({}^{a})\cdot[\exp({}^{a}\cdot u)-1]/{}^{a} + 1_{\{0\}}({}^{a})\cdot u \equiv u^{[{}^{a}]},$$

u ∈ ***R***, can be used to transform real-valued data. By applying a modified exponential transformation to real-valued data, the range of the transformation is not ***R*** except for the identity transformation. Furthermore, when the range of the transformation is not ***R***, the standardized errors in the transformation model with heteroscedastic errors have possibly different supports and thus are not necessarily identically distributed. One way to tackle this problem is only to assume that all standardized errors in the transformation model with heteroscedastic errors have mean 0 and variance 1, but not necessarily normally and/or identically distributed.

三、結果與討論

In this section, we propose the GTP model with heteroscedastic errors for modeling longitudinal data as follows: For each subject i in a random sample of size n from the population of all possible subjects, assume that repeated measurements are available for prespecified times $t(i,1)$, $t(i,2)$, …, $t(i,m(i))$, where $t(i,j)$ is the time of the jth measurement of subject i for j = 1, 2, …, $m(i)$ and $m(i)$ is the total number of repeated measurements of subject i. For each subject i, let {$x_{i,t}$: t ≥ 0} be its actual path for t ≥ 0. Then the sample path $y_{i,t(i,1)}$, $y_{i,t(i,2)}$, …, $y_{i,t(i,m(i))}$ of subject i for t = $t(i,1)$, $t(i,2)$, …, $t(i,m(i))$ is given by

$$h(y_{i,t(i,j)}; {}^{a}_0, {}^{a}_i)$$

$$= h(x_{i,t(i,j)}; {}^{a}_0, {}^{a}_i) + \text{£}_{i,t(i,j)}{}^*$$

$$= \text{«}(t(i,j);\ _0,\ _i)$$

$$+ g(\text{«}(t(i,j);\ _0,\ _i), t(i,j); \text{¥}_0, \text{¥}_i)\cdot\text{£}_{i,t(i,j)},$$

j = 1, 2, …, $m(i)$; i = 1, 2, …, n, where h is a known strictly monotone transformation indexed by a fixed-effects parameter vector $^{a}{}_0$ and a random-effects parameter vector $^{a}{}_i$, $^{a}{}_i$'s are *i.i.d.* with mean 0 and unknown covariance matrix p_{a}, £$_{i,t(j)}$*'s are independent errors with mean 0, « is a known positive regression function indexed by a fixed-effects parameter vector $_0$ and a random-effects parameter vector $_i$, $_i$'s are *i.i.d.* with mean 0 and unknown covariance matrix p , «(t; $_0$, $_i$) = $h(x_{i,t}; {}^{a}_0, {}^{a}_i)$ for t ≥ 0, g is a known positive weight function indexed by a fixed-effects parameter vector ¥$_0$ and a random-effects parameter vector ¥$_i$, ¥$_i$'s are *i.i.d.* with mean 0 and unknown covariance matrix $p_{¥}$, and $£_{i,t(i,j)}$'s are independent standardized errors with mean 0 and variance 1, and independent of $_i$'s, $ª_i$'s and $¥_i$'s.

For each fixed ($ª_0$, $ª_i$) pair, let $h^{-1}(\cdot; ª_0, ª_i)$ be the inverse function of $h(\cdot; ª_0, ª_i)$ and set $(h^{-1})^{(k)}(v; ª_0, ª_i) = \partial^k h^{-1}(v; ª_0, ª_i)/\partial v^k$ for k = 1, 2, …. Note that, for each (i,j) pair, both the conditional mean and median of $y_{i,t(i,j)}$ given $_i$, $ª_i$ and $¥_i$ are in general unavailable.

As an example,

$$h(u; ª_0, ª_i) = u^{(ª\_0+ª\_i)},\ u > 0,$$

$$«(t;\ _0,\ _i) = (\ _{01}+\ _{i1}) + (\ _{02}+\ _{i2})\cdot t,\ t \ge 0,$$

and

$$g(«(t;\ _0,\ _i),t;¥_0,¥_i)$$
$$= \exp\{¥_{01}+¥_{02}\cdot\log[«(t;\ _0,\ _i)]+¥_i\},\ t \ge 0,$$

i = 1, 2, …, n, can be used for modeling positive longitudinal data, where $_0$ = ($_{01}$, $_{02}$) and $¥_0 = (¥_{01},¥_{02})$. As another example,

$$h(u; ª_0, ª_i) = u^{[ª\_0+ª\_i]},\ u \in \boldsymbol{R},$$

$$«(t;\ _0,\ _i) = (\ _{01}+\ _{i1}) + (\ _{02}+\ _{i2})\cdot t,\ t \ge 0,$$

and

$$g(«(t;\ _0,\ _i),t;¥_0,¥_i)$$
$$= \exp[¥_{01}+¥_{02}\cdot|«(t;\ _0,\ _i)|+¥_i],\ t \ge 0,$$

i = 1, 2, …, n, can be used for modeling real-valued longitudinal data, where $_0$ = ($_{01}$, $_{02}$) and $¥_0 = (¥_{01},¥_{02})$.

