ニューラルネットによる生存時間解析

大阪電気通信大学 総合情報学部 辻谷将明※
西宮市立中央病院 外科 左近賢人
※印が発表者

要 旨 従来、生存時間解析では Cox **比例ハザードモデル**が広範に活用されてきた。特に、共変量の値が時間とともに変動する**時間依存型**データが含まれる場合、その近似解法として **Mayo updated モデル**や**ヨーロッパ new version モデル**が広範に活用されてきた。しかし、それらのモデルには、ベースライン生存関数やベースライン累積ハザード関数の推定などに問題点が残されている。本稿では、**部分**ロジスティック回帰モデルを援用した**部分ロジスティックモデル**および**ニューラルネットモデル**を提案し、ブートストラップ法による統計的推測を系統的に行う。実際例として、PBC(原発性胆汁性肝硬変)データを取上げる。肝移植を念頭においた、観測期間の任意時点における 6 ヶ月後の条件付き生存率の予後予測を通じ、提案手法を既存手法と数値的に比較する。

1. はじめに

生存時間解析には、Cox 比例ハザードモデルが広範に活用されてきた(大橋,浜田,1995;中村, 2001;Kalbfleish and Prentice, 2002;Klein and Moeschberger, 2003)。Cox 比例ハザードモデルでは、生存時間(観測時点)t のハザード関数を

$$h(\boldsymbol{x},t)=\lambda_0(t)\exp\left(\sum_{i=1}^{I}\beta_i x_i\right) \tag{1}$$

と表す。ここで、ベースラインハザード関数 $\lambda_0(t)$ は観測時点 t の関数であって、共変量 $\boldsymbol{x}=(x_1,\cdots,x_I)$ を含んでいない。このモデルでは、共変量の観測は各患者について 1 回のみで、時間が変化しても不変である。しかし、反復して測定される検査値の生存時間への影響の評価、あるいは観測期間中に薬剤の投与量を変化させたとき、その効果の有無の検証が必要な場面もでてくる(Christensen et al., 1986)。このように、共変量の値が時間とともに変動する**時間依存型**データが含まれる場合、その近似解法として Mayo updated モデル(Murtaugh et al., 1994)やヨーロッパ new version モデル(Christensen et al., 1986, 1993; Altman and De Stavola, 1994)が広範に活用されてきた。

しかし、共変量の値が時間とともに変動する時間依存型データが含まれる場合、各患者に対する生存関数はベースラインハザード関数のみならず共変量の値にも依存し、厳密にはベースライン生存関数のベキ乗の形式では表現できない(Collett, 1994, 7.2 節)。また、ベースライン生存関数は、明確で有効な解釈をもたない(Kalbfleish and Prentice, 2002, 6.4.1 節; Marubini and Valsecchi, 1995, 6.8.2 節)。

この点を回避するため、PBC データの解析を目的にした Mayo updated モデルでは、一人の患者から生成された複数個の観測ベクトルを互いに独立として取扱い、推定されたベースライン生存関数から予後予測を行っている。そのため、各患者に関し、時間と共に変動する共変量の履歴が解析に反映されていない点が問題である(Altman and De Stavola, 1994)。

一方、時間依存型共変量が含まれる場合でも、Breslow(Nelson-Aalen)推定量を拡張すれば、ベースライン累積ハザード関数は求められる(Arjas,1988; Kalbfleish and Prentice, 2002, 6.4.1 節)。当初開発されたヨーロッパ new version モデルは、時間と共に変動する共変量の履歴を解析に反映させ、ハザード関数に基づく予後予測の近似が試みられている。しかし、ベースラインハザード関数が局所的に一定という仮定が必要である。また、その一定値をグラフから目算するため、解析者による恣意性が生じてくる。

本稿では、二項分布に基づく部分ロジスティック回帰モデル(Efron, 1988)を援用し、時間依存型共変量を伴う生存データの解析について検討する。Cox 比例ハザードモデルでは、生存時間の順位を用いているが、生存時間値そのものを取込むことにより、*i*) **部分尤度**(Cox, 1975)から構築される**部分ロジスティックモデル**が、**グループ化されていない**データ(ungrouped data)に適用可能であること

を示す。また近年、Cox 比例ハザードモデルに対抗して、ニューラルネットによる生存時間解析が注目されつつある(Biganzoli et al., 1998, 2002)。時間依存型共変量を伴う生存データの解析に *ii*) 柔軟な**ニューラルネットモデル**が威力を発揮することを明確にし、部分ロジスティックモデルとニューラルネットモデルとの関係についても触れる。そして、*iii*) ブートストラップ法を援用し、想定した部分ロジスティックモデルやニューラルネットモデルの包括的な適合度を検定する。ニューラルネットモデルにおける隠れ層のユニット数も系統的に決定できる。実際例として、Mayo クリニックに来院した 312 例の PBC データを取上げる(Therneau and Grambsch, 2000, 5.6 節)。PBC は、他の臓器に重篤な合併症がないことなどから、肝移植の良い適応となる疾患である(市田,谷川, 1991)。そこで、*iv*) 肝移植時期の決定を念頭においた、観測期間の任意時点での 6 ヶ月後の生存率の予測が、本稿の究極の狙いである。

