福山平成大学経営学部紀要
第１３号（2016），*－**頁

社会システム分析のための統合化プログラム３１
－生存時間分析－

福井正康[*1]、呉暁娜[*2]
[*1] 福山平成大学経営学部
[*2] 福山平成大学大学院経営学研究科

要旨：我々は教育分野での利用を目的に社会システム分析に用いられる様々な手法を統合化したプログラム College Analysis を作成してきた。今回は、中途打ち切りを含むデータから死亡率や生存確率分布を予測する分析手法である生存時間分析のプログラムについて報告する。プログラムには生存時間分布表、Kaplan-Meier のグラフと予測曲線、Cox 比例ハザードモデル、最尤法による Weibull ハザードモデルと混合 Weibull ハザードモデルなどが含まれる。

キーワード：College Analysis、多変量解析、生存時間分析、混合 Weibull 分布、ハザードモデル、EM アルゴリズム
URL：http://www.heisei-u.ac.jp/ba/fukui/

1. はじめに

生存時間分析は中途打ち切りを含むデータから死亡危険率や生存確率分布を予測する分析手法である。この分析は生物の生存時間だけでなく、機械の故障までの時間などにも利用できる。そのため、死亡という言葉は、あるイベントが発生するまでの時間とした方が的を射ているが、ここでは慣例的に使われてきた死亡や生存という言葉を使うことにする。

生存時間分析では生存数、期間死亡数、期間打ち切り数を元に、期間死亡率、期間生存率、累積生存確率（以後これを表す関数を生存関数と呼ぶ）、ハザードなどを計算し、その生存関数を視覚的にグラフで表す。また、その生存関数から累積死亡確率を求め、それに指数分布や Weibull 分布などの分布関数を当てはめ、モデルを作成する。また、対数ハザードに関して、死亡の原因と考えられる変数の影響を線形関数として導入し、その変数のハザードへの影響を議論する。さらに、分布を仮定して生存関数自体をこれらの変数で予測することも考える。このようなモデルによる分析を、ハザードモデルという。特に、分布によらず、変数の影響を考える有名なモデルには Cox 比例ハザードモデルがある。

人間の死亡や製品の故障のデータを、全時間を通して単一の指数分布や Weibull 分布で表すことは難しい。我々はパラメータの異なるこれらの分布を重ね合わせた混合指数分布や混合 Weibull 分布も考える。これはハザードモデルについても同様であり、混合 Weibull ハザードモデルも考えるようにした。

2. 生存時間分析の理論

2.1 生存時間分析の基礎

時刻$t=0$に$l(0)$個の個体があり、死亡により時刻tに個体数が$l(t)$個になっているものとする。時刻tからの単位時間の間に死亡する割合$p(t)=-dl(t)/dt$は、以下で与えられると仮定する。

$$-\frac{dl(t)}{dt}=\mu(t)l(t)$$

ここに$\mu(t)$を時刻tにおける死力という。

上式を時刻tと時刻$t+h$の間で定積分すると以下の関係を得る。

$$\log l(t+h)-\log l(t)=-\int_t^{t+h}\mu(\tau)d\tau=-\int_0^h\mu(t+\tau)d\tau$$

これより、

$$l(t+h)=l(t)\exp\left[-\int_0^h\mu(t+\tau)d\tau\right]$$

ここで、$p(h;t)=\exp\left[-\int_0^h\mu(t+\tau)d\tau\right]$とおくと、$p(h;t)$は時刻$t$から$t+h$の間の期間生存率と呼ばれる。この期間生存率は以下のようにも書ける。

$$p(h;t)=\frac{l(t+h)}{l(t)}$$

同様にして、期間死亡率$q(h;t)$も以下のように与えられる。

$$q(h;t)=1-p(h;t)=\frac{l(t)-l(t+h)}{l(t)}=\frac{d(h;t)}{l(t)}$$

ここに$d(h;t)=l(t)-l(t+h)$は期間死亡数を表す。特に、$h=1$とした期間生存率、期間死亡率を単に時刻tでの生存率$p(t)$、死亡率$q(t)$という。

時刻t以降の生存時間の合計$T(t)$を個体数で割った量を平均余命$e(t)$という。

$$e(t)=\int_t^\infty l(\tau)d\tau\Big/l(t)=T(t)/l(t)$$

また、$t=0$での平均余命を平均寿命という。

死亡の発生までの時間を確率変数Tとする確率分布を考え、その密度関数を$f(t)$、分布関数を$F(t)$とすると、これらには以下の関係がある。

$$F(t)=P(0\le T\le t)=\int_0^t f(\tau)d\tau$$

分布関数$F(t)$は累積死亡関数である。これに対して、時刻tまで生きる確率を表す関数$S(t)$を生存関数といい、以下で表される。

$$S(t)=P(T>t)=1-F(t)=\int_t^\infty f(t)dt$$

時刻tにおける死亡発生危険率をハザード関数（故障率関数）$\lambda(t)$といい、以下のように定義される。

$$\lambda(t)=\frac{f(t)}{S(t)}=-\frac{d}{dt}\log S(t)$$

このハザード関数を積分した累積ハザード関数$\Lambda(t)$は以下のように与えられる。

$$\Lambda(t)=\int_0^t \lambda(\tau)d\tau=-\log S(t)$$

逆に生存関数は、以下のように表される。

$$S(t)=e^{-\Lambda(t)}$$

生存関数は$t\to\infty$で$S(t)\to 0$であるから、累積ハザード関数は$t\to\infty$で$\Lambda(t)\to\infty$でなければならない。

累積死亡分布には、指数分布やWeibull分布が仮定される。指数分布の確率密度関数は以下で与えられる。

$$f(t)=\lambda e^{-\lambda t}\qquad (t\ge 0)$$

分布関数と生存関数はそれぞれ以下で与えられる。

$$F(t)=1-e^{-\lambda t},\quad S(t)=e^{-\lambda t}\qquad (t\ge 0)$$

確率変数の平均と分散はそれぞれ以下で与えられる。

$$E[T]=\frac{1}{\lambda},\quad V[T]=\frac{1}{\lambda^2}$$

ハザード関数は定数で与えられる。

$$\lambda(t)=\frac{f(t)}{S(t)}=\frac{\lambda e^{-\lambda t}}{e^{-\lambda t}}=\lambda$$

Weibull分布の確率密度関数は以下で与えられる。

$$f(t)=(a/b)\left(t/b\right)^{a-1}\exp\left[-\left(t/b\right)^{a}\right]\qquad (t\ge 0)$$

分布関数と生存関数はそれぞれ以下で与えられる。

$$F(t)=1-\exp\left[-\left(t/b\right)^{a}\right],\quad S(t)=\exp\left[-\left(t/b\right)^{a}\right]$$

確率変数の平均と分散はそれぞれ以下で与えられる。

$$E[T]=b\,\Gamma\left(1+1/a\right),\quad V[T]=b^2\left[\Gamma(2+1/a)-\Gamma(1+1/a)\right]^2$$

ここに、$\Gamma(x)$はガンマ関数である。ハザード関数は以下で与えられる。

$$\lambda(t)=\frac{f(t)}{S(t)}=\frac{(a/b)\left(t/b\right)^{a-1}\exp\left[-\left(t/b\right)^{a}\right]}{\exp\left[-(t/b)^{a}\right]}=(a/b)\left(t/b\right)^{a-1}=at^{a-1}/b^{a}$$

実際のハザード関数は、初期段階で値が大きく、しばらく時間が経つと安定期に入り、最終的な段階でまた値が大きくなる。安定期では指数分布が使われ、初期段階では Weibull 分布がよく利用される。最終段階ではあまり当てはまりが良くないと言われることもあるが、我々は Weibull 分布を当てはめてみる。全体への当てはめの分布としては、後に述べる混合 Weibull 分布を考えてみることにする。

2.2 Kaplan-Meier 推定と log-rank 検定

観測対象 $\lambda = 1, \cdots, N$ に対して、生存時間を $t_\lambda = 0$ から $t_\lambda = T_\lambda$（打ち切りのないデータ）、$t_\lambda = 0$ から $t_\lambda = T_\lambda^+$（打ち切りのあるデータ、実際のデータでは 17+ 等と表記）とする。この終了時刻 T_λ を 0 から順番に並べた時刻を $t_0 = 0, t_1, \cdots, t_m$（同一のものもある）とし、$t_m$ ですべて死亡および打ち切りが確認されたものとする。これに対して、一定の時間間隔で時刻を取る方法もある。各時点での生存数を l_i、$t_i < t \le t_{i+1}$ の間に死亡した数を d_i、打ち切りになった数を w_i とする。これらを使って、死亡のリスクにさらされた数を $r_i = l_i - w_i/2$ とする。

死亡の期間発生率 q_i と期間生存率 p_i は以下で与えられる。

$$q_i = d_i / r_i, \quad p_i = 1 - q_i$$

生存関数 S_i、密度関数 f_i、ハザード関数 λ_i は以下のように計算される。

$$S_i = \prod_{k=0}^{i-1} p_k, \quad f_i = q_i S_i / (t_i - t_{i-1}), \quad \lambda_i = f_i / S_i = q_i / (t_i - t_{i-1})$$