In the following, we propose an estimation method for all parameters and a time-to-failure distribution in the GTP model with heteroscedastic errors as follows:

To simplify the notation, for each (i,j) pair and k = 1, 2, …, set $\ = (\ _0{}^T,\ _1{}^T,\ldots,\ _n{}^T)^T$, $ª = (ª_0{}^T, ª_1{}^T,\ldots, ª_n{}^T)^T$, $¥ = (¥_0{}^T,¥_1{}^T,\ldots,¥_n{}^T)^T$, $¯ = (\ ^T, ª^T,¥^T)^T$, $«_{ij}(\ ) = «(t(i,j);\ _0,\ _i)$, $h_{ij}(ª) = h(y_{i,t(i,j)}; ª_0, ª_i)$, $g_{ij}(\ ,¥) = g(«_{ij}(\ ),t(i,j);¥_0,¥_i)$, $H_{ij1}(¯) = h_{ij}(ª) - «_{ij}(\ ,ª)$, $H_{ij2}(¯) = H_{ij1}{}^2(¯) - g_{ij}{}^2(\ ,¥)$, $H_{ij}(¯) = (H_{ij1}(¯),H_{ij2}(¯))^T$, $h^{(k)}(u; ª_0, ª_i) = \partial^k h(u; ª_0, ª_i)/\partial u^k$, $g^{(k)}(u,t(i,j);¥_0,¥_i) = \partial^k g(u,t(i,j);¥_0,¥_i)/\partial u^k$, $h_ª(u; ª_0, ª_i) = \partial h(u; ª_0, ª_i)/\partial ª$, $h_ª{}^{(k)}(u; ª_0, ª_i) = \partial^k h_ª(u; ª_0, ª_i)/\partial u^k$, $h_0(v; ª_0, ª_i) = h_ª(h^{-1}(v; ª_0, ª_i); ª_0, ª_i)$, $h_0{}^{(0)}(v; ª_0, ª_i) = h_0(v; ª_0, ª_i)$ and $h_0{}^{(k)}(v; ª_0, ª_i) = \partial^k h_0(v; ª_0, ª_i)/\partial v^k$.

By the method of fixed-sample optimal estimating functions (Heyde, 1988), we may wish to utilize the root ¯* of the fixed-sample optimal estimating equation $S(¯)|_{¯=¯*} = 0$ to estimate ¯, where

$$S(¯)$$
$$= \Sigma_{i=1}{}^n\Sigma_{j=1}{}^{m(i)}E(-\partial H_{ij}{}^T(¯)/\partial¯|\ _i, ª_i,¥_i)$$
$$Cov^{-1}(H_{ij}(¯)|\ _i, ª_i,¥_i)H_{ij}(¯).$$

Since both $E(-\partial H_{ij}{}^T(¯)/\partial¯|\ _i, ª_i,¥_i)$ and $Cov(H_{ij}(¯)|\ _i, ª_i,¥_i)$ are unavailable for each (i,j) pair in the GTP model with heteroscedastic errors, both $S(¯)$ and ¯* are unavailable.

By the method of generalized estimating equations (GEE) (Liang and Zeger, 1986), we may wish to utilize the root ¯** of the generalized estimating equation $G^{**}(¯)|_{¯=¯**}$ = 0 to estimate ¯ instead of the unavailable estimator ¯*, where

$$G^{**}(¯)$$
$$= \Sigma_{i=1}{}^n\Sigma_{j=1}{}^{m(i)}E(-\partial H_{ij}{}^T(¯)/\partial¯|\ _i, ª_i,¥_i)\,W_{ij}{}^{-1}(¯)$$
$$H_{ij}(¯),$$

with working covariance matrices $W_{ij}(¯) = \mathrm{diag}\{g_{ij}{}^2(\ ,¥),2\cdot g_{ij}{}^4(\ ,¥)\}$. Note that if all standardized errors have the same first four moments as those of the standard normal distribution, then $W_{ij}(¯) = Cov(H_{ij}(¯)|\ _i, ª_i,¥_i)$ and thus both ¯* and ¯** are the same. Since $E(-\partial H_{ij}{}^T(¯)/\partial¯|\ _i, ª_i,¥_i)$ is unavailable for each (i,j) pair, both $G^{**}(¯)$ and ¯** are in general unavailable.

For each (i,j) pair, by deleting all of the terms involving unmodeled $E(£_{i,t(i,j)}{}^3)$, $E(£_{i,t(i,j)}{}^4)$, …, we may approximate both $E(\partial h_{ij}(ª)/\partial ª|\ _i, ª_i,¥_i)$ and $E(H_{ij1}(¯)\cdot\partial h_{ij}(ª)/\partial ª|\ _i, ª_i,¥_i)$ by the Taylor approximation, respectively. Then, by the method of unbiased estimating functions (e.g., see Godambe, 1991), we may utilize the root ¯^ of the unbiased estimating equation $G(¯)|_{¯=¯^}$ = 0 to estimate ¯ instead of the generally unavailable estimator ¯**, where

$$G(¯) = \Sigma_{i=1}{}^n\Sigma_{j=1}{}^{m(i)}D_{ij}{}^T(¯)\,W_{ij}{}^{-1}(¯)H_{ij}(¯),$$

where $D_{ij}(¯)$ is the Taylor approximation of $E(-\partial H_{ij}{}^T(¯)/\partial¯|\ _i, ª_i,¥_i)$.