2. モデル構築

2.1 定式化

患者#*d* の来院日(観測時点) $t_1^{<d>},...,t_l^{<d>},...,t_{l_d}^{<d>}$ において、l 番目の時間区間に対する中央値を $a_l^{<d>}=\left(t_{l-1}^{<d>}+t_l^{<d>}\right)/2$、観測起点を $t_0^{<d>}\equiv 0$ とし、時間依存型共変量を $\boldsymbol{X}_l^{<d>}=\left(a_l^{<d>},\boldsymbol{x}_l^{<d>}\right)$ とする。ここに、$\boldsymbol{x}_l^{<d>}=\left(x_{l1}^{<d>},\cdots,x_{lI}^{<d>}\right)$ は時間依存型共変量である。例えば、表 1 は Murtaugh et al.(1994)が取上げた患者#9(死亡例)の共変量の値を示している。患者は原則として 6 ヶ月目、12 ヶ月目、その後、1 年おきに来院する。そして、来院ごとに、年齢と共にプロトロンビン時間、ビリルビン値、アルブミン値、エデマ・スコアを測定する。ただし、

$$\text{エデマ・スコア}=\begin{cases}0:\text{浮腫なし、かつ浮腫に対する利尿剤の投与(–)}\\0.5:\text{利尿剤の投与がない状況で浮腫がみられる、あるいは利尿剤で浮腫が軽快する}\\1.0:\text{利尿剤投与を投与しても浮腫がみられる}\end{cases}$$

とする。表 1 では、1 人の患者が、7 個の観測ベクトルを生成し、共変量の値は時間区間と共に変動している。このようなデータについて、ある観測時点 t まで生存していた患者の、次の Δt (例えば、6 ヶ月)後の条件付き生存率を予測するため、従来は、Mayo updated モデルやヨーロッパ new version モデルが広範に活用されてきた(市田, 谷川, 1991; 井上, 1994)。

表 1. 患者#9(死亡例)に関する共変量の値

	時間区間 l (来院日 $t_l^{<9>}$)						
	1(0)	2(184)	3(361)	4(723)	5(1027)	6(1396)	7(2278)
中央値(日) $a_l^{<9>}$	92.0	272.5	542.0	875.0	1211.5	1837.0	2339.0
年齢(歳) $x_{l1}^{<9>}$	42.5	43.0	43.5	44.5	45.4	46.4	48.8
ﾌﾟﾛﾄﾛﾝﾋﾞﾝ時間 $x_{l2}^{<9>}$	11.0	12.5	11.2	14.1	11.5	11.5	13.0
ﾋﾞﾘﾙﾋﾞﾝ値 $x_{l3}^{<9>}$	3.2	7.0	4.2	13.5	12.0	16.2	14.8
ｱﾙﾌﾞﾐﾝ値 $x_{l4}^{<9>}$	3.08	3.64	3.10	2.87	2.96	2.99	2.41
ｴﾃﾞﾏ･ｽｺｱ $x_{l5}^{<9>}$	0	0	0	0	0	0.5	1

(1) Mayo updated モデル

観測時点(来院日)t における共変量 $\boldsymbol{x}(t)=\left(x_1(t),x_2(t),...,x_I(t)\right)$ の値が与えられたとき、Mayo updated モデルを用い、PBC データを解析しよう。表 1 の患者#9(死亡例)の場合、初診時 $t_0^{<9>}=0$ から次の来院日(来院回数 1) $t_1^{<9>}=184$ までの時間 184 を $l=1$ に対する生存時間(打切り)、$t_1^{<9>}=184$ から次の来院日(来院回数 2) $t_2^{<9>}=361$ までの時間 177(=361-184)を $l=2$ に対する生存時間(打切り)など

とする。そして、最終の$l=7$に対する生存時間は122$(=2400-2278)$で、死亡となる。時間区間lにおける生存時間は、共変量として$x_{l1}^{<9>}, x_{l2}^{<9>}, ..., x_{l5}^{<9>}$をもつ。よって、$n=312$例について、合計1945個の観測ベクトルが得られ、これらを従来の時間固定(time-fixed)型の比例ハザードモデルで解析する。生存関数は、$S\left(\boldsymbol{x}(t),t\right)=\left\{S_0(t)\right\}^{\exp(PI)}$と書ける。ここに、予後指数$PI$ (Prognostic Index)を

$$PI=\beta_1\left\{x_1(t)-\bar{x}\right\}+\beta_2\left\{x_2(t)-\bar{x}_2\right\}+\cdots+\beta_I\left\{x_I(t)-\bar{x}_I\right\}$$

