特集　社会的リスクのOR

比例ハザードモデルによる統計解析

鎌倉　稔成

1. はじめに

統計解析の対象となるデータは，変動的な測定値の集合として定義されるが，ここで重要なことは測定可能ということである．測定可能な対象の変動の規則を探ることが統計解析の目的である．測定されるデータの属性の違いによって解析の方法が異なってくるが，本稿では，主として人間あるいは機械などの寿命データ(生存時間データ)の統計的扱いについて述べる．その中でも特に，最近脚光を浴びている Cox の比例ハザードモデル[2]に焦点をあてて解説を行なう．

変動するデータから規則性を見いだすには，母集団の分布を何らかの方法で推定することが必要となる．寿命分布を扱うさいには，分布関数を1から引いた形の生存時間分布関数(Survival function)のほうが都合がよい．生存時間を T としたとき，生存時間分布関数 $S(t)$は，

$$S(t)=Pr\{T>t\}$$

で定義され，この $S(t)$ を用いると，ハザード関数，$\lambda(t)$，密度関数，$f(t)$ はそれぞれ次のように表現される．

$$\lambda(t)=-d\log S(t)/dt,$$

$$f(t)=\lambda(t)S(t).$$

従来，母集団分布の推定に当っては様々なパラメトリックモデルが提案され議論されてきた．たとえば生存時間分布関数としては指数分布，ワイブル分布，対数正規分布，ロジスチック分布，ガンマ分布などが用いられている．また Kaplan-Meier[7]などのノンパラメトリックな生存時間分布関数の推定方法などもよく利用されている．

2. 比例ハザードモデル

機械やシステムの寿命分布を考えるときには，ある一定の条件下での分布ということで，環境条件の均一性が重要な仮定となる．しかしながら，実際にはデータを収集する場では，環境を一定にすることはなかなか困難であることが多い．特にフィールドデータでは，環境のコントロールはまず不可能であるとみてよい．また人間の寿命を対象とする場合にもむずかしいといえる．そこで逆に，積極的に変動する環境条件を計測し，条件の違いをモデルに導入することを考えるのである．環境条件を表わす変数を共変量(covariate)と呼び，機械などの寿命を考えるときには，温度，湿度，圧力，等の使用条件がそれにあたる．人間の寿命を考える場合には環境条件だけでなく，性別，年齢などの内的な変数も寿命分布に影響を与える変数であり，共変量として扱われる．

比例ハザードモデルは，寿命分布に回帰型のモデルを導入し，環境による分布の不均一性を，環境の変動を表わす共変量によって説明しようというものである．モデルは，ハザード関数を用いて次のように表現される．

かまくら　としなり　統計数理研究所

$$\lambda(t)=\lambda_0(t)\exp(z\beta).$$

ただし$\lambda_0(t)$は不特定の非負の関数であり，基準ハザード関数(base-line hazard function)と呼ばれる．zは前述した共変量で，ベクトル値であってかまわない．このモデルは共変量の影響を積の形でハザードに取り入れている点に特徴がある．積の形でモデル化することの利点は回帰係数βの推定が基準ハザード関数$\lambda_0(t)$によらず簡単に行なうことができる点にある．Coxの提案した回帰係数の推定方法は$\lambda_0(t)$に関する情報を無視した部分尤度にもとづくものである．部分尤度は次のようにしてつくられる．まず比例ハザードモデルのハザード関数を用いて密度関数を表現し，与えられたデータ$(t_i,\delta_i,z_i)(i=1,\cdots,n)$から比例ハザードモデルの尤度，$L(\beta,\lambda_0(t))$をつくる．ただし$t_i$, δ_i, z_iはそれぞれi番目の個体に対する故障時間(死亡時間)，打ち切りデータかそうでないかを示す変数(1：故障，0：打ち切り)，共変量ベクトルである．$L(\beta,\lambda_0(t))$を$\lambda_0(t)$を含まない$L_P(\beta)$と残りの$L_R(\beta,\lambda_0(t))$の2つの因数に分け，第2因数を無視することによる情報損失はβの推定にさいしては少ないとして得られるのである[3][4]．ここに$L_P(\beta)$は次のように表現される．

$$L_P(\beta)=\prod_{i:\delta_i=1}\frac{\exp(z_i\beta)}{\sum_{j\in R(t_{(i)})}\exp(z_j\beta)}$$

ただし，$R(t)$はリスク集合と呼ばれ，

$$R(t_i)=\{j:t_i\leqq t_j\},$$

であり，$t_{(i)}$は$\{t_i:\delta_i=1\}$を順序づけたときのi番目の値である．ここまでの尤度の取扱いでは，タイ(同順位)のデータを考慮に入れていないし，連続モデルではタイがおこるのは理論的には確率測度0であるが，実際のデータでは観測精度などの点からタイがおこることが十分におこりうる．タイの扱いを含めた回帰係数の推定法としてはBreslow-Petoの方法がある[1]．観測時点で不連続なステップ関数としてノンパラメトリックに基準ハザード関数を与え，尤度関数を書き下すものである．この方法では，尤度は，

$$L_{BP}(\beta)=\prod_{i=1}^{k}\frac{\prod_{j\in\phi_{0i}}\exp(z_j\beta)}{[\sum_{j\in R(t_{(i)})}\exp(z_j\beta)]^{d_i}}$$

