競合モデルに関する統計的手法

宮村　鐵夫

1. はじめに

競合モデルは“competing risk model”の日本語訳であって，その起源は18世紀の数学者ベルヌーイ（D. Bernoulli）の研究までさかのぼることができる．ベルヌーイは，もし天然痘の病気がなくなればその影響は寿命にどのように表われるかを，ある仮定のもとで数学的に求めている．

ベルヌーイの考えたモデルを，今日の競合モデルの言葉を用いて表わすと次のようになる．人間の死亡原因としてのリスク（すなわち病気，事故など）がk種類あったとき，そのリスクの1つである天然痘が撲滅されたならば人類の寿命分布はどのように変化するであろうか．一方，データ解析的な立場からは，死亡原因別の寿命データから各リスク（病気，事故）ごとの死亡率，生存分布などを求めることも重要である．このように競合モデルを用いる解析の方向は上の2つの場合に大別することができる．

前者の場合は，人口統計学，保険統計学などでよく研究されて関心をもたれているものであってcompeting risksという言葉よりもmultiple decrementsという言葉がよく用いられている．後者の場合は，次節の例でも示されるように，得られたデータのモデルとして競合モデルを考えて，このモデルにもとづいてデータを解析しようという立場に立つものである．ここでは，後者の立場に立って競合モデルを把えてその統計的手法に関する話を集めて紹介しよう．

みやむら　てつお　茨城大学

2. 例と問題

競合モデルのデータ解析への応用例を3つほどあげてその具体的な意味を明らかにしておこう．

例1　接着強度の推定

ICに金線を熱圧着する場合に，その接着強度を推定することが問題になることがある（嶋田ら[10]）．金線の接着強度を測定するには，ICに熱圧着した金線を引張ることが必要になる．このとき，接着強度そのものを直接観測することは不可能である．すなわち，X_1を金線の接着強度，X_2をその引張り強度とすれば，観測できるのはこれらの最小値$T=\min(X_1, X_2)$と，金線またはその接着部分のいずれが破断したかの故障原因$J=1$（$X_1 \leqq X_2$のとき），2（$X_1 > X_2$のとき）ということになる．このようなデータから接着強度を推定するにはどうすればよいだろうか．

例2　故障モードと故障率

信頼性の手法の1つであるFMEA（故障モードとその影響解析）で定量的解析を行なうためには故障モードごとの部品の故障率の推定が必要となる．いま故障モードがk種類あって故障モードiが単独で存在した場合の部品の故障までの時間を仮想的に考えてこれをX_iとすれば，部品の寿命データは故障時間：$T=\min(X_1, \cdots, X_k)$，故障原因：$J=\{j | X_j \leq X_i, i=1, \cdots, k\}$と表わすことができる．この部品の寿命データから故障モードご

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

との部品の故障率を推定するにはどうすればよいか．

例 3　医療データの解析

医学の分野では，患者の術後の追跡調査を行なって手術の効果の有無を調べることがよく行なわれる．このときすべての患者について最後まで追跡できればよいが，途中で追跡不可能となる場合が少なくない．このような途中で観測不可能となったデータを含む場合の解析のモデルとしては次のように考えることができる．X_1 を患者の寿命，X_2 をその患者についての観測可能時間（ある確率分布にしたがうものと考える）として，$T=\min(X_1, X_2), J=\{j|X_j\leq X_i, i=1,2\}$ を観測データと考える．これは，例 1，2 と同じ形をしている．

これらの例のように，ある個体の故障となる原因（これをリスクと呼ぶ）がいくつかあり，このリスクが同時に作用して個体が故障に至る過程をモデル化したのが競合モデルである．競合モデルでは $\{T$…故障時間，J…故障原因$\}$ の形のデータから，リスクが単独で作用したときの個体の故障時間の分布について調べることが 1 つの重要な問題となる．以下このデータの解析法について調べてみよう．