このような生存関数の推定法を Kaplan-Meier の product-limit 推定法（以後 Kaplan-Meier 推定法と呼ぶ）という。生存関数 S_i のばらつきを表す標準誤差 $S.E.[S_i]$ は近似的に以下で与えられることが知られている。

$$S.E.[S_i] = S_{i-1} \sqrt{\sum_{k=1}^{i-1} \frac{d_k}{l_k (l_k - d_k)}} \qquad (i \ge 2)$$

指数分布や Weibull 分布の見極めは、累積ハザード関数に関する以下の関係を利用し、グラフが直線になるか否かで判断することができる。

指数分布　$-\log S(t) = \lambda t$

Weibull 分布　$\log(-\log S) = a \log(t/b) = a \log t - a \log b$

指数分布や Weibull 分布のパラメータの最小 2 乗推定は、以下の式によって与えられる。

指数分布　$S(t) = e^{-\lambda t}$

$$\lambda = -\sum_{i=0}^{m-1} t_i \log S_i \Big/ \sum_{i=0}^{m-1} t_i^2$$

Weibull 分布　$S(t)=\exp\left[-(t/b)^{a}\right]$

$t_i'=\log t_i$，　$S_i'=\log(-\log S_i)$ として、

$$a=\sum_{i=1}^{m-1}(t_i'-\bar{t}')(S_i'-\bar{S}')\Big/\sum_{i=1}^{m-1}(t_i'-\bar{t}')^2 \quad ， \quad b=\exp\left[-(\bar{S}'-a\bar{t}')/a\right]$$

分類数G の個体群について、生存時間データの差の検定を行うには以下の性質を用いる。第 r 分類群のt_i 時点での期間死亡数をd_i^r 、生存数をl_i^r として

$$O_r=\sum_{i=0}^{m-1}d_i^r \ ， \quad E_r=\sum_{i=0}^{m-1}l_i^r(d_i/l_i)， \quad \text{ここに、} l_i=\sum_{r=1}^{G}l_i^r \ ， \quad d_i=\sum_{r=1}^{G}d_i^r$$

を計算し、以下の近似的な関係を用いて群間の差を検定する。

$$\chi^2=\sum_{r=1}^{G}\frac{(O_r-E_r)^2}{E_r}\sim\chi^2_{G-1}$$

ここに、O_r は分類群r の実測累積死亡数、E_r は分類群r の予測累積死亡数である。この検定を Peto & Peto の log-rank 検定という。

2.3　パラメータの最尤推定

1）指数分布に基づく最尤推定

最初に通常の指数分布の最尤推定を考える。指数分布の確率密度関数と生存関数は以下で与えられる。

$f(t)=\lambda\exp(-\lambda\,\mathrm{t})$　　（$t\geq 0$）

$S(t)=\exp(-\lambda\,\mathrm{t})$　　（$t\geq 0$）

指数分布の最尤推定で、尤度$L(\lambda)$ は以下で与えられる。

$$L(\lambda)=\prod_{i=1}^{N}f(t_i)^{\delta_i}S(t_i)^{1-\delta_i}$$

ここで、打ち切りデータと非打ち切り（死亡）データをそれぞれ$\delta_i=0,1$ としている。

ハザード関数は以下で与えられる。

$$\lambda(t)=\frac{f(t)}{S(t)}=\lambda$$

対数尤度は以下となる。

$$\log L(\lambda)=\sum_{i=1}^{N}\left[\delta_i\log\lambda(t_i)+\log S(t_i)\right]=\sum_{i=1}^{N}\left[\delta_i\log\lambda-\lambda t_i\right]$$

対数尤度を微分してスコアベクトルに相当するものを作成するが、この場合はスカラーである。これを仮にスコアと呼ぶ。

$$\frac{\partial}{\partial\lambda}\log L=\sum_{i=1}^{N}\left[\delta_i/\lambda-t_i\right]=\frac{1}{\lambda}\sum_{i=1}^{N}\delta_i-\sum_{i=1}^{N}t_i=0$$

$$\lambda=\sum_{i=1}^{N}\delta_i\Big/\sum_{i=1}^{N}t_i$$

スコアをもう一度微分して、情報行列 $\Im$ に相当するものを作成する。この場合もスカラーである。

$$\Im=-\frac{\partial^2}{\partial\lambda^2}\log L=\frac{1}{\lambda^2}\sum_{i=1}^{N}\delta_i$$

この逆数は、推定値の分散を与える。

2) Weibull 分布に基づく最尤推定

最初に通常の Weibull 分布の最尤推定を考える。weibull 分布の確率密度関数と生存関数は以下で与えられる。

$$f(t)=(a/b)\left(t/b\right)^{a-1}\exp\left[-\left(t/b\right)^{a}\right]\qquad(t\geq 0)$$

$$S(t)=\exp\left[-\left(t/b\right)^{a}\right]\qquad(t\geq 0)$$

Weibull 分布の最尤推定で、尤度 $L(a,b)$ は以下で与えられる。

$$L(a,b)=\prod_{i=1}^{N}f(t_i)^{\delta_i}S(t_i)^{1-\delta_i}$$

ここで、打ち切りデータと非打ち切り（死亡）データをそれぞれ $\delta_i=0,1$ としている。

ハザード関数は以下で与えられる。

$$\lambda(t)=\frac{f(t)}{S(t)}=\frac{(a/b)\left(t/b\right)^{a-1}\exp\left[-\left(t/b\right)^{a}\right]}{\exp\left[-(t/b)^{a}\right]}=(a/b)\left(t/b\right)^{a-1}=at^{a-1}b^{-a}$$

対数尤度は以下となる。

$$\begin{aligned}\log L(a,b)&=\sum_{i=1}^{N}\left[\delta_i\log\lambda(t_i)+\log S(t_i)\right]\\&=\sum_{i=1}^{N}\left[\delta_i\log\left(at_i^{a-1}b^{-a}\right)-t_i^{a}b^{-a}\right]=\sum_{i=1}^{N}\left[\delta_i\log\left(at_i^{a-1}e^{\beta}\right)-t_i^{a}e^{\beta}\right]\\&=\sum_{i=1}^{N}\left[\delta_i\left(\log a+(a-1)\log t_i+\beta\right)-t_i^{a}e^{\beta}\right]\end{aligned}$$

ここで、$b^{-a}=e^{\beta}$ （$b=e^{-\beta/a}$，$e^{\beta}=b^{-a}\rightarrow\exp({}^{t}\mathbf{x}\boldsymbol{\beta})$ に相当）としている。

これを微分して、スコアベクトル $\mathbf{U}$ と情報行列 $\Im$ をもとめると以下となる。

$$\boldsymbol{\beta}' = \begin{pmatrix} a \\ \beta \end{pmatrix},\quad \mathbf{U} = \begin{pmatrix} \partial \log L/\partial a \\ \partial \log L/\partial \beta \end{pmatrix},\quad \Im = -\begin{pmatrix} \partial^2 \log L/\partial a^2 & \partial^2 \log L/\partial a \partial \beta \\ \partial^2 \log L/\partial a \partial \beta & \partial^2 \log L/\partial \beta^2 \end{pmatrix}$$

ここに、

$$\frac{\partial}{\partial a}\log L = \sum_{i=1}^{N}\left[\delta_i(1/a + \log t_i) - \log t_i\, t_i^a e^{\beta}\right]$$

$$\frac{\partial}{\partial \beta}\log L = \sum_{i=1}^{N}\left[\delta_i - t_i^a e^{\beta}\right]$$

$$\frac{\partial^2}{\partial a^2}\log L = -\sum_{i=1}^{N}\left[\delta_i/a^2 + (\log t_i)^2 t_i^a e^{\beta}\right]$$

$$\frac{\partial}{\partial a \partial \beta}\log L = -\sum_{i=1}^{N}\log t_i\, t_i^a e^{\beta}$$

$$\frac{\partial^2}{\partial \beta^2}\log L = -\sum_{i=1}^{N} t_i^a e^{\beta}$$

これらを用いてニュートン・ラフソン法でパラメータの推定を行う。

$$\boldsymbol{\beta}'^{(m+1)} = \boldsymbol{\beta}'^{(m)} + (\Im^{(m)})^{-1}\mathbf{U}^{(m)}$$

ここに右肩の添え字はニュートン・ラフソン法のループの段階を表している。

この情報行列の逆行列の対角成分はパラメータの分散を与える。

3）混合分布に基づく最尤推定

混合分布の最尤推定で、尤度 $L(\lambda)$ は以下で与えられる。

$$L(\lambda) = \prod_{i=1}^{N} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

K 種混合分布では、それぞれの密度関数を $f_k(t)$、生存関数を $S_k(t)$ として、全体の密度関数と生存関数は以下となる。ここに、π_k は分布の重ね合わせの確率である。

$$f(t) = \sum_{k=1}^{K}\pi_k f_k(t),\quad S(t) = \sum_{k=1}^{K}\pi_k S_k(t)$$

混合分布の最尤推定で、尤度 $L(\boldsymbol{\theta}, \boldsymbol{\pi})$ は以下で与えられる。

$$L(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{N} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = \prod_{i=1}^{N}\left(\sum_{k=1}^{K}\pi_k f_k(t_i)\right)^{\delta_i}\left(\sum_{k=1}^{K}\pi_k S_k(t_i)\right)^{1-\delta_i}$$