と定義する。ベースライン生存関数$S_0(t)$は、すべての共変量$\boldsymbol{x}(t)=\left(x_1(t),x_2(t),...,x_I(t)\right)$がそれぞれの平均値$\bar{\boldsymbol{x}}=\left(\bar{x}_1,\cdots,\bar{x}_I\right)$を取ったときの生存関数である。このベースライン生存関数$S_0(\Delta t)$を用い、観測時点tまで生存していた患者の、Δt後の条件付き生存率

$$\Pr\left(t,t+\Delta t\right)=\left\{S_0(\Delta t)\right\}^{\exp(PI)} \tag{2}$$

を予測する。

(2)ヨーロッパ new version モデル

時間と共に変動する共変量が含まれる場合でも、Breslow(Nelson-Aalen)推定量を拡張すれば、ベースライン累積ハザード関数は求められる(Arjas,1988; Kalbfleish and Prentice, 2002,6.4.1節)。ヨーロッパ new version モデルでは、ハザード関数に基づく予後予測の近似が試みられている。表1の患者#9に関し、観測時点tまで生存していた患者の、Δt後の条件付き生存率$\Pr\left(t,t+\Delta t\right)$を予測しよう。(1)式においてベースラインハザード関数$\lambda_0(t)$が一定λ^*(すなわち、Δtが短期間で、ベースライン累積ハザード関数$\Lambda_0(t)$がtの一次関数として線形補間できる)と仮定したハザード関数の近似式

$$h\left(\boldsymbol{x}(t),t\right)=\lambda^*\exp\left\{\sum_{i=1}^{I}\beta_i x_i(t)\right\} \tag{3}$$

がヨーロッパ new version モデルである。よって、Δt後の条件付き生存率$\Pr\left(t,t+\Delta t\right)$は

$$\Pr\left(t,t+\Delta t\right)=\exp\left[-\lambda^*\cdot\Delta t\cdot\exp\left\{-\sum_{i=1}^{I}\beta_i x_i(t)\right\}\right] \tag{4}$$

となる。

2.2 部分ロジスティックモデル

二項分布に基づく部分ロジスティック回帰モデル(Efron, 1988)をベルヌーイ分布に援用し、時間依存型共変量をもつグループ化されていないデータについて、(離散)ハザード関数に対する部分ロジスティックモデル

$$h_l^{<d>}=\frac{1}{1+\exp\left\{-\left(\gamma a_l^{<d>}+\sum_{i=0}^{I}\beta_i x_{li}^{<d>}\right)\right\}},\ x_{l0}^{<d>}\equiv 1;\ l=1,2,\cdots,l_d;\ d=1,2,\cdots,n \tag{5}$$

を提案する。ここに、回帰係数は$\boldsymbol{\beta}=\left(\gamma,\beta_0,\beta_1,...,\beta_I\right)$、$n$は総患者数、$l_d$は患者#$d$の時間区間の個数である(本例の場合、$n=312$で、例えば表1の患者#9について、$l_9=7$となる)。(5)式では、Efron(1988)に準拠し、l番目の時間区間に対する中央値$a_l^{<d>}=\left(t_{l-1}^{<d>}+t_l^{<d>}\right)/2$を共変量として考慮している。

さて、患者#dの時間区間lについて

$$\delta_l^{<d>}=\begin{cases}1:\text{患者 \#}d\text{が時間区間}l\text{で死亡}\\0:\text{その他}\end{cases}\quad、\quad \delta_l'^{<d>}=\begin{cases}1:\text{患者\#}d\text{が時間区間}l\text{で打切り}\\0:\text{その他}\end{cases}$$

と定義する。患者#dの最初の時間区間$l-1$までの死亡あるいは打切りに関する履歴

$\boldsymbol{v}_l^{<d>} = \left(\delta_1^{<d>}, \delta_1'^{<d>}, \delta_2^{<d>}, \delta_2'^{<d>}, \cdots, \delta_{l-1}^{<d>}, \delta_{l-1}'^{<d>}\right)$は、$(0,0,...,0)$となる。$\delta_l^{<d>}$が含まれるまで$\boldsymbol{v}_l^{<d>}$を拡張した量$\boldsymbol{v}_l'^{<d>} = \left(\boldsymbol{v}_l^{<d>}, \delta_l^{<d>}\right)$は$\left(0,0,...,0,\delta_l^{<d>}\right)$である。履歴$\boldsymbol{v}_l^{<d>} = (0,0,...,0)$が与えられたとき、$\delta_l^{<d>}$の分布はベルヌーイ分布$B_e\left(1, h_l^{<d>}\right)$に従う。ただし、$h_l^{<d>}$はハザード関数(5)である。$\boldsymbol{v}_l'^{<d>} = \left(0,0,...,0,\delta_l^{<d>}\right)$が与えられたとき$\delta_l'^{<d>}$は、ある分布$p\left(\delta_l'^{<d>} \middle| \boldsymbol{v}_l'^{<d>}\right)$に依存するが、回帰係数$\boldsymbol{\beta}$には依存しないと仮定する。患者#$d$に対する$\left(\delta_1^{<d>}, \delta_1'^{<d>}, \delta_2^{<d>}, \delta_2'^{<d>}, ..., \delta_{l_d-1}^{<d>}, \delta_{l_d-1}'^{<d>}, \delta_{l_d}^{<d>}, \delta_{l_d}'^{<d>}\right) = \left(0,0,...,0,\delta_{l_d}^{<d>}, \delta_{l_d}'^{<d>}\right)$の分布は