3.　競合モデルとその基本的性質

k 個のリスクを $C_1, \cdots, C_k$ で表わし，リスク C_i が単独で個体に作用したときの故障までの時間を X_i とすれば，k 個のリスクが同時に作用したときの個体の故障までの時間 T は，

$$T=\min(X_1, \cdots, X_k)$$

で表わされ，その原因 J は，

$$J=\{C_j|X_j\leq X_i, i=1, \cdots, k\}$$

と書くことができる．

X_i の分布関数，密度関数をそれぞれ $F_i(x)$，$f_i(x)$，$X_1, \cdots, X_k$ の同時生存分布を，

$$Q(x_1, \cdots, x_k)=P(X_1>x_1, \cdots, X_k>x_k)$$

として，$\{T$…故障時間，J…故障原因$\}$ のデータから同時生存分布 $Q(\cdot)$，周辺分布 $\{F_i(\cdot)\}$ が推定可能か調べてみよう．そのため，$\lambda(t), \lambda_j(t), h_j(t)$ を，

$$\lambda(t)=\lim_{\Delta t\to 0}P(t\leq T<t+\Delta t|T\geq t)/\Delta t$$

$$\lambda_j(t)=\lim_{\Delta t\to 0}P(t\leq T<t+\Delta t,\quad J=C_j|T\geq t)\Delta t$$

$$h_j(t)=\lim_{\Delta t\to 0}P(t\leq X_j<t+\Delta t|X_j\geq t)/\Delta t$$

とおけば，$\lambda(t)=\sum_{j=1}^{k}\lambda_j(t)$ となるが，一般には $\lambda_j(t)\neq h_j(t)(j=1, \cdots, k)$ である．

データ（t_i…故障時間，j_i…故障原因）が観測されることは，個体が時間 t_i でリスク j_i の原因により故障したことを意味するから，このようなデータが n 個観測されたときの尤度 L は，

$$(1)\quad L=\prod_{i=1}^{n}\{\lambda_{j_i}(t_i)\prod_{j=1}^{k}\exp(-\int_0^{t_i}\lambda_j(u)du)\}$$

と書くことができる．(1) より，尤度 L は $\{\lambda_j(\cdot), j=1, \cdots, k\}$ の関数として表わされているから，$\{\lambda_j(\cdot)\}$ は推定可能であるが，分布型を仮定しなければ同時生存分布 $Q(\cdot)$，周辺分布 $\{F_j(\cdot), j=1, \cdots, k\}$ は推定不可能である．たとえば，$k=2$ として $\{\lambda_j(\cdot)\}$ を，

$$\lambda_j(t)=\alpha_j[1+\alpha_{12}\exp\{\alpha_{12}(\alpha_1+\alpha_2)t\}],\ j=1,2$$

とすれば，このような $\{\lambda_j(\cdot)\}$ をもつ同時分布としては，

$$(2)\quad Q(x_1, x_2)=\exp[1-\alpha_1x_1-\alpha_2x_2-\exp\{\alpha_{12}(\alpha_1x_1+\alpha_2x_2)\}]$$

$$(3)\quad Q(x_1, x_2)=\exp[1-\alpha_1x_1-\alpha_2x_2-\sum_{j=1}^{2}\alpha_j\exp\{\alpha_{12}(\alpha_1+\alpha_2)x_j\}(\alpha_1+\alpha_2)^{-1}]$$

などがあるから，$\{\lambda_j(\cdot)\}$ より同時分布を一意的に決めることはできない．また一般には $h_j(t)\neq\lambda_j(t)(j=1, \cdots, k)$ であるから周辺分布 $\{F_j(\cdot)\}$ も推定不可能である．

両者を推定可能にするには，個体に k 個のリスクが互いに独立に作用すること，すなわち $X_1, \cdots, X_k$ が互いに独立であることを仮定する必要がある．すると，$h_j(t)=\lambda_j(t)$ $(j=1, \cdots, k)$ となって $\{F_j(\cdot)\}$ は推定可能になる．また $Q(x_1, \cdots, x_k)=\prod_{j=1}^{k}\bar{F}(x_j)$ $(\bar{F}(x_j)=1-F(x_j))$ であるから，$Q(\cdot)$ も推定可能となる．ところで，故障時間と故障原因のデータから $X_1, \cdots, X_k$ が独立であるこ