ここで、打ち切りデータと非打ち切り（死亡）データをそれぞれ $\delta_i = 0,1$ としている。

対数尤度は以下となる。

$$\log L(\boldsymbol{\theta},\boldsymbol{\pi}) = \sum_{i=1}^{N}\left[\delta_i \log\sum_{k=1}^{K}\pi_k f_k(t_i) + (1-\delta_i)\log\sum_{k=1}^{K}\pi_k S_k(t_i)\right]$$

$$= \sum_{i=1}^{N}\left[\delta_i \log\sum_{k=1}^{K} q_k^{(i)}\frac{\pi_k f_k(t_i)}{q_k^{(i)}} + (1-\delta_i)\log\sum_{k=1}^{K} q_k^{(i)}\frac{\pi_k S_k(t_i)}{q_k^{(i)}}\right]$$

$$\geq \sum_{i=1}^{N}\left[\sum_{k=1}^{K} q_k^{(i)}\delta_i \log\frac{\pi_k f_k(t_i)}{q_k^{(i)}} + \sum_{k=1}^{K} q_k^{(i)}(1-\delta_i)\log\frac{\pi_k S_k(t_i)}{q_k^{(i)}}\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\left[\delta_i \log f_k(t_i) + (1-\delta_i)\log S_k(t_i) + \log\pi_k - \log q_k^{(i)}\right]$$

上式の不等号は、$q_k^{(i)}$ の値によって、等号になることが知られている。

パラメータの推定には以下の手順①と②をパラメータ値が収束するまで繰り返す。このような 2 段階の推定法を EM アルゴリズムという。

①パラメータ $q_k^{(i)}, \pi_k$ の最適化

この $q_k^{(i)}$ について、$\sum_{k=1}^{K} q_k^{(i)} = 1$ の条件をつけて右辺を最大化するために、ラグランジュの未定定数法を用いる。

$$\frac{\partial}{\partial q_k^{(i)}}\left[\log L(\boldsymbol{\theta},\boldsymbol{\pi}) - \sum_{i=1}^{N}\eta_i\left(\sum_{k=1}^{K} q_k^{(i)} - 1\right)\right]$$

$$= \delta_i \log\frac{\pi_k f_k(t_i)}{q_k^{(i)}} + (1-\delta_i)\log\frac{\pi_k S_k(t_i)}{q_k^{(i)}} + 1 - \eta_i$$

$$= \log\frac{\pi_k f_k(t_i)^{\delta_i} S_k(t_i)^{1-\delta_i}}{q_k^{(i)}} - 1 - \eta_i = 0$$

これより、

$$q_k^{(i)} = e^{-(1+\eta_i)}\pi_k f_k(t_i)^{\delta_i} S_k(t_i)^{1-\delta_i} = \frac{\pi_k f_k(t_i)^{\delta_i} S_k(t_i)^{1-\delta_i}}{\sum_{k=1}^{K}\pi_k f_k(t_i)^{\delta_i} S_k(t_i)^{1-\delta_i}}$$

これを書き換えて、以下のようにすることもできる。

$$q_k^{(i)} = \pi_k f_k(t_i) \Big/ \sum_{k=1}^{K}\pi_k f_k(t_i) \qquad \text{for } \delta_i = 1$$

$$q_k^{(i)} = \pi_k S_k(t_i) \Big/ \sum_{k=1}^{K}\pi_k S_k(t_i) \qquad \text{for } \delta_i = 0$$

この $q_k^{(i)}$ を群 k への帰属度という。

次に、この尤度関数をパラメータ π_j で微分して 0 と置き、パラメータの推定を行うが、

$\sum_{k=1}^{K}\pi_k = 1$ の条件をつけるために、ラグランジュの未定定数法を用いる。

$$\frac{\partial}{\partial \pi_j}\left[\log L(\boldsymbol{\theta},\boldsymbol{\pi}) - \eta\left(\sum_{k=1}^{K}\pi_k - 1\right)\right] = \sum_{i=1}^{N} q_j^{(i)} \big/ \pi_j - \eta = 0$$

より、

$$\pi_k = \frac{1}{\eta}\sum_{i=1}^{N} q_k^{(i)}, \qquad \sum_{k=1}^{K}\pi_k = \frac{1}{\eta}\sum_{k=1}^{K}\sum_{i=1}^{N} q_k^{(i)} = \frac{1}{\eta}\sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)} = \frac{1}{\eta}\sum_{i=1}^{N} 1 = \frac{N}{\eta} = 1$$

となり、以下の関係を得る。

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} q_k^{(i)}$$

②パラメータ $\boldsymbol{\theta}$ の推定

パラメータ $\boldsymbol{\theta}$ の最尤法による推定では、$q_k^{(i)}, \pi_k$ は①の方法で求められた既知の定数として計算する。この部分の計算については具体的な関数形を用いて考える。

4）混合指数分布に基づく最尤推定

指数分布の確率密度関数と生存関数の以下の具体的な表式を代入すると

$$f_k(t) = \lambda_k \exp(-\lambda_k t), \qquad S_k(t) = \exp(-\lambda_k t)$$

対数尤度は以下のようになる。

$$\log L(\boldsymbol{\lambda},\boldsymbol{\pi}) \geq \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\left[\delta_i\left(\log\lambda_k - \lambda_k t_i\right) - (1-\delta_i)\lambda_k t_i + \log\pi_k - \log q_k^{(i)}\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\left[\delta_i \log\lambda_k - \lambda_k t_i + \log\pi_k - \log q_k^{(i)}\right]$$

これより、群 k への帰属度は以下となる。

$$q_k^{(i)} = \pi_k\lambda_k \exp(-\lambda_k t_i) \Big/ \sum_{k=1}^{K}\pi_k\lambda_k \exp(-\lambda_k t_i) \qquad \text{for } \delta_i = 1$$

$$q_k^{(i)} = \pi_k \exp(-\lambda_k t_i) \Big/ \sum_{k=1}^{K}\pi_k \exp(-\lambda_k t_i) \qquad \text{for } \delta_i = 0$$

対数尤度を微分して、スコアベクトルを求め、それを 0 とする。

$$\frac{\partial}{\partial\lambda_j}\log L = \sum_{i=1}^{N} q_j^{(i)}\left(\delta_i/\lambda_j - t_i\right) = \frac{1}{\lambda_j}\sum_{i=1}^{N} q_j^{(i)}\delta_j - \sum_{i=1}^{N} q_j^{(i)} t_j = 0$$

これより、

$$\lambda_j = \sum_{i=1}^{N} q_j^{(i)} \delta_j \Big/ \sum_{i=1}^{N} q_j^{(i)} t_j$$

スコアをもう一度微分して、情報行列 $\Im$ に相当するものを作成する。

$$\Im_{jk} = -\frac{\partial^2}{\partial\lambda_j \partial\lambda_k} \log L = \frac{\delta_{jk}}{\lambda_j^2} \sum_{i=1}^{N} q_j^{(i)} \delta_i$$

この逆行列の対角成分は、推定値の分散を与える。

5) 混合 Weibull 分布に基づく最尤推定

K 種混合 Weibull では、以下となる。

$$f(t) = \sum_{k=1}^{K} \pi_k f_k(t) = \sum_{k=1}^{K} \pi_k a_k t^{a_k - 1} b_k^{-a_k} \exp(-t^{a_k} b_k^{-a_k})$$

$$= \sum_{k=1}^{K} \pi_k a_k t^{a_k - 1} e^{\beta_k} \exp(-t^{a_k} e^{\beta_k})$$

$$S(t) = \sum_{k=1}^{K} \pi_k S_k(t) = \sum_{k=1}^{K} \pi_k \exp(-t^{a_k} b_k^{-a_k}) = \sum_{k=1}^{K} \pi_k \exp(-t^{a_k} e^{\beta_k})$$

混合 Weibull 分布の対数尤度は以下となる。

$$\log L(\mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\pi}) \geq \sum_{i=1}^{N} \sum_{k=1}^{K} q_k^{(i)} \Big[\delta_i \left(\log a_k + (a_k - 1) \log t_i + \beta_k \right) - t_i^{a_k} e^{\beta_k} + \log \pi_k - \log q_k^{(i)} \Big]$$

これより、群 k への帰属度は以下となる。

$$q_k^{(i)} = \frac{\pi_k a_k t^{a_k - 1} e^{\beta_k} \exp(-t^{a_k} e^{\beta_k})}{\sum_{k=1}^{K} \pi_k a_k t^{a_k - 1} e^{\beta_k} \exp(-t^{a_k} e^{\beta_k})} \qquad \text{for } \delta_i = 1$$

$$q_k^{(i)} = \frac{\pi_k \exp(-t^{a_k} e^{\beta_k})}{\sum_{k=1}^{K} \pi_k \exp(-t^{a_k} e^{\beta_k})} \qquad \text{for } \delta_i = 0$$