$$\left(1-h_1^{<d>}\right) p\left(\delta_1'^{<d>} \middle| \boldsymbol{v}_1'^{<d>}\right) \times \left(1-h_2^{<d>}\right) p\left(\delta_2'^{<d>} \middle| \boldsymbol{v}_2'^{<d>}\right) \times \cdots \times \left(1-h_{l_d-1}^{<d>}\right) p\left(\delta_{l_d-1}'^{<d>} \middle| \boldsymbol{v}_{l_d-1}'^{<d>}\right) \times \left(h_{l_d}^{<d>}\right)^{\delta_{l_d}^{<d>}} \left(1-h_{l_d}^{<d>}\right)^{1-\delta_{l_d}^{<d>}} p\left(\delta_{l_d}'^{<d>} \middle| \boldsymbol{v}_{l_d}'^{<d>}\right)$$
$$= \left\{\prod_{l=1}^{l_d-1}\left(1-h_l^{<d>}\right)\right\} \times \left(h_{l_d}^{<d>}\right)^{\delta_{l_d}^{<d>}} \left(1-h_{l_d}^{<d>}\right)^{1-\delta_{l_d}^{<d>}} \times \prod_{l=1}^{l_d} p\left(\delta_l'^{<d>} \middle| \boldsymbol{v}_l'^{<d>}\right)$$

となる。すべての患者に関する対数尤度は

$$\ln L = \ln L(\boldsymbol{\beta}) + \sum_{d=1}^{n}\sum_{l=1}^{l_d} \ln p\left(\delta_l'^{<d>} \middle| \boldsymbol{v}_l'^{<d>}\right) \tag{6}$$

と表せる。ただし、

$$\ln L(\boldsymbol{\beta}) = \sum_{d=1}^{n}\left\{\sum_{l=1}^{l_d-1}\ln\left(1-h_l^{<d>}\right) + \delta_{l_d}^{<d>}\ln h_{l_d}^{<d>} + \left(1-\delta_{l_d}^{<d>}\right)\ln\left(1-h_{l_d}^{<d>}\right)\right\} \tag{7}$$

とする。よって、独立なベルヌーイ分布の(部分)対数尤度(7)式を最大化すれば、(5)式の回帰係数$\boldsymbol{\beta}$が推定される。

部分ロジスティックモデルの包括的な適合度を評価する手順はないが、ブートストラップ法を援用し、逸脱度

$$Dev = 2\left(\ln L_f - \ln L_c\right) \tag{8}$$

の棄却点を算出することができる。$\ln L_c$は、現行(current)の部分ロジスティックモデルのもとでの対数尤度(7)式、$\ln L_f$は飽和(full)モデルのもとでのそれで

$$\ln L_f = \sum_{d=1}^{n}\sum_{l=1}^{l_d}\left\{\delta_l^{<d>}\ln\delta_l^{<d>} + \left(1-\delta_l^{<d>}\right)\ln\left(1-\delta_l^{<d>}\right)\right\} = 0$$

となる。しかし、グループ化されていない二値データの場合、(8)式の漸近カイ二乗性は成立たない(Collett, 2003, 3.8.3節; Landwehr et al., 1984)。更に、患者1人が複数個の観測値(すなわち、患者#dはl_d個)を生成しており、自由度が計算できない。そこで、ブートストラップ法に基づいて、逸脱度の棄却点を算出する:

Step 1 初期標本$\boldsymbol{X} = \left(\boldsymbol{X}^{<1>}, \boldsymbol{X}^{<2>}, ..., \boldsymbol{X}^{<d>}, ..., \boldsymbol{X}^{<n>}\right)$からリサンプリング(ペア・サンプリング)によりブートストラップ標本$\boldsymbol{X}^* = \left(\boldsymbol{X}^{<1>*}, \boldsymbol{X}^{<2>*}, ..., \boldsymbol{X}^{<n>*}\right)$を生成する。ただし、$\boldsymbol{X}$の成分$\boldsymbol{X}^{<d>}$は$\boldsymbol{X}^{<d>} = \left(\boldsymbol{X}_l^{<d>}, \delta_1^{<d>}, \delta_2^{<d>}, ..., \delta_{l_d}^{<d>}\right)$である。

Step 2 $b(=1,\cdots,B)$番目のブートストラップ標本を$\boldsymbol{X}_b^*$とし、逸脱度

$$Dev(b) = 2\left\{\ln L_f - \ln L\left(\boldsymbol{X}_b^*; \hat{\boldsymbol{\beta}}\left(\boldsymbol{X}_b^*\right)\right)\right\} = -2\ln L\left(\boldsymbol{X}_b^*; \hat{\boldsymbol{\beta}}\left(\boldsymbol{X}_b^*\right)\right) \tag{9}$$

を計算する。ここに、$\hat{\boldsymbol{\beta}}\left(\boldsymbol{X}_b^*\right)$は$b$番目のブートストラップ標本$\boldsymbol{X}_b^*$から推定される回帰係数で、その推定値から算出される$\boldsymbol{X}_b^*$の対数尤度が$\ln L\left(\boldsymbol{X}_b^*; \hat{\boldsymbol{\beta}}\left(\boldsymbol{X}_b^*\right)\right)$である。