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

とを検証することは可能であろうか．答は否である．(2)で表わされる同時生存分布はX_1とX_2が独立でないとき，(3)のは独立な場合である．しかしながら同一の$\lambda_j(t)$をもつから，このようなデータでは独立性の検定は不可能であることが示される．したがって，リスクの独立性は他の手段によって検証することが必要である．

次に故障時間と故障原因の同時分布を，

$$G_j(x)=P(T\leqq x, J=C_j)$$

とおけば，リスクが互いに独立な場合には$F_j(x)$は，

$$(4)\quad F_j(x)=1-\exp\{-\int_0^x(1-\sum_{i=1}^{k}G_i(t))^{-1}\,dG_j(t)\}$$

となることが知られている（Berman [1]）．(4)を用いれば，故障原因と故障時間が独立であるためには，$F_j(x)$が，

$$F_j(x)=1-\exp\{-p_jH(x)\},\quad j=1,\cdots,k$$

となることが必要十分であることが示される．すなわち，$F_j(x)$が比例ハザード（proportional hazard）をもつときである．

4. リスクが互いに独立な場合の推定法

$k=2$としても一般性を失わないので，$k=2$としてリスクが互いに独立な場合の周辺分布$F_j(x)$の推定法について述べてみよう．

4.1 ノン・パラメトリックな場合の推定法

故障時間と故障原因

$$T=\min(X_1,X_2)$$

$$\delta=\begin{cases}1, & T=X_1\text{のとき}\\ 0, & T=X_2\text{のとき}\end{cases}$$

のデータより，分布形を仮定しないで$P_t=P(X_1>t)=1-F_1(t)$を推定することを考える．ここで区間$[0,t)$をl個の区間$[u_1=0,u_2),[u_2,u_3),\cdots,[u_l,u_{l+1}=t)$に分割して，$P_i=P(X_1>u_{i+1})$ $(i=1,\cdots,l)$とおけば，P_tは，

$$P_t=P_1\cdot\frac{P_2}{P_1}\cdots\frac{P_t}{P_{t-1}}$$

と書き直すことができる．いまn個のデータが得られているとき，故障時間のデータを小さい順に並べたものを，

$$t_1\leqq t_2\leqq\cdots\leqq t_n$$

として，これに対応する故障原因を$\delta_1,\delta_2,\cdots,\delta_n$とする．すると，$p_i=P_i/P_{i-1}$は時間$u_i$で故障していない個体が時間$u_{i+1}$でも故障していない条件付確率であるから，この自然な推定量としては，

$$\hat{p}_i=(n_i-\xi_i)/n_i$$

が考えられる．ここで，n_iは時間u_iの直前に故障していない個体数，ξ_iは区間$[u_i,u_{i+1})$でリスク1が原因で故障した個体数を示す．特に$u_2=t_1$，$u_3=t_2,\cdots$と考えれば，P_tの推定量として，

$$\hat{P}_t=\hat{p}_1\cdots\hat{p}_t=\prod_{i:t_i\leqq t}\left(\frac{n-i}{n-i+1}\right)^{\delta_i}$$

を導くことができる（Kaplan と Meier [6]）．この推定量については理論的にもよく研究されていて，

(a) $\hat{P}_t$は最尤推定量であり，漸近的に正規分布にしたがう（Kaplan と Meier [6]）．

(b) $\sqrt{n}(\hat{P}_t-P_t)$は，$F_1(t),F_2(t)$が連続ならばガウス過程に弱収束する（Breslow とCrowley [2]）．

などの性質がわかっている．

信頼性では，生存確率P_tのかわりに累積ハザード$H(t)$を推定することに興味がある場合もある．

よく知られているように，P_tと$H(t)$のあいだには，

$$P_t=\exp(-H(t))$$

の関係がある．Nelson [9] は$H(t)$の推定量として，

$$\hat{H}(t)=\sum_{i:t_i\leqq t}\frac{\delta_i}{n-i+1}$$

を与えている．ハザード確率紙を用いてデータを解析する場合には，P_tのかわりに$H(t)$を推定するのが便利なことがある．この推定量についても$\hat{P}_t$とほぼ同じような性質をもつことが調べられている．