ここで、$b_k^{-a_k} = e^{\beta_k}$ （$b_k = e^{-\beta_k / a_k}$ に相当）としている。

$$\frac{\partial}{\partial a_j} \log L(\mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{i=1}^{N} q_j^{(i)} \Big[\delta_i \left(1/a_j + \log t_i \right) - \log t_i \, t_i^{a_j} e^{\beta_j} \Big]$$

$$\frac{\partial}{\partial \beta_j} \log L(\mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{i=1}^{N} q_j^{(i)} \Big[\delta_i - t_i^{a_j} e^{\beta_j} \Big]$$

$$\frac{\partial^2}{\partial a_j \partial a_k}\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi}) = -\delta_{jk}\sum_{i=1}^{N} q_j^{(i)}\left[\delta_i / a_j^2 + (\log t_i)^2\, t_i^{a_j} e^{\beta_j}\right]$$

$$\frac{\partial^2}{\partial a_j \partial \beta_k}\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi}) = -\delta_{jk}\sum_{i=1}^{N} q_j^{(i)} \log t_i\, t_i^{a_j} e^{\beta_j}$$

$$\frac{\partial^2}{\partial \beta_j \partial \beta_k}\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi}) = -\delta_{jk}\sum_{i=1}^{N} q_j^{(i)} t_i^{a_j} e^{\beta_j}$$

2.4　比例ハザードモデル

比例ハザードモデルはハザート関数に対して、説明変数 $\mathbf{x} = {}^t(1, x_1, x_2, \cdots, x_p)$ とパラメータ $\boldsymbol{\beta} = {}^t(\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$ を用いて、以下の仮定を行う。

$$\lambda(t \mid \mathbf{x}, \boldsymbol{\beta}) = \lambda_0(t)\exp({}^t\mathbf{x}\boldsymbol{\beta}) \quad \text{ここに、} {}^t\mathbf{x}\boldsymbol{\beta} = \sum_{i=1}^{p} x_i\beta_i + \beta_0$$

Cox の比例ハザードモデルでは $\lambda_0(t)$ と定数項 β_0 について議論しないが、Weibull ハザードモデルでは以下のように時間に関して Weibull 分布のハザード関数を仮定する。

$$\lambda(t \mid \mathbf{x}, \boldsymbol{\beta}) = (a/b)\left(t/b\right)^{a-1} = at^{a-1}b^{-a} = at^{a-1}\exp({}^t\mathbf{x}\boldsymbol{\beta})$$

1）Cox の比例ハザードモデル

Cox の比例ハザードモデルでは、尤度関数に対して近似的な部分尤度関数を考えて処理を行う。その対数尤度は以下で与えられる[3]。

$$\log L'(\boldsymbol{\beta}) = \sum_{i=0}^{m-1}\left[\sum_{j\in D_i} {}^t\mathbf{x}_j\boldsymbol{\beta} - d_i \log\sum_{j\in R_i}\exp({}^t\mathbf{x}_j\boldsymbol{\beta})\right]$$

ここに、$\boldsymbol{\beta}$ は定数項を除いた偏回帰係数ベクトル、D_i は $t_i < t \le t_{i+1}$ で亡くなった個体の集合、R_i は時刻 t_i で生存が確認されている個体の集合である。これを最大化するようにニュートン・ラフソン法を使って $\boldsymbol{\beta}$ を求める。ここでは $w_j = \exp({}^t\mathbf{x}_j\boldsymbol{\beta})$ として以下の値を示しておく。

$$\mathbf{U} \equiv \frac{\partial}{\partial\boldsymbol{\beta}}\log L'(\boldsymbol{\beta}) = \sum_{i=1}^{m-1}\left[\sum_{i\in D_i}\mathbf{x}_j - d_i\sum_{j\in R_i} w_j\mathbf{x}_j \Big/ \sum_{j\in R_i} w_j\right]$$

$$\mathfrak{I} \equiv -\frac{\partial^2}{\partial\boldsymbol{\beta}\partial{}^t\boldsymbol{\beta}}\log L'(\boldsymbol{\beta})$$

$$= \sum_{i=1}^{m-1} d_i\left[\sum_{j\in R_i} w_j\mathbf{x}_j\,{}^t\mathbf{x}_j \Big/ \sum_{j\in R_i} w_j - \sum_{j\in R_i} w_j\mathbf{x}_j \sum_{j\in R_i} w_j\,{}^t\mathbf{x}_j \Big/ \left(\sum_{j\in R_i} w_j\right)^2\right]$$

この $\mathbf{U}$ をスコアベクトル、$\mathfrak{I}$ を情報行列という。$\boldsymbol{\beta}$ の推定値は以下の計算を繰り返して求める。

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathfrak{I}^{(m)})^{-1}\mathbf{U}^{(m)}$$

ここに右肩の添え字はニュートン・ラフソン法のループの段階を表している。

2) Weibull ハザードモデル

Weibull ハザードモデルは、ハザード関数に対して以下の仮定を行う。

$$\lambda(t) = \frac{f(t)}{S(t)} = (a/b)\left(t/b\right)^{a-1} = at^{a-1}/b^{a} = at^{a-1}\exp({}^{t}\mathbf{x}\boldsymbol{\beta})$$

通常の Weibull 分布との関係は以下である。

$$b^{-a} = e^{\beta} \rightarrow \exp({}^{t}\mathbf{x}\boldsymbol{\beta}) \quad （\beta \rightarrow {}^{t}\mathbf{x}\boldsymbol{\beta} \equiv \sum_{i=1}^{p} x_i\beta_i + \beta_0）$$

これより、$b = \exp(-{}^{t}\mathbf{x}\boldsymbol{\beta}/a)$ であるから、$\mu \equiv E[T] = b\,\Gamma\left(1+1/a\right)$ より、

$$\eta \equiv {}^{t}\mathbf{x}\boldsymbol{\beta} = -a\log b = -a\log\left(\mu/\Gamma(1+1/a)\right)$$

となり、右辺が一般化線形モデルの連結関数となる。

この関係を用いて、密度関数と生存関数を求めると以下となる。

$$f(t) = at^{a-1}\exp({}^{t}\mathbf{x}\boldsymbol{\beta})\exp\left[-t^{a}\exp({}^{t}\mathbf{x}\boldsymbol{\beta})\right]$$

$$S(t) = \exp\left[-\left(t/b\right)^{a}\right] = \exp\left[-t^{a}b^{-a}\right] = \exp\left[-t^{a}\exp({}^{t}\mathbf{x}\boldsymbol{\beta})\right]$$

打ち切りデータと非打ち切り（死亡）データをそれぞれ $\delta_i = 0,1$ と区別し、尤度を求めると以下となる。添え字 i について、ここでは個体の番号として使っている。

$$L(\alpha,\boldsymbol{\beta}) = \prod_{i=1}^{N} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

さらに、対数尤度は以下となる。

$$\begin{aligned}\log L(\alpha,\boldsymbol{\beta}) &= \sum_{i=1}^{N}\left[\delta_i \log\lambda(t_i) + \log S(t_i)\right] \\ &= \sum_{i=1}^{N}\left[\delta_i \log\left(at_i^{a-1}\exp({}^{t}\mathbf{x}_i\boldsymbol{\beta})\right) - t_i^{a}\exp({}^{t}\mathbf{x}_i\boldsymbol{\beta})\right] \\ &= \sum_{i=1}^{N}\left[\delta_i\left(\log a + (a-1)\log t_i + {}^{t}\mathbf{x}_i\boldsymbol{\beta}\right) - t_i^{a}\exp({}^{t}\mathbf{x}_i\boldsymbol{\beta})\right]\end{aligned}$$

対数尤度を微分してスコアベクトル $\mathbf{U}$ と情報行列 $\mathfrak{I}$ を求めると以下となる。

$$\boldsymbol{\beta}' = \begin{pmatrix} a \\ \boldsymbol{\beta} \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} \partial\log L/\partial a \\ \partial\log L/\partial\boldsymbol{\beta} \end{pmatrix},$$

$$\Im = -\begin{pmatrix} \partial^2 \log L/\partial a^2 & \partial^2 \log L/\partial a \partial {}^t\boldsymbol{\beta} \\ \partial^2 \log L/\partial a \partial \boldsymbol{\beta} & \partial^2 \log L/\partial \boldsymbol{\beta} \partial {}^t\boldsymbol{\beta} \end{pmatrix}$$

ここに、

$$\frac{\partial}{\partial a}\log L = \sum_{i=1}^{N}\left[\delta_i(1/a + \log t_i) - \log t_i t_i^a \exp({}^t\mathbf{x}_i\boldsymbol{\beta})\right]$$

$$\frac{\partial}{\partial \boldsymbol{\beta}}\log L = \sum_{i=1}^{N}\left[\delta_i\mathbf{x}_i - \mathbf{x}_i t_i^a \exp({}^t\mathbf{x}_i\boldsymbol{\beta})\right]$$

$$\frac{\partial^2}{\partial a^2}\log L = \sum_{i=1}^{N}\left[-\delta_i/a^2 - (\log t_i)^2 t_i^a \exp({}^t\mathbf{x}_i\boldsymbol{\beta})\right]$$

$$\frac{\partial^2}{\partial a \partial \boldsymbol{\beta}}\log L = -\sum_{i=1}^{N}(\log t_i)\mathbf{x}_i t_i^a \exp({}^t\mathbf{x}_i\boldsymbol{\beta})$$

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial {}^t\boldsymbol{\beta}}\log L = -\sum_{i=1}^{N}\mathbf{x}_i\,{}^t\mathbf{x}_i t_i^a \exp({}^t\mathbf{x}_i\boldsymbol{\beta})$$