Step 3 $Dev \geq Dev^*$なら、有意水準αでモデルは妥当でないとみなす。ここに、Devは初期標本に対

する(8)式の逸脱度で、

$Dev^{*} = Dev(b)$を小さい順に並べたときの第j番目の値、$\alpha = 1 - j/(B+1)$

とする。

次に、患者#d の来院時間(観測時点)$t_1^{<d>},...,t_l^{<d>},...,t_{l_d}^{<d>}$ について、(5)式のハザード関数$h_l^{<d>}$を用いると生存関数は

$$S\left(\boldsymbol{x}_l^{<d>}, a_l^{<d>}\right) = \prod_{1 \le k \le l} \left(1 - h_k^{<d>}\right) \tag{10}$$

となる。よって、$t_l^{<d>}$ まで生存した患者が、次の短期間Δt 後も生存する条件付き生存率は、(10)式から

$$\Pr\left(t_l^{<d>}, t_l^{<d>} + \Delta t\right) = \frac{S\left(\boldsymbol{x}_l^{<d>}, a_l^{<d>} + \Delta t\right)}{S\left(\boldsymbol{x}_l^{<d>}, a_l^{<d>}\right)}$$

と予測できる。ちなみに、共変量の係数β_i の有意性を検定するには、尤度比検定統計量

$$-2\ln\lambda = -2\left\{\ln L\left(\boldsymbol{X}_{[i]}; \hat{\boldsymbol{\beta}}'\right) - \ln L\left(\boldsymbol{X}; \hat{\boldsymbol{\beta}}\right)\right\}, \tag{11}$$

を計算すればよい。ここに、$\ln L\left(\boldsymbol{X}; \hat{\boldsymbol{\beta}}\right)$ は初期標本 $\boldsymbol{X}$ に対する対数尤度であり、$\ln L\left(\boldsymbol{X}_{[i]}; \hat{\boldsymbol{\beta}}'\right)$ は i 番目の共変量を除去したときのそれである。帰無仮説が真なら、(11)式の$-2\ln\lambda$ は漸近的に自由度 $d.f.$=1 のカイ二乗分布に従う。

2.3 ニューラルネット

データ解析の目的が予測にあるなら、母集団構造の非線形性を摘出するだけでは不十分で、適切な非線形モデルで記述しなければならない。本節では、階層型ニューラルネットワークモデル(図1)を想定し、ブートストラップ法(Tsujitani and Koshimizu, 2000)に基づく隠れユニット数の決定およびモデルの妥当性の検証を系統的に行う。隠れユニット数の決定にクロス・バリデーション法(Shibata, 1997)を適用することもできるが、そのための再計算に要する時間は膨大になり、かつその方法では、モデルの妥当性の検証ができない。

Biganzoli et al.(1998)は、時間依存型でない共変量をもつ生存時間を“月”あるいは“週”単位で離散化し、ニューラルネットによる生存解析を試みた(平野ら, 1998; 本田ら, 2000)。その離散化の仕方により結果が異なることもある。しかし、表1のように共変量が時間と共に変動する場合、離散化の仕方は一意に決まり、観測期間の任意時点におけるΔt 後の条件付き生存率の予後予測が威力を発揮する。更に、ニューラルネットは部分ロジスティックモデルの拡張とみなせるので、各個体の尤度への寄与が、時間区間で独立であるという(6)式は、ニューラルネットにおいても成立する。

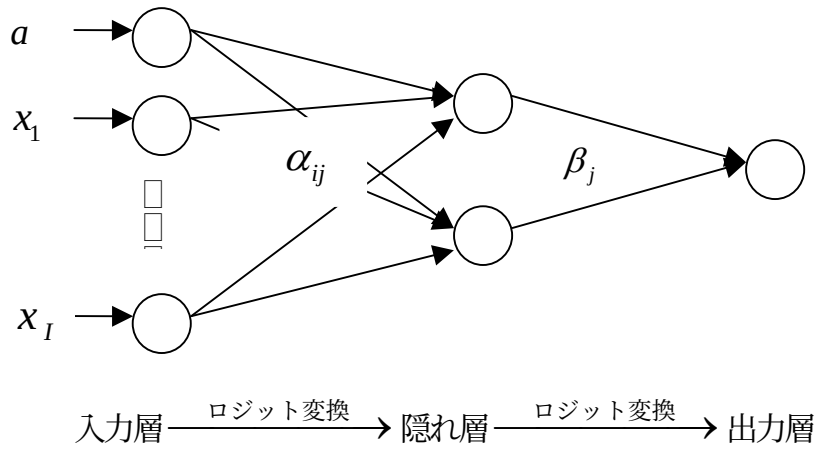


図1. 階層型ニューラルネット

(1) 定式化

階層型ニューラルネットによって生存データを取扱う場合、ある入力パターンを提示したときの出力値は、ベイズの事後確率と考えられる。いま、患者#$d\,(=1,2,...,n)$の入力$\boldsymbol{X}_l^{<d>}$が与えられたとき、隠れ層の第$j(=1,2,...,J)$ユニットの活性値

$$u_{lj}^{<d>} = \alpha a_l^{<d>} + \sum_{i=0}^{I} \alpha_{ij} x_{li}^{<d>}, x_{0j} \equiv 1; l = 1,...,l_d; d = 1,...,n; j = 1,...,J \tag{12}$$