4.2 パラメトリックな場合の推定法

リスク1が原因の故障時間のデータがn_1個(t_{11},

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

$\cdots, t_{1n_1}$)，リスク2が原因の故障時間のデータがn_2個($t_{21}, \cdots, t_{2n_2}$) 観測されていれば，尤度Lは比例定数を除いて，

$$L=\prod_{i=1}^{n_1} f_1(t_{1i})\bar{F}_2(t_{1i})\cdot\prod_{j=1}^{n_2}\bar{F}_1(t_{2j})f_2(t_{2j})$$

$$=\{\prod_{i=1}^{n_1}f_1(t_{1i})\prod_{j=1}^{n_2}\bar{F}_1(t_{2j})\}\cdot\{\prod_{j=1}^{n_2}f_2(t_{2j})\prod_{i=1}^{n_1}\bar{F}_2(t_{1i})\}$$

と書くことができる．この尤度の表現をみると，$f_1(x)$の母数の最尤推定量は，この母数が$f_2(x)$と共通のものでないかぎり，$f_2(x)$ の分布型には関係しないということがわかる．

このように，リスクが互いに独立な場合には，$f_1(x)$ の母数の最尤推定量を求めるには，他の分布形$f_2(x)$ の知識を必要としないが，求められた最尤推定量の性質は $f_2(x)$にも依存することになる．

具体的な例として，$X_j(j=1,2)$ が故障率 λ_j の指数分布にしたがう場合を考えてみよう．このとき尤度Lは，

$$L=\lambda_1^{n_1}\exp(-\lambda_1 u)\cdot\lambda_2^{n_2}\exp(-\lambda_2 u)$$

ここで　$u=\sum_{i=1}^{n_1}t_{1i}+\sum_{j=1}^{n_2}t_{2j}$

となるから，λ_1 の最尤推定量 $\hat{\lambda}_1$ は，

$$\hat{\lambda}_1=n_1/u$$

となる．この推定量は「リスク1が原因の故障数／総試験時間」という通常の指数分布の故障率の推定の場合によくみられる形をしている．$\hat{\lambda}_1$ の性質を調べるため，その平均，分散を求めると，それぞれ，

$$\mathbf{E}(\hat{\lambda}_1)=\frac{n}{n-1}\lambda_1,$$

$$\mathrm{Var}(\hat{\lambda}_1)=n\{(n-1)\lambda\lambda_1+\lambda_1^2\}/(n-1)^2(n-2),$$

ここで　$n=n_1+n_2$,　$\lambda=\lambda_1+\lambda_2$,

となる．したがって，$\hat{\lambda}_1$ は不偏推定量ではなく，

$$\tilde{\lambda}_1=\frac{n-1}{n}\lambda_1$$

がλ_1の不偏推定量であって，この分散は，

$$\mathrm{Var}(\tilde{\lambda}_1)=\{(n-1)\lambda\lambda_1+\lambda_1^2\}/n(n-2)$$

となる．このように，$\hat{\lambda}_1, \tilde{\lambda}_1$ の分散は λ_1 のみでなくλ_2の影響をうける．なお推定量の平均，分散を求めるときには指数分布が比例ハザードモデルになることから，n_1とuが独立になることを用いている．

このように，指数分布の場合には，これが比例ハザードモデルに属することから，推定量の性質を調べることは比較的容易であるが，その他の分布の場合には漸近的性質しか調べられていない場合が多い(たとえば，David と Moeschberger [4])．

5. リスクが独立でない場合の推定法

リスクが互いに独立でない場合には，3節で述べたように故障時間と故障原因のデータから周辺分布 $\{F_j(\cdot)\}$ のノン・パラメトリックな推定は一般に不可能である．最近 Langberg ら [8] が，ある条件の下では $\{F_j(\cdot)\}$ の推定が可能であることを示しているが，この条件は独立であるという条件とほぼ同じほど厳しく，実際に成立しているかどうか確認することはむずかしいように思われる．

この条件は，$k=2$で同時分布が絶対連続のときには，任意の $x(>0)$ に対して，

(5)　$P(X_2\geqq x|X_1=x)=P(X_2>x|X_1>x)$

となる．この条件が成立していれば $F_1(x)$ は 4.1 の Kaplan と Meier の方法によって推定可能である．(5) の条件を満たす分布のクラスがどのようなものであるか，またこのクラスが十分に広いものであるかの研究は未だなされていないようである．