これらを用いてニュートン・ラフソン法でパラメータの推定を行う。

$$\boldsymbol{\beta}'^{(m+1)} = \boldsymbol{\beta}'^{(m)} + (\Im^{(m)})^{-1}\mathbf{U}^{(m)}$$

ここに右肩の添え字はニュートン・ラフソン法のループの段階を表している。

求められたパラメータを使って、個人の予想寿命を以下のように求めることができる。

$$\mu \equiv E[T] = b\,\Gamma\left(1+1/a\right) = \exp(-{}^t\mathbf{x}\boldsymbol{\beta}/a)\Gamma\left(1+1/a\right)$$

この値を実際の寿命と比較することで相関係数等を求めることもできる。

3) 混合 Weibull ハザードモデル

K 種混合 Weibull ハザードモデルでは以下を仮定する。

$$f(t) = \sum_{k=1}^{K}\pi_k f_k(t) = \sum_{k=1}^{K}\pi_k a_k t^{a_k-1}\exp({}^t\mathbf{x}\boldsymbol{\beta} + \gamma_k)\exp(-t^{a_k}\exp({}^t\mathbf{x}\boldsymbol{\beta} + \gamma_k))$$

$$S(t) = \sum_{k=1}^{K}\pi_k S_k(t) = \sum_{k=1}^{K}\pi_k \exp(-t^{a_k}\exp({}^t\mathbf{x}\boldsymbol{\beta} + \gamma_k))$$

通常の Weibull 分と比較すると、ここでは以下を仮定している。

$$b_k^{-a_k} = e^{\beta_k} \to \exp({}^t\mathbf{x}\boldsymbol{\beta} + \gamma_k) \quad (\beta_k \to {}^t\mathbf{x}\boldsymbol{\beta} + \gamma_k \equiv \sum_{i=1}^{p} x_i\beta_i + \gamma_k)$$

これより、$b_k \to \exp\left[-({}^t\mathbf{x}\boldsymbol{\beta} + \gamma_k)/a_k\right]$ であるから、

$$\mu \equiv E[T] = \sum_{k=1}^{K}\pi_k b_k \Gamma\left(1+1/a_k\right)$$

となる。連結関数については、以下の関数の逆関数である。

$$\mu = \sum_{k=1}^{K} \pi_k \exp\left[-({}^t\mathbf{x}\boldsymbol{\beta} + \gamma_k)/a_k\right]\Gamma\left(1+1/a_k\right)$$

$$= \sum_{k=1}^{K} \pi_k \exp\left[-(\eta + \gamma_k)/a_k\right]\Gamma\left(1+1/a_k\right)$$

混合 Weibull 分布の対数尤度は以下となる。

$$\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\pi}) \geq \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\left[\delta_i\left(\log a_k + (a_k - 1)\log t_i + {}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_k\right)\right.$$

$$\left. - t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_k) + \log\pi_k - \log q_k^{(i)}\right]$$

これより、群 k への帰属度は以下となる。

$$q_k^{(i)} = \frac{\pi_k a_k t_i^{a_k - 1}\exp({}^t\mathbf{x}\boldsymbol{\beta} + \gamma_k)\exp(-t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_k))}{\sum_{k=1}^{K}\pi_k a_k t_i^{a_k - 1}\exp({}^t\mathbf{x}\boldsymbol{\beta} + \gamma_k)\exp(-t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_k))} \quad \text{for } \delta_i = 1$$

$$q_k^{(i)} = \frac{\pi_k \exp(-t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_k))}{\sum_{k=1}^{K}\pi_k \exp(-t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_k))} \quad \text{for } \delta_i = 0$$

ここで、$b_k^{-a_k} \to \exp({}^t\mathbf{x}\boldsymbol{\beta} + \gamma_k)$ （$b_k \to \exp\left[-({}^t\mathbf{x}\boldsymbol{\beta} + \gamma_k)/a_k\right]$）としている。

対数尤度を微分してスコアベクトル $\mathbf{U}$ と情報行列 $\mathfrak{I}$ を求めると以下となる。

$$\boldsymbol{\beta}' = \begin{pmatrix}\mathbf{a}\\ \boldsymbol{\gamma}\\ \boldsymbol{\beta}\end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix}\partial\log L/\partial\mathbf{a}\\ \partial\log L/\partial\boldsymbol{\gamma}\\ \partial\log L/\partial\boldsymbol{\beta}\end{pmatrix},$$

$$\mathfrak{I} = -\begin{pmatrix}\partial^2\log L/\partial\mathbf{a}\partial^t\mathbf{a} & \partial^2\log L/\partial\mathbf{a}\partial^t\boldsymbol{\gamma} & \partial^2\log L/\partial\mathbf{a}\partial^t\boldsymbol{\beta}\\ \partial^2\log L/\partial\boldsymbol{\gamma}\partial^t\mathbf{a} & \partial^2\log L/\partial\boldsymbol{\gamma}\partial^t\boldsymbol{\gamma} & \partial^2\log L/\partial\boldsymbol{\gamma}\partial^t\boldsymbol{\beta}\\ \partial^2\log L/\partial\boldsymbol{\beta}\partial^t\mathbf{a} & \partial^2\log L/\partial\boldsymbol{\beta}\partial^t\boldsymbol{\gamma} & \partial^2\log L/\partial\boldsymbol{\beta}\partial^t\boldsymbol{\beta}\end{pmatrix}$$

ここに、

$$\frac{\partial}{\partial a_j}\log L = \sum_{i=1}^{N} q_j^{(i)}\left[\delta_i\left(1/a_j + \log t_i\right) - \log t_i\, t_i^{a_j}\exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_j)\right]$$

$$\frac{\partial}{\partial \gamma_j}\log L = \sum_{i=1}^{N} q_j^{(i)}\left[\delta_i - t_i^{a_j}\exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_j)\right]$$

$$\frac{\partial}{\partial \boldsymbol{\beta}}\log L = \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\left[\delta_i\mathbf{x}_i - \mathbf{x}_i t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_j)\right]$$

$$\frac{\partial^2}{\partial a_j \partial a_k}\log L = -\delta_{jk}\sum_{i=1}^{N} q_j^{(i)}\left[\delta_i / a_j^2 + (\log t_i)^2\, t_i^{a_j}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_j)\right]$$

$$\frac{\partial^2}{\partial a_j \partial \gamma_k}\log L = -\delta_{jk}\sum_{i=1}^{N} q_j^{(i)}\log t_i\, t_i^{a_j}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_j)$$

$$\frac{\partial^2}{\partial a_j \partial \boldsymbol{\beta}}\log L = -\sum_{i=1}^{N} q_j^{(i)}\log t_i\, \mathbf{x}_i t_i^{a_j}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_j)$$

$$\frac{\partial^2}{\partial \gamma_j \partial \gamma_k}\log L = -\delta_{jk}\sum_{i=1}^{N} q_j^{(i)} t_i^{a_j}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_j)$$

$$\frac{\partial^2}{\partial \gamma_j \partial \boldsymbol{\beta}}\log L = -\sum_{i=1}^{N} q_j^{(i)}\mathbf{x}_i t_i^{a_j}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_j)$$

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial {}^t\boldsymbol{\beta}}\log L = -\sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\mathbf{x}_i\,{}^t\mathbf{x}_i t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_k)$$

これらを用いてニュートン・ラフソン法でパラメータの推定を行う。

$$\boldsymbol{\beta}'^{(m+1)} = \boldsymbol{\beta}'^{(m)} + (\Im^{(m)})^{-1}\mathbf{U}^{(m)}$$

ここに右肩の添え字はニュートン・ラフソン法のループの段階を表している。

3. プログラムの利用法

メニュー［分析－多変量解析他－生存時間分析］を選択すると、図１のような分析実行メニューが表示される。

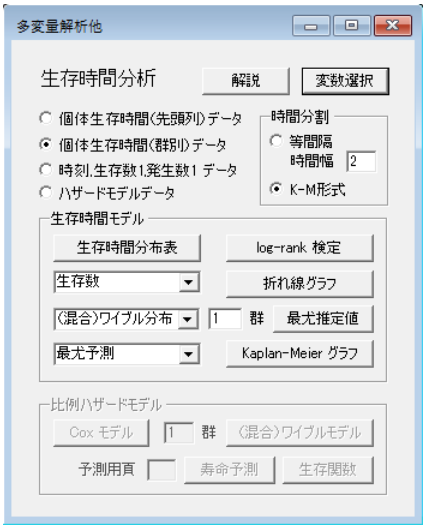


図１　分析実行メニュー

この分析のデータ形式は大きく分けて３種類ある。１つは個体の生存時間を元にしたデータで、これは先頭列で分類される形式とすでに群別に並べられている形式に分けられる。これらの形式は基本統計のデータ形式に類似している。次に、すでに生存時間分布表に近い形式になっているデータである。これは、観測時刻、その時点での生存個体数、その時点より後で次の時点までに死亡する期間発生数が、すでに表の形式になっているデータである。変数としては、時刻，生存個体数 1，期間発生数 1，生存個体数 2，期間発生数 2，… のようになっており、生存個体数と期間発生数は複数組入力が可能である。但し、これは実際の処理では、あまり使われることがないと思われる。最後は、ハザードモデルデータで、重回帰分析などと同様の形式である。最初と最後の形式で、通常のデータと異なる部分は、観測の打ち切りデータが含まれる点である。打ち切りデータは、観測を打ち切られた時点の数値の後ろに+記号を付けて表す。観測が打ち切られた際の扱いは、生存時間分布表や Kaplan-Meier 推定においては、生存数から打ち切られたデータ数の半分を引いて、死亡リスクに晒されたデータ数として処理している[1]。