に比例するという形で与えられる．ここにd_iは各時点におけるタイの数，ϕ_{0i}はt_i時点で死亡した個体の集合である．また，$t_{(1)},<t_{(2)}<\cdots<t_{(k)}$は$\{t_{(i)}\}$のうち異なる故障時間を順序づけたものである．タイがある場合の部分尤度は次のように表現される．

$$\prod_{i=1}^{k}\frac{\prod_{j\in\phi_{0i}}\exp(z_j\beta)}{\sum_{\phi\in\Phi_i}\prod_{j\in\phi}\exp(z_j\beta)}$$

ただしΦ_iは$t_{(i)}$時点におけるリスク集合，$R(t_{(i)})$の要素から，その時点での死亡あるいは故障の数(タイの数)d_i個をとり出してつくる組合せからなる集合である．ϕはその集合の任意の組合せの1つである．タイがおきない場合にはBreslow-Petoの尤度のほうが原点側に偏った推定量を与えることが証明されている[6]．しかしながらタイが大きい場合には部分尤度の計算が困難であり，Breslow-Petoの方法が用いられていることが多い．現在，比例ハザードモデルを含むプログラム・パッケージのほとんどは計算の容易な尤度，$L_{BP}(\beta)$の最大化によって回帰係数の推定を行なっている．たとえばSAS, BMDP, SURVREGなどで$L_{BP}(\beta)$が利用されている．$L_{BP}(\beta)$から得られる回帰係数βの推定量は，部分尤度のそれと比較してコンサーバティブな推定量となるので，βの有意性検定のさいには都合がよい．タイが極端に多い場合を除けば，計算のしやすさを考えて，$L_{BP}(\beta)$の利用で十分である．次に述べる例でもBreslow-Petoの方法を用いている．

3. 解析例

表1は，[2]にみられる白血病の寛解時間のデータである．2標本問題となっており，標本1と標本2の生存時間に有意な差があるかどうかが問題である．両者の違いをみるためには，まず予備

表 1　白血病患者の寛解時間(単位：週)　*印はフォロー・アップが完全でないことを示す.

標本1	6*	6	6	6	7	9*	10*	10	11*	13	16	17*	19*	20*	22	23	25*	32*	32*	34*	35*
標本2	1	1	2	2	3	4	4	5	5	8	8	8	11	11	12	12	15	17	22	23	

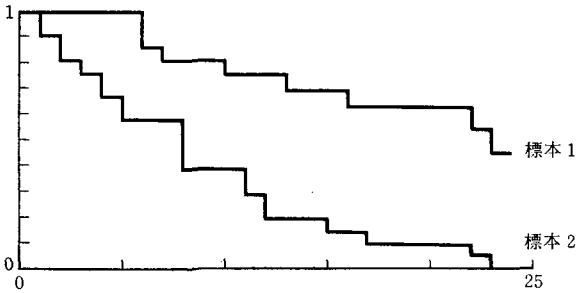

図 1　Kaplan-Meier 推定値

的な解析として，ノンパラメトリックな方法で生存時間分布関数を推定してグラフ化することである．標本1と標本2のそれぞれのKaplan-Meierの推定値を図示したものが図1である．図からは，標本1のほうが2に比べて寛解時間が長そうであることが見てとれるが，標本1に対しての優越性を示す客観的尺度は明らかでない．しかしながら，データ全体の情報を視覚的にとらえることも，分布の形状を既存のモデルにとらわれることなく把握するうえで重要なことである．

ここでは，1変数の共変量を導入して比例ハザードモデルの適用について述べる．共変量zの値は，標本1から得られたデータには1，標本2からのものには0を与えることにする．また，データの中で*印を識別するために，フォロー・アップが完全の場合には1，不完全の場合には0となる変数を用いて表1のデータを整理する．次に，$L_{BP}(\beta)$ の尤度をつくり最尤法によって標本を識別する回帰係数βを推定すれば，$\beta=-1.50919$，その正規偏位 (Normal Deviate) は-3.68487，片側確率は0.00011となる．したがって，比例ハザードモデルによる解析では，両標本には有意な差があるとみることができる．

4. おわりに

紙面の都合で共変量がベクトル値の場合について述べることができなかったが，共変量がリスク因子と混乱因子に分れるような場合には，混乱因子を調整する形でリスク因子の影響をみることが可能である．また，生存時間データだけでなく順序のあるカテゴリーデータの分割表の解析も比例ハザードモデルで行なうことができる[8]．応用範囲の広いフレキシブルなモデルであるが，モデルの適合性についてはまだ十分な検討がなされていない．今後の研究を待ちたい．

参考文献

[1] Breslow, N. E. : Covariance analysis of censored survival data. *Biometrics*, Vol. 30, No. 1(1974), 89-99

[2] Cox, D. R. : Regression models and life tables(with discussion). *J. R. Statist. Soc.*, Vol. 34, No. 2(1972)187-220

[3] Cox, D. R. : Partial likelihood. *Biometrika*, Vol. 62, No. 2(1975), 269-276.

[4] Johansen, S. : An extension of Cox's Regression Model. *Int. Statist. Rev.* Vol. 51, No. 1 (1983), 165-174

[5] Kalbfleisch, J. D. and Prentice, R. L. : *The statistical Analysis of Failure Time Data*. New York, Wiley, 1980

[6] Kamakura, T. and Yanagimoto, T. : Evaluation of the regression parameter estimators in the proportional hazard model. *Biometrika*, Vol. 70, No. 2, 530-533

[7] Kaplan, E. L. and Meier, P. : Nonparametric Estimation from Incomplete Observations. *J. Amer. Statist. Assoc.* Vol. 53 (1958), 457-481

[8] 柳本武美・清水央子：2次元分割表における比例ハザードモデルの適用，応用統計学12, 1 (1983), 17-29