が定まる。(12)式の活性値$u_{lj}^{<d>}$に**ロジット(シグモイド)**変換

$$y_{lj}^{<d>} = \frac{1}{1+\exp\left(-\varepsilon u_{lj}^{<d>}\right)} \tag{13}$$

を施すと、隠れ層の第jユニットの出力値y_{lj}が決まる。ここに、ロジット関数の傾きεは、非線形性を有効に作用させるための係数である。εが小さければ線形関数に近い形状となり、逆に大きくなれば非線形性が強くなる。傾きεの決定方法については、安居ら(1993,4.4節)を参照されたし。

ロジット変換された(13)式の$y_{lj}^{<d>}$が、出力層への活性値

$$z_l^{<d>} = \sum_{j=0}^{J} \beta_j y_{lj}^{<d>}, y_{l0}^{<d>} \equiv 1$$

となる。この$z_l^{<d>}$にロジット変換を施すと、最終出力(ハザード関数)

$$h_l^{<d>} = \frac{1}{1+\exp(-\varepsilon z_l^{<d>})} \tag{14}$$

が得られる。2.2節と同様にして、$\delta_l^{<d>}, \delta_l'^{<d>}, \boldsymbol{v}_l^{<d>}, \boldsymbol{v}_l'^{<d>}$を定義する。(5)式の代わりに(14)式を用いれば、対数尤度が得られる。この対数尤度を最大にする$\hat{h}_l^{<d>}$ (すなわち、リンク荷重$\hat{\boldsymbol{\theta}} = \left\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right\}$, $\hat{\boldsymbol{\alpha}} = \{\hat{\alpha}, \hat{\alpha}_{ij}\}, \hat{\boldsymbol{\beta}} = \{\hat{\beta}_j\}$が最尤推定量となる。

(2) 隠れユニット数の決定

ニューラルネットは、母集団モデルの近似であるため、真のモデルと想定したそれとは分離している(model misspecified)と考えるべきである(Anders and Korn,1999)。この点から、想定したモデル族に真のモデルが含まれていることを前提にした***AIC***(Akaike, 1973)によるモデル選択は妥当ではない。競合するモデルが複数個あるとき、対数尤度の比較によってモデルを選択すると、自由なパラメータ数の大きいモデルほど選ばれやすい。***EIC***(Ishiguro et al., 1997; Konishi and Kitagawa, 1991)は、対数尤度のバイアスをブートストラップ法を用いて直接推定している。ブートストラップ法に基づくバイアス推定は次の通りである：

Step 1 初期標本$\boldsymbol{X}$からリサンプリングによりブートストラップ標本$\boldsymbol{X}^*$を生成する。

Step 2 $b(=1,\cdots,B)$番目のブートストラップ標本$\boldsymbol{X}_b^*$について、ブートストラップの推定値に基づく初期標本の対数尤度$\ln L\left(\boldsymbol{X}; \hat{\boldsymbol{\theta}}\left(\boldsymbol{X}_b^*\right)\right)$、およびブートストラップの推定値に基づくブートストラップ標本の対数尤度$\ln L\left(\boldsymbol{X}_b^*; \hat{\boldsymbol{\theta}}\left(\boldsymbol{X}_b^*\right)\right)$を計算する。

Step 3 ***Step 1***、***2***をB回繰り返す。

Step 4 バイアスのブートストラップ推定は

$$C^* \cong \frac{1}{B}\sum_{b=1}^{B}\left\{\ln L\left(\boldsymbol{X}_b^*; \hat{\boldsymbol{\theta}}\left(\boldsymbol{X}_b^*\right)\right) - \ln L\left(\boldsymbol{X}; \hat{\boldsymbol{\theta}}\left(\boldsymbol{X}_b^*\right)\right)\right\}$$

となる。このとき、ブートストラップ法に基づく情報量規準

$$EIC=-2\ln L\left(\boldsymbol{X};\hat{\boldsymbol{\theta}}\left(\boldsymbol{X}\right)\right)+2C^{*} \tag{15}$$

が得られ、*EIC* が最小のモデルを最適として選択することができる。

(3) 部分ロジスティックモデルとニューラルネットとの関係

部分ロジスティックモデルとニューラルネットとの関係について考察しよう。図1のニューラルネットにおいて、隠れユニットが 1 個の場合を考える。入力 $\boldsymbol{X}_l^{<d>}$ が与えられたとき、隠れユニットの活性値

$$u_l^{<d>}=\alpha a_l^{<d>}+\sum_{i=0}^{I}\alpha_i x_{li}^{<d>},\ x_{0j}\equiv 1; l=1,...,l_d; d=1,...,n$$