リスクが独立でないとノン・パラメトリックな推定はむずかしいが，未知の母数を含む同時分布 $Q(\cdot)$ の形を仮定することができれば，尤度が(1)のように書き表わされることを利用して母数の最尤推定量を求めることは可能である．

6. 共変数を含む場合の推定法

リスクが互いに独立であることを仮定して，共変数 (covariate) を含む場合の推定法について考えてみよう．

© 日本オペレーションズ・リサーチ学会. 無断複写・複製・転載を禁ず.

寿命試験データ，医療データなどでは，観察を始めてから完了するまでの時間的データ

$$t=(t_1,\cdots,t_n)$$

の他に，これに付随したデータ（共変数）

$$W=(z_1,\cdots,z_n)$$

ここで，$z_i=(z_{i1},\cdots,z_{ip})'$

が得られる場合がある．たとえば，医療データについては，治療を始めてから全快あるいは死亡に至るまでの時間とともに，患者の状態に関するデータ（年齢，性別，喫煙の有無，その他病理学的データ等）が得られることが多い．

さて，故障率 $\lambda(t;z)$ が時間に影響される項と共変数に影響される項の積の形

$$\lambda(t;z)=\lambda_0(t)h(z)$$

に書くことができるとする．一般的には z そのものも時間の関数と考えられるが，ここでは z は時間にかかわらず一定の値をとるものとする．

さらに $h(z)$ が，

(6) $h(z)=\exp(\boldsymbol{\beta}z)$

$$=\exp(\beta_1z_1+\cdots+\beta_pz_p)$$

ここで $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)$

と書くことができる場合，このモデルは Cox [3] の regression model と呼ばれる．したがって故障率は $\lambda(t;z)=\lambda_0(t)\exp(\boldsymbol{\beta}z)$ と書くことができて，$\lambda_0(t)$ は $z=\mathbf{0}$ のときの基準故障率を示すことになる．

(6) の形で故障率が与えられているとして，故障率に対する共変数の影響を調べるため，母数 $\boldsymbol{\beta}$ を推定する方法について考えてみよう．もし基準故障率 $\lambda_0(t)$ の形を仮定できれば，最尤法の考え方を援用できる．これが未知の場合には，Cox の部分尤度（partial likelihood）の考えを用いれば $\boldsymbol{\beta}$ を推定できる．いま時間 t_i の直前で故障していない個体の集合を R_i（リスク・セットと呼ぶ）としたとき，次の微小時間内に共変数 z_i の個体が故障する条件付確率を求めると，これは，

$$\exp(\boldsymbol{\beta}z_i)/\sum_{l\in R_i}\exp(\boldsymbol{\beta}z_l)$$

となって，$h_0(t)$ の影響をうけない．この条件付確率の n 個の積

$$L_p=\prod_{i=1}^{n}\{\exp(\boldsymbol{\beta}z_i)/\textstyle\sum_{l\in R_i}\exp(\boldsymbol{\beta}z_l)\}$$

で部分尤度と呼ばれるもので，これを最大にするように $\boldsymbol{\beta}$ を推定しようというのが Cox の考え方である．

いままで述べてきた考え方を競合モデルに適用するには次のようにすればよい．共変数 z をもつ個体に k 個のリスクが同時に作用しているときのリスク j による故障率を，

$$\lambda(t,j,z)$$

$$=\lim_{\Delta t\to 0}P(t\leqq T<t+\Delta t,\ J=C_j\,|\,T\geqq t;z)/\Delta t$$

とする．Lagakos [7] は $\lambda(t,j,z)$ を，

$$\lambda(t,j,z)=\lambda_j\exp(\boldsymbol{\beta}^{(j)}z)$$

と仮定して，また Holt [5] はこれを一般化して，

$$\lambda(t,j,z)=h_{0j}(t)\exp(\boldsymbol{\beta}^{(j)}z)$$

と仮定して，母数 $\lambda_j,\boldsymbol{\beta}^{(j)}=(\beta_1^{(j)},\cdots,\beta_p^{(j)})$ の推定法ならびに検定法について調べている．これらのモデルは Cox の regression model の 1 つである．