最初に図２の単独データを元に説明をする。

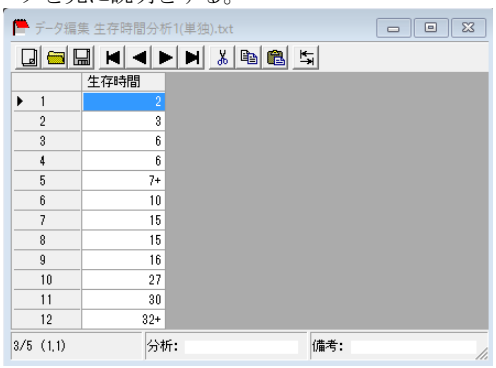


図２　単独データ（生存時間分析 1(単独).txt 3 頁目）

ここに、データのファイル名はホームページ[4]上にあるサンプル内のファイル名である。このデータでは、２個体が観測を打ち切られている。

「個体生存時間(群別)データ」ラジオボタンを選択し、変数選択を実行して、「生存時間分布表」ボタンをクリックすると図３のような結果が表示される。

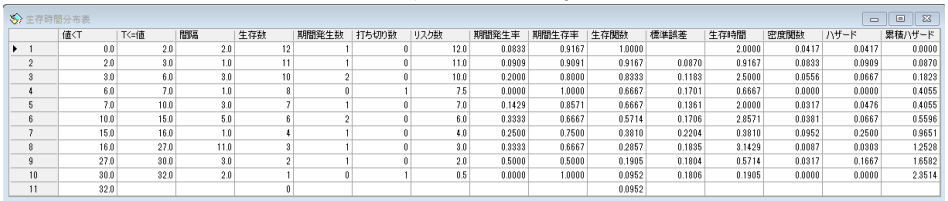


図３　生存時間分布表結果

図３では、様々な指標が区切られた時点毎に表示されている。ここで特に大切な指標は、「生存関数」と「ハザード」である。これらはそれぞれ、その時点まで生存している確率とその時点での死亡の危険率の意味を持つ。

図 3 の生存時間分布表の中で、生存数、生存関数、ハザード関数、累積ハザード関数については、コンボボックスで設定して、「折れ線グラフ」ボタンをクリックすると表示される。ここでは生存関数とハザード関数についてのグラフを図 4a と図 4b に示す。

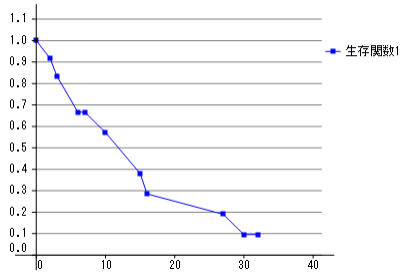


図 4a　生存関数

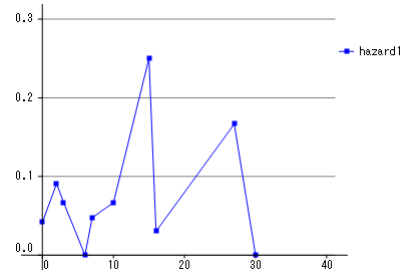


図 4b　ハザード関数

また、同じコンボボックスで「指数分布確認」または「Weibull 分布確認」を選択すると、図 5a と図 5b のような図が表示される。

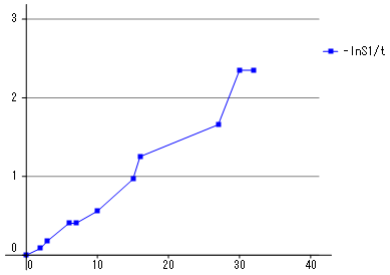


図 5a　生存関数確認

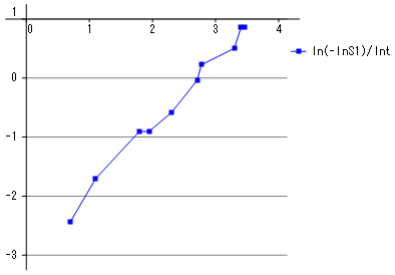


図 5b　ハザード関数確認

生存時間が指数分布または Weibull 分布に従うならば、それぞれの生存関数の時間依存性からこの点列は直線状に並ぶ。指数分布は Weibull 分布の特殊な場合であるので、指数分布が成り立つ場合は Weibull 分布も成り立つ。但し、Weibull 分布の場合の横軸は時間の対数である。

指数分布または Weibull 分布の確認の場合、「折れ線グラフ」をクリックすると、上図と共に分布の当てはまりの良さを示す、図 6a や図 6b のような指標も表示される。

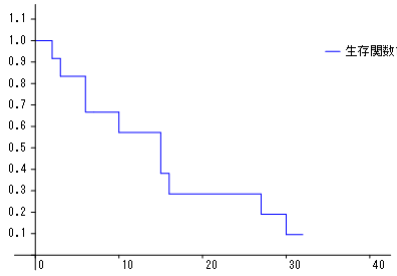
	メジアン	平均	直線性R	直線性R^2
▶ 群1	15.000	15.226	0.992	0.985

図 6a　指数分布の指標

	メジアン	平均	直線性R	直線性R^2
▶ 群1	15.000	15.226	0.993	0.986

図 6b　Weibull 分布の指標

生存関数の Kaplan-Meier 推定のグラフは、「Kaplan-Meier グラフ」ボタンをクリックして表示される。その際、左のコンボボックスで指定して、指数分布または Weibull 分布の予想曲線を描くこともできる。予想曲線のないグラフと、Weibull 分布の予想曲線を付けて描いたグラフを図 7a と図 7b に示す。



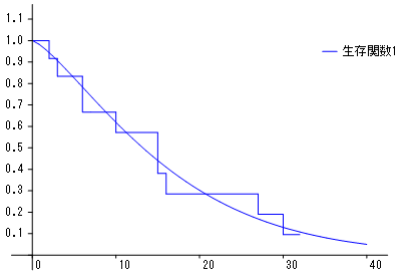


図 7a　Kaplan-Meier 生存関数グラフ　　図 7b　予想曲線付き Kaplan-Meier グラフ

これらの予想曲線では最小２乗法によるものと最尤法によるものとが選択できる。上図は最尤法によるものである。

また、予想曲線は混合指数分布や混合 Weibull 分布についても表示することができる。その際は分布を選んだコンボボックスの右のテキストボックスで混合する数を指定する。図８に２種の混合 Weibull 分布による予想曲線を付けた Kaplan-Meier グラフを表示する。サンプルでは２つの時期に危険度が高くなっている。

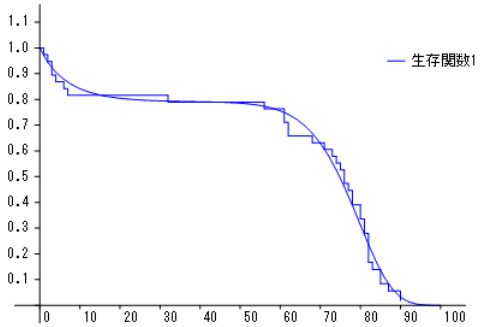


図８　２種混合分布による予測（生存時間分析 1(単独).txt 8 頁目）

このパラメータの値については、上と同じ設定で「最尤推定値」ボタンをクリックすると、図９のように表示される。

混合ワイブル分布推定結果

	推定値	標準偏差	5%下限	5%上限
▶ 生存時間	R	0.993	R^2	0.986
出現確率1	0.790			
a1	10.545	1.623	7.364	13.726
b1=exp(-β/a)	80.564			
β1	-46.283	7.171	-60.339	-32.227
出現確率2	0.210			
a2	0.937	0.233	0.480	1.394
b2=exp(-β/a)	6.964			
β2	-1.819	0.681	-3.154	-0.483