One way to obtain the estimator ¯^ is to utilize the following iteration method: We first choose a good initial value $¯^{(0)}$ and then iterate the following equations

$$¯^{(k+1)} = ¯^{(k)} + J^1(¯^{(k)})G(¯^{(k+1)}),$$

k = 0, 1, 2, …, until $¯^{(k)}$'s converge to ¯^, where

$$J^1(¯) = \Sigma_{i=1}{}^n\Sigma_{j=1}{}^{m(i)}D_{ij}{}^T(¯)\,W_{ij}{}^{-1}(¯)D_{ij}(¯).$$

Finally, the remaining parameters ρ , ρ_{a} and $\rho_{\text{¥}}$ could be estimated by $\rho\wedge$ = $(\Sigma_{i=1}{}^{n}\ {}_{i}\wedge\ {}_{i}\wedge^{T})/n$, $\rho_{a}\wedge = (\Sigma_{i=1}{}^{n}\, a_{i}\wedge\, a_{i}\wedge^{T})/n$, $\rho_{\text{¥}}\wedge$ = $(\Sigma_{i=1}{}^{n}\text{¥}_{i}\wedge\text{¥}_{i}\wedge^{T})/n$, respectively.

For each subject i, the failure time $T[i]$ is defined as the time when the actual path of subject i crosses the critical level D for some known $D > 0$. Since we only observe the sample paths of subjects, we never observe the actual failure times. For each subject i, we may utilize the root $T[i]\wedge$ of equation $h^{-1}(\text{«}(t;\ {}_{0}\wedge,\ {}_{i}\wedge);\, a_{0}\wedge, a_{i}\wedge)|_{t=T[i]\wedge} = D$ to predict the unobserved failure time $T[i]$.

In some longitudinal studies, we are interested in estimating the time-to-failure distribution F_D for the population of all possible subjects. Since the failure times $T[i]$'s are *i.i.d.* F_D, we may utilize the estimator $F_D\wedge$ to estimate the time-to-failure distribution F_D, where

$$F_D\wedge(t) = [\Sigma_{i=1}{}^{n}1_{(-\infty,t]}(T[i]\wedge)]/n,\ t \in \boldsymbol{R}.$$

四、計畫成果自評

Sometimes, it is inappropriate to assume that the GP model holds for modeling longitudinal data. For example, in practice the transformation h in the GP model is in general unknown. Moreover, different subjects may need different transformations for modeling longitudinal data. Thus, it is better to assume that the transformation h in the GP model depend on fixed-effects and/or random-effects parameters for modeling longitudinal data. For some longitudinal data, we may need not only a transformation but also a weight function in order approximately to achieve homogeneity. This is particularly true when the variance depends on time. Hence, a weight function depending on time t may be needed for transformed longitudinal data. Therefore, it is better to utilize the GTP model with heteroscedastic errors for modeling longitudinal data instead of the GP model.

五、參考文獻

[1] Atkinson, A. C. and Cox, D. R. (1988). Transformations (update). In *Encyclopedia of Statistical Sciences*, Volume 9, Kotz, S. and Johnson, N. L. (Ed.), John Wiley & Sons, New York.

[2] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* 26, 211-243.

[3] Dowling, N. E., (1993). *Mechanical Behavior of Materials*. Prentice Hall, New York.

[4] Godambe, V. P. (Ed.) (1991). *Estimating Functions*. Oxford, New York.

[5] Heyde, C. C. (1988). Fixed sample and asymptotic optimality for classes of estimating functions, *Contemp. Math.* **80**, 241-247.

[6] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

[7] Lu, C. J. and Meeker, W. Q. (1993). Using degradation measures to estimate a time-to-failure distribution, *Technometrics* **35**, 161-174.

[8] Taylor, J. M. G. (1998). Transformations (update). In *Encyclopedia of Statistical Sciences*, Update Volume 2, Kotz, S. (Ed.), John Wiley & Sons, New York.

附件：封面格式

行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※※

※　　　　　　　　　　　　　　　　　　　　　　　　　※

※　　　　**用 Gamma 過程分析單調長期追蹤資料**　　　　※

※　　　　　　　　　　　　　　　　　　　　　　　　　※

※※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：□個別型計畫　　□整合型計畫

計畫編號：NSC 90-2118-M-009-009

執行期間：90 年 8 月 1 日至 91 年 7 月 31 日

計畫主持人：陳志榮

共同主持人：

計畫參與人員：

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

□出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

執行單位：國立交通大學統計所

中　華　民　國　91 年 10　月 30　日