に**恒等変換**を施すと、隠れ層の出力値が決まる。これが、出力層への活性値

$$z_l^{<d>}=\beta_0+\beta_1 u_l^{<d>}$$

となる。$z_l^{<d>}$ にロジット変換を施すと、最終出力(ハザード関数)

$$h_l^{<d>}=\frac{1}{1+\exp\left\{-\sum_{j=0}^{1}\beta_j\left(\alpha a_l^{<d>}+\sum_{i=0}^{I}\alpha_{ij}x_{li}^{<d>}\right)\right\}}=\frac{1}{1+\exp\left\{\xi a_l^{<d>}+\sum_{i=0}^{I}\xi_i x_{li}^{<d>}\right\}}$$

が得られる。これは、グループ化されていないデータに対する部分ロジスティックモデルと等価になる。

3. 解析結果

3.1 Mayo updated モデルとヨーロッパ new version モデル

Mayo updated モデルでは、時間区間 l における生存時間は、共変量として $x_{l1}^{<d>}, x_{l2}^{<d>}, ..., x_{l5}^{<d>}$ をもつ。よって、$n=312$ 例について、合計 1945 個の観測ベクトルが得られ、これらを従来の時間固定(time-fixed)型の比例ハザードモデルで解析する。回帰係数の推定値および標準誤差の推定値を表 2 に与えておく。なお、Mayo updated モデルおよびヨーロッパ new version モデルでは、プロトロンビン時間、ビリルビン値、アルブミン値について、対数変換値が用いられる。

表2. 回帰係数、および標準誤差の推定値

共変量	Mayo updated ﾓﾃﾞﾙ		ﾖｰﾛｯﾊﾟ new version ﾓﾃﾞﾙ		部分ﾛｼﾞｽﾃｨｯｸﾓﾃﾞﾙ	
	回帰係数	標準誤差	回帰係数	標準誤差	回帰係数	標準誤差
年齢(歳)	0.044	0.009	0.043	0.009	0.077	0.018
ﾌﾟﾛﾄﾛﾝﾋﾞﾝ時間	2.682	0.580	2.817	0.627	0.199	0.052
ﾋﾞﾘﾙﾋﾞﾝ値	1.184	0.114	1.068	0.110	0.139	0.015
ｱﾙﾌﾞﾐﾝ値	-3.496	0.459	-3.666	0.492	-1.688	0.227
ｴﾃﾞﾏ･ｽｺｱ	0.669	0.226	0.801	0.233	0.931	0.288

ベースライン生存関数 $S_0(\Delta t)$ は、表3のように推定される。$S_0(\Delta t)$ を用い、観測時点 t まで生存していた患者の、6 ヶ月後の条件付き生存率を予測する。例えば、来院回数 1 の *PI* は $0.044\times(42.5-52.43)+2.682\times(\ln 11.0-2.39)+1.184\times(\ln 3.2-0.6)-3.496\times(\ln 3.08-1.21)+0.669\times(0-0.182)=0.4472$ となる。よって、来院回数 1 のとき、6 ヶ月後の条件付き生存率は、表 3 と(2)式から $0.992^{\exp(0.4472)}=0.9875$ と予測される。

表 3.　ベースライン生存関数の推定値

Δt(月)	0	3	6	9	12
$S_0(\Delta t)$	1	0.996	0.992	0.991	0.990

次に、ヨーロッパ new version モデルでは、ハザード関数に基づく予後予測の近似が試みられている。(3)式の回帰係数および標準誤差の推定値を表 2 に示しておく。ヨーロッパ new version モデルにおけるベースライン累積ハザード関数 $\Lambda_0(t)$ は、図 2 のように推定される。

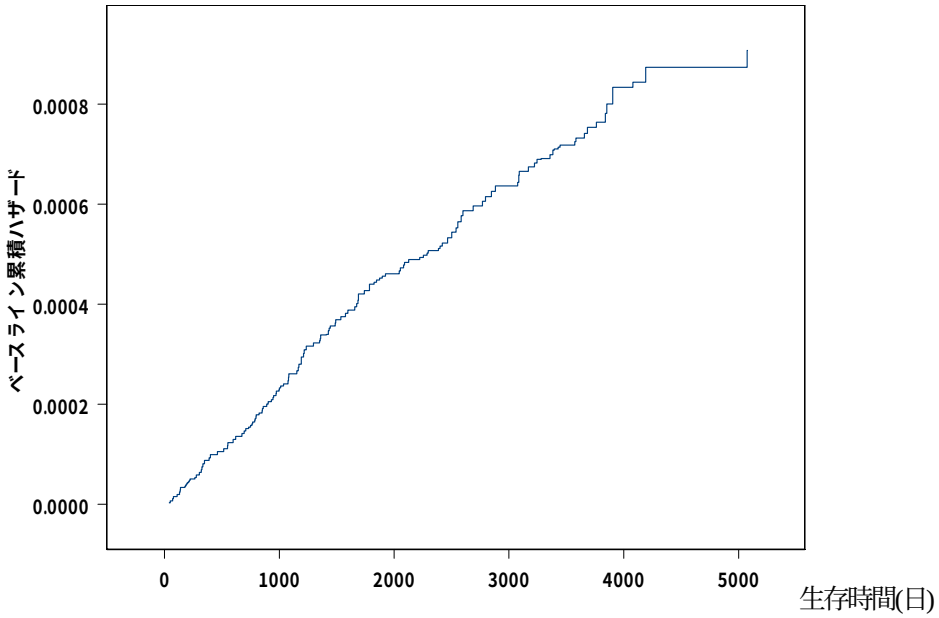


図 2.　ベースライン累積ハザード関数

図 2 に直線を当てはめ、目算により(4)式の傾きを推定すると、$\hat{\lambda}^* = 2.28\times10^{-7}$ となる。よって、例えば来院回数 1 のときの 6 ヶ月後の生存率は