この場合でも，(1) の尤度の表現を用いて母数の最尤推定量を求めることは可能である．また尤度比検定を用いた母数の検定も，データ数が多い漸近的な場合には可能である．ただし，Holt のモデルでは $h_{0j}(t)$ の形を仮定しないと最尤法は用いられないので，仮定できない場合には部分尤度を用いることになる．

このようなことから，共変数を含む場合でも，計算は複雑になるとしても母数の推定値を求めることはそれほどむずかしくないが，その性質を調べることは容易でなく漸近的なときの研究しか行なわれていない．

7. おわりに

競合モデルの基本的性質と統計的手法について簡単に述べてきた．競合モデルはその定義からもわかるように，非常に簡単なモデルであるが，実際に得られるデータの背景には 2 節でみたように競合モデルの考え方が適用できる場合が少なくな

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

いと思われる．データを解析する場合には，競合モデルのように，それが得られた背景についても十分考察する必要があると思う．

参考文献

[1] Berman, S.M.: Note on Extreme Values, Competing Risks and Semi-Markov Processes. *Ann. Math. Stat.*, 34 (1963), 1104-1106

[2] Breslow, N and Crowley, J.: A Large Sample Study of the Life Table and Product Limit Estimate under Random Censorship. *Ann. Stat.*, 2 (1974), 437-453

[3] Cox, D. R.: Regression Models and Life Tables, *J. Roy. Stat. Soc.*, B. 34 (1972), 187-220

[4] David, H. A, and Moeschberger M. L.: *The Theory of Competing Risks*, Charles Griffin, London, 1978

[5] Holt, J. D.: Competing Risk Analysis with Special Reference to Matched Pair Experiments. *Biometrika*, 65 (1978), 159-166

[6] Kaplan, E. L. and Meier, P.: Nonparametric Estimation from Incomplete Observations. *J. Amer Stat. Assoc.*, 282 (1958), 451-481

[7] Lagakos, S. W.: A Covariate Model for Partially Censored Data Subject to Competing Causes of Failure. *Appl. Stat.*, 27 (1978), 235-241

[8] Langberg, N., Proschan, F. and Quinzi, A. J.: Estimating Dependent Life Length with Applications to the Theory of Competing Risks. *Ann. Stat.*, 9 (1981), 157-167

[9] Nelson, W.: Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics*, 14 (1972), 945-966

[10] 嶋田正三，芳賀敏郎，岩田誠一：2つの分布の小さい方の値のみが観測される場合の母平均と標準偏差の推定．応用統計学，**5** (1977), 63-83

●ご利用ください●差し上げます●

下記の雑誌は交換等によって，日本OR学会にほぼ定期的に送られてきているものです．学会事務局で保管しておりますので，どうぞご利用ください．下記のもの以外にも大学の論叢等があります．なお，1981年中に発行のものは，ご希望があれば差し上げますので事務局までお申出ください．

(1)	ＩＥ	(社)日本能率協会
(2)	運輸と経済	(財)運輸調査局
(3)	ENGINEERS	(財)日本科学技術連盟
(4)	技術と経済	(社)科学技術と経済の会
(5)	計測と制御	(社)計測自動制御学会
(6)	研究実用化報告	電々公社武蔵野通研
(7)	高速道路と自動車	(財)高速道路調査会
(8)	産業能率	大阪府立産業能率研究所
(9)	数理科学	(株)サイエンス社
(10)	テレトピア	日本電々公社
(11)	電子通信学会誌	(社)電気通信学会誌
(12)	電子通信学会論文誌	〃
(13)	土木学会誌	(社)土木学会
(14)	日本機械学会誌	(社)日本機械学会
(15)	標準化ジャーナル	(社)日本規格協会
(16)	標準化と品質管理	〃
(17)	労働研究	兵庫県労働部労働調査室
(18)	統計数理研究所彙報	統計数理研究所
(19)	Annals of The Institute of Statistical Mathematics	〃
(20)	理論経済学	理論・計量経済学会

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．