図９　２種混合 Weibull 予測（生存時間分析 1(単独).txt 8 頁目）

ここでは表示されていないが、混合がない場合には、右端に最小２乗推定による推定値も表示される。

複数群の生存時間分布表は、先頭列で群分けデータ（生存時間分析 2(2 群比較).txt）または群別データを元に図 10 のように縦に並べて表示される。

	値<T	T<=値	間隔	生存数	期間発生数	打ち切り数	リスク数	期間発生率	期間生存率	生存関数	標準誤差	生存時間	密度関数	ハザード	累積ハザード
1	0.0	1.0	1.0	12	0	0	12.0	0.0000	1.0000	1.0000		1.0000	0.0000	0.0000	0.0000
2	1.0	2.0	1.0	12	1	0	12.0	0.0833	0.9167	1.0000	0.0000	1.0000	0.0833	0.0833	0.0000
3	2.0	3.0	1.0	11	1	0	11.0	0.0909	0.9091	0.9167	0.0870	0.9167	0.0833	0.0909	0.0870
4	3.0	6.0	3.0	10	2	0	10.0	0.2000	0.8000	0.8333	0.1183	2.5000	0.0556	0.0667	0.1823
5	6.0	7.0	1.0	8	1	0	8.0	0.1250	0.8750	0.6667	0.1701	0.6667	0.0833	0.1250	0.4055
6	7.0	9.0	2.0	7	0	0	7.0	0.0000	1.0000	0.5833	0.1627	1.1667	0.0000	0.0000	0.5390
7	9.0	10.0	1.0	7	1	0	7.0	0.1429	0.8571	0.5833	0.1423	0.5833	0.0833	0.1429	0.5390
8	10.0	15.0	5.0	6	2	0	6.0	0.3333	0.6667	0.5000	0.1684	2.5000	0.0333	0.0667	0.6931
9	15.0	16.0	1.0	4	1	0	4.0	0.2500	0.7500	0.3333	0.2041	0.3333	0.0833	0.2500	1.0986
10	16.0	22.0	6.0	3	0	0	3.0	0.0000	1.0000	0.2500	0.1667	1.5000	0.0000	0.0000	1.3863
11	22.0	27.0	5.0	3	1	0	3.0	0.3333	0.6667	0.2500	0.1250	1.2500	0.0167	0.0667	1.3863
12	27.0	30.0	3.0	2	1	0	2.0	0.5000	0.5000	0.1667	0.1614	0.5000	0.0278	0.1667	1.7918
13	30.0	32.0	2.0	1	1	0	1.0	1.0000	0.0000	0.0000	0.1596	0.0000	0.0000	0.5000	0.0000
14	32.0			0						0.0000					
1	0.0	1.0	1.0	9	4	0	9.0	0.4444	0.5556	1.0000		1.0000	0.4444	0.4444	0.0000
2	1.0	2.0	1.0	5	1	0	5.0	0.2000	0.8000	0.5556	0.2981	0.5556	0.1111	0.2000	0.5878
3	2.0	3.0	1.0	4	2	0	4.0	0.5000	0.5000	0.4444	0.2070	0.4444	0.2222	0.5000	0.8109
4	3.0	6.0	3.0	2	0	0	2.0	0.0000	1.0000	0.2222	0.2772	0.6667	0.0000	0.0000	1.5041
5	6.0	7.0	1.0	2	0	0	2.0	0.0000	1.0000	0.2222	0.1386	0.2222	0.0000	0.0000	1.5041
6	7.0	9.0	2.0	2	1	0	2.0	0.5000	0.5000	0.2222	0.1386	0.4444	0.0556	0.2500	1.5041
7	9.0	10.0	1.0	1	0	0	1.0	0.0000	1.0000	0.1111	0.2095	0.1111	0.0000	0.0000	2.1972
8	10.0	15.0	5.0	1	0	0	1.0	0.0000	1.0000	0.1111	0.1049	0.5556	0.0000	0.0000	2.1972
9	15.0	16.0	1.0	1	0	0	1.0	0.0000	1.0000	0.1111	0.1049	0.1111	0.0000	0.0000	2.1972
10	16.0	22.0	6.0	1	1	0	1.0	1.0000	0.0000	0.0000	0.1049	0.0000	0.0000	0.1667	0.0000
11	22.0	27.0	5.0	0						0.0000					
12	27.0	30.0	3.0												
13	30.0	32.0	2.0												
14	32.0														

図 10　2 群の生存時間分布表

これ以外に、もっと群の違いを比較できる方法を考えて行きたい。

複数群の生存関数と Kaplan-Meire 生存関数グラフを図 11 と図 12 に示す。

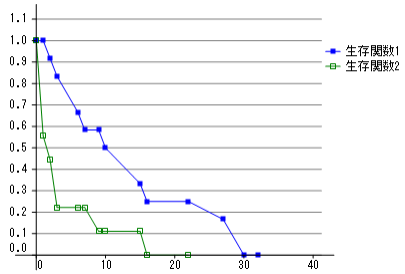


図 11　2 種類の生存関数グラフ

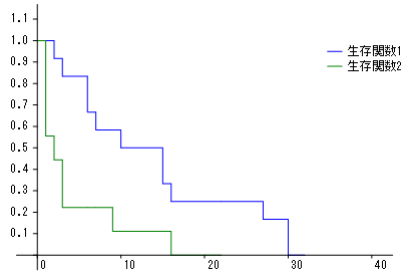


図 12　2 種類の Kaplan-Meier グラフ

複数群の生存関数間の差の log-rank 検定結果は、「log-rank 検定」ボタンをクリックすると図 13 のように表示される。

図 13　log-rank 検定結果

最後に、比例ハザードモデルの分析結果について示しておく。データは図 14 のような重回帰分析などと同じデータ形式である。

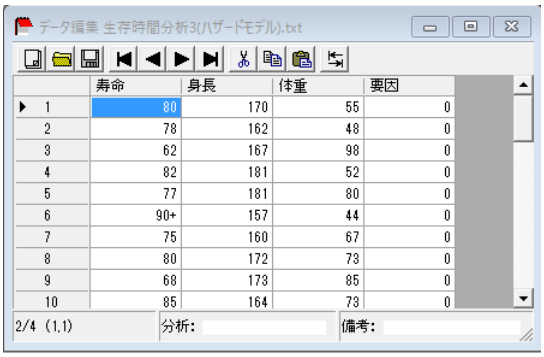


図 14　比例ハザードモデルデータ（生存時間分析 3(ハザードモデル).txt）

ハザードモデルでは Cox 比例ハザードモデルと Weibull 比例ハザードモデルを組み込んでいる。ハザード関数について、2 つのモデルとも、前章でのべたように、以下の形を仮定する。

$$\lambda(t\,|\,\mathbf{x},\boldsymbol{\beta}) = \lambda_0(t)\exp({}^t\mathbf{x}\boldsymbol{\beta}) \quad \text{ここに、} {}^t\mathbf{x}\boldsymbol{\beta} = \sum_{i=1}^{p} x_i\beta_i + \beta_0$$

Cox 比例ハザードモデルは $\lambda_0(t)$ や β_0 の推定は行わないが、分布の形に依存しない利点がある。Weibull ハザードモデルでは、時間部分に Weibull 分布を仮定し、その 1 つのパラメータを説明変数で推定するという一般化線形モデルの形式を採用している。

$$\lambda(t\,|\,\mathbf{x},\boldsymbol{\beta}) = (a/b)\left(t/b\right)^{a-1} = at^{a-1}b^{-a} = at^{a-1}\exp({}^t\mathbf{x}\boldsymbol{\beta})$$

「Cox モデル」ボタンをクリックした結果を図 15 に、「Weibull モデル」ボタンをクリックした結果を図 16 に示す。

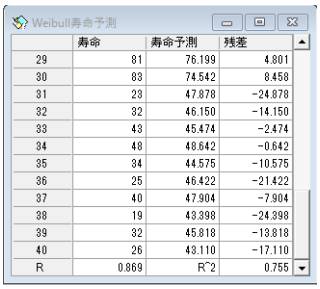
	偏回帰係数	標準偏差	両側確率	2.5%下限	2.5%上限	EXP(b)
身長	-0.0038	0.0227	0.8669	-0.0483	0.0407	0.9962
体重	0.0294	0.0133	0.0272	0.0033	0.0555	1.0299
要因	2.4581	0.4826	0.0000	1.5123	3.4039	11.6825

図 15　Cox 比例ハザードモデル結果

	偏回帰係数	標準偏差	両側確率	2.5%下限	2.5%上限	EXP(b)
身長	-0.0040	0.0216	0.8547	-0.0464	0.0385	0.9960
体重	0.0211	0.0119	0.0773	-0.0023	0.0445	1.0213
要因	2.0780	0.3916	0.0000	1.3104	2.8455	7.9884
切片	-21.4299	4.3297	0.0000	-29.9161	-12.9437	4.933E-10
a	4.7354	0.6323	0.0000	3.4961	5.9747	

図 16　Weibull 比例ハザードモデル

最後に Weibull 比例ハザードモデルが予想する生存時間の平均値と実際の観測値との比較を行ってみる。「予測用頁」テキストボックスを空欄のまま、「寿命予測」ボタンをクリックすると図 17a と図 17b の結果が示される。


Weibull寿命予測

	寿命	寿命予測	残差
29	81	76.199	4.801
30	83	74.542	8.458
31	23	47.878	−24.878
32	32	46.150	−14.150
33	43	45.474	−2.474
34	48	48.642	−0.642
35	34	44.575	−10.575
36	25	46.422	−21.422
37	40	47.904	−7.904
38	19	43.398	−24.398
39	32	45.818	−13.818
40	26	43.110	−17.110
R	0.869	R^2	0.755

図 17a　寿命予測図

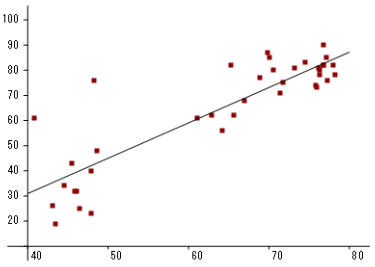


図 17b　実測/予測散布図

これには非打ち切り（死亡）データのみが用いられている。また、寿命予測の結果の最後に、予測値と実測値の相関係数の値とその 2 乗の値を表示している。寿命予測には、予測したいデータを別の頁に作っておき、「予測用頁」テキストボックスにその頁番号を入力して、「寿命予測」ボタンをクリックする方法もある。

このWeibullハザードモデルと混合Weibull分布のKaplan-Meier推定とを比較してみる。「予測用頁」テキストボックスを空欄のまま、「生存関数」ボタンをクリックし、各個体の生存関数を描画すると図 18 のようになる。また Weibull 分布を使った Kaplan-Meier 推定は図 19 のようになる。

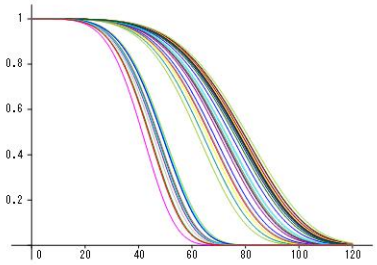


図 18　各個体の生存関数

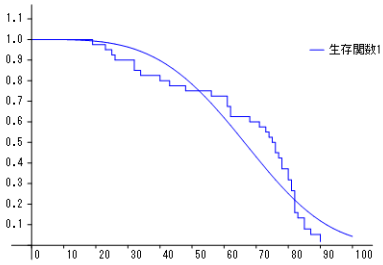


図 19　Weibull 分布による推定

Kaplan-Meier 推定はあまり適合しておらず、各個体の生存関数も当然のことであるが、寿命が長いか短いかしか分からず、あまり面白みがない。

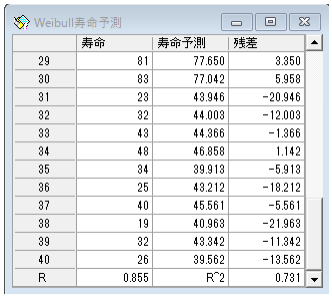
次に、このモデルに例えば 2 種混合 Weibull ハザードモデルを適用してみよう。この場合、比例ハザードモデルの中の「群」テキストボックスに「2」を入れて、「(混合) Weibull モデル」ボタンをクリックする。図 20 に結果を示す。

混合ワイブル比例ハザードモデル結果

	偏回帰係数	標準偏差	両側確率	2.5%下限	2.5%上限	EXP(b)
▶ 身長	0.0274	0.0206	0.1830	−0.0129	0.0678	1.0278
体重	0.0526	0.0132	0.0001	0.0266	0.0785	1.0540
要因	5.2765	0.7407	0.0000	3.8248	6.7283	195.6922
出現確率1	0.2497					
a1	29.7025	5.6218		18.6837	40.7212	
γ1	−140.1803	25.9021		−190.9484	−89.4123	
出現確率2	0.7503					
a2	7.3947	1.0264		5.3829	9.4065	
γ2	−40.1365	6.2861		−52.4574	−27.8157	

図 20　混合 Weibull ハザードモデル（生存時間分析 3(ハザードモデル).txt 2 頁目）

このモデルによる実測・予測値と重相関係数 R の値、及びそのグラフを表示するには、「予測用頁」テキストボックスを空欄のまま、「寿命予測」ボタンをクリックする。結果は図 21 のようになる。



	寿命	寿命予測	残差
29	81	77.650	3.350
30	83	77.042	5.958
31	23	43.946	−20.946
32	32	44.003	−12.003
33	43	44.366	−1.366
34	48	46.858	1.142
35	34	39.913	−5.913
36	25	43.212	−18.212
37	40	45.561	−5.561
38	19	40.963	−21.963
39	32	43.342	−11.342
40	26	39.562	−13.562
R	0.855	R^2	0.731

図 21a　寿命予測図

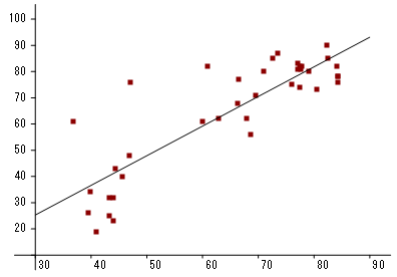


図 21b　実測/予測散布図

このモデルと 2 種混合 Weibull 分布の Kaplan-Meier 推定とを比較してみる。「予測用頁」テキストボックスを空欄のまま、「生存関数」ボタンをクリックし、各個体の生存関数を描画すると図 22 のようになる。また 2 種混合 Weibull 分布を使った Kaplan-Meier 推定は図 23 のようになる。

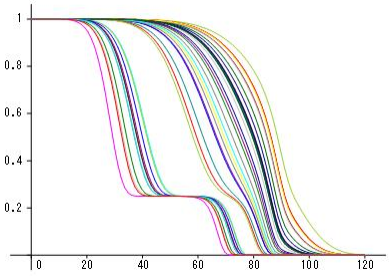


図 22　各個体の生存関数

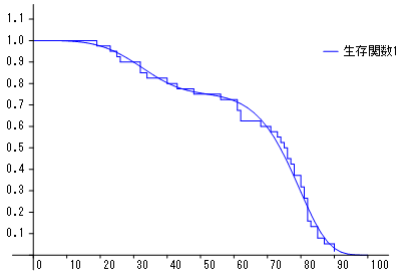


図 23　混合 Weibull 分布による推定

図 20 で生存関数が途中で折れ曲がっている個体は、寿命に及ぼす影響が大きいと思われる、データ中の「要因」因子が「1」の個体である。このようにして見ると、「要因」因子が「1」の個体と「0」の個体で生存関数が大きく異なっていることが分かる。

それでは、図 22 の個体の生存関数グラフとハザードモデルを考えない図 23 の混合 Weibull 分布モデルのグラフとの関係はどのようになっているのであろうか。図 23 のすべての曲線の時間ごとの平均を取ると図 24 のようになり、図 23 のハザードモデルを考えない形に非常に近くなることが分かる。

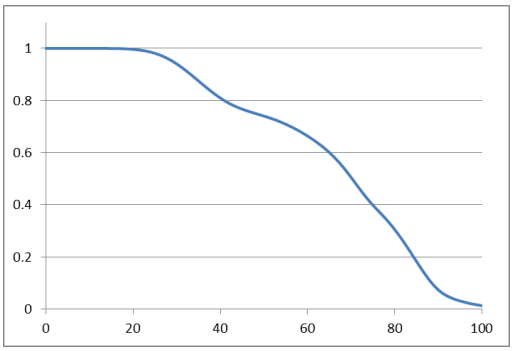


図 24　各個体の生存関数の平均

この結果はグラフのデータ出力機能を利用し、結果を Excel に貼り付けて平均を計算し描画したものである。

4.　おわりに

この論文では、Kaplan-Meier 推定法による生存分析表や生存関数グラフ、log-rank 検定、分布の最尤推定、Cox の比例ハザードモデル、Weibull ハザードモデルなど、生存時間分析法とそのプログラムについて解説して来た。特徴としては、混合 Weibull 分布をハザードモデルに適用し、使い易いプログラムにした部分である。これによって、寿命に影響を及ぼす変数がある場合、生存関数グラフが大きな変更を受ける可能性が見えてきた。視覚的なグラフにより危険性が示されることは、ハザードや平均寿命の数値だけによる表示に比べて注意を促しやすい。これがどの程度示されるのか、今後詳しく調べてみる必要がある。

このプログラムでは、生存時間分布表や生存関数の推定方法を、最もよく利用される Kaplan-Meier 推定法に限定している。また、それに当てはめる分布も、指数分布と Weibull 分布及び、それらの混合分布だけであり、他の分布は考えていない。今後これらの点については必要に応じて追加して行く予定である。また、生存関数、ハザード、推定されるパラメータの 95%信頼区間についても、すべてを取り上げていない。この問題についても今後の課題として残っている。改良すべき点を、段階を追って改良し、使い易いプログラムとして行きたい。

参考文献

[1] 打波守, Excel で学ぶ生存時間解析, オーム社, 2005.
[2] 柳井晴夫, 高木廣文編著, 多変量解析ハンドブック, 現代数学社, 1986.
[3] Annete J. Dobson, 田中豊他訳, 一般化線形モデル入門 原著第 2 版, 共立出版, 2008.
[4] ホームページ http://www.heisei-u.ac.jp/ba/fukui/analysis.html 内のサンプルデータ Samples.zip 内のファイル

社会システム分析のための統合化プログラム３１
－生存時間分析－

Multi-purpose Program for Social System Analysis 31
- Survival Analysis -

Masayasu FUKUI[*1] and Xiaona WU[*2]

[*1] *Department of Business Administration, Faculty of Business Administration, Fukuyama Heisei University*
[*2] *Division of Business Information Study of Graduate School of Business Administration, Fukuyama Heisei University*

Abstract: We have been constructing a unified program on the social system analysis for the purpose of education. This time, we report about the program of survival analysis which is a method to predict mortality rate and survival probability distribution from sample including abort data. The program includes survival time table, Kaplan-Meier graph including prediction curve, Cox proportional hazard model, Weibull hazard model by maximum likelihood method and mixed Weibull hazard model.

Keywords: College Analysis, multivariate analysis, survival analysis, compound Weibull distribution, Cox proportional hazard model, Weibull hazard model, EM Algorithm
URL: http://www.heisei-u.ac.jp/ba/fukui/