$$\sum_{i=1}^{5}\hat{\beta}_i x_i(t) = 0.04343\times42.5 + 2.8174\times\ln 11.0 + 1.0684\times\ln 3.2 - 3.666\times\ln 3.08 + 0.8010\times 0 = 5.744$$

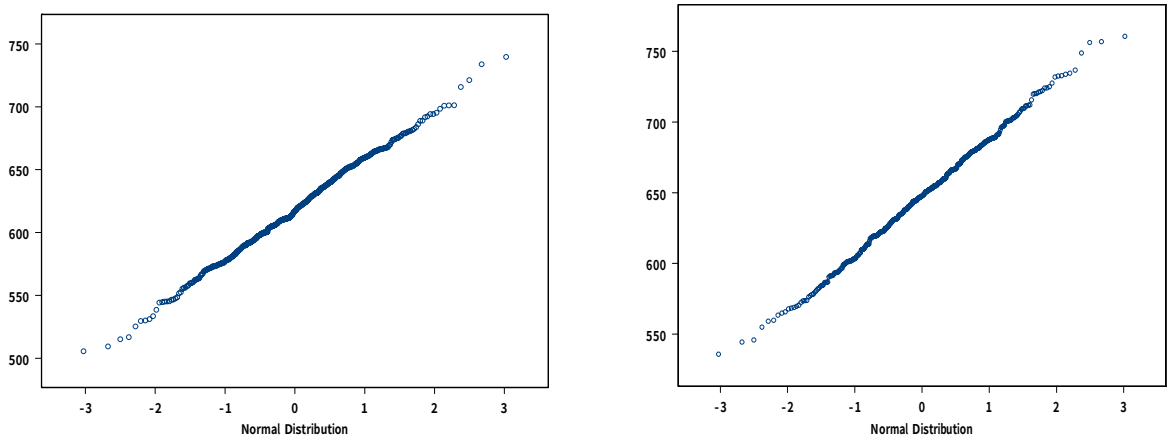
より

$$\Pr(t, t+\Delta t) = \exp\left\{-2.3\times10^{-7}\times\frac{365}{2}\times\exp(5.744)\right\} = 0.9871$$

と予測される。このデータの場合、図3に線形性が見られるので、$\lambda_0(t) = \lambda^*$ の仮定が成立つ。しかし、この方式では、ベースラインハザード関数が局所的に一定という仮定が必要である。また、その一定値をグラフから目算するため、解析者による恣意性が生じてくる。

3.2 部分ロジスティックモデル

部分ロジスティックモデルで解析しよう。$B = 400$ としたとき、ブートストラップ法による(9)式の Q-Q プロットが図 3(a)である。$Dev(b)$ の分布はかなり正規分布に近いことが分かる。PBC データの 95%点 Dev^* の推定値は、$Dev^* = 680.71$ となる。(8)式の逸脱度は $Dev = 614.14$ より、部分ロジスティックモデルの妥当性が示唆される。次に共変量の係数の有意性を検定する。部分ロジスティックモデル(5)の回帰係数 $\boldsymbol{\beta}$、およびその標準誤差の推定値も表 2 に併記しておく。



(a) 部分ロジスティックモデル　　(b) ニューラルネット

図 3. $Dev(b)$ の *Q-Q* プロット

参考までに、時間依存型比例ハザードモデル(Kalbfleisch and Prentice, 2002, 6.4 節, Lawless, 2003, 7.1.8 節)に基づく Wald 検定のカイ二乗値、および初診時(来院回数 1)の値のみを用いた場合(Collett, 1994)の共変量の係数の有意性検定(Wald 検定)を表 4 示す。部分ロジスティックモデルおよび時間依存型比例ハザードモデルでは、すべての共変量が高度に有意であることが分かる。共変量の初診時の値のみを用いた場合では、エデマ・スコアのみ 5%有意で、他は高度に有意になる。

表 4.　共変量の係数の有意性検定

共変量	部分ﾛｼﾞｽﾃｨｯｸﾓﾃﾞﾙ (Wald ｶｲ二乗値)	時間依存型 比例ﾊｻﾞｰﾄﾞﾓﾃﾞﾙ	共変量の初診時の値のみを用いた場合
年齢	26.22(25.15)	23.82	23.07
ﾌﾟﾛﾄﾛﾝﾋﾞﾝ時間	13.32(14.65)	20.16	17.02
ﾋﾞﾘﾙﾋﾞﾝ値	86.62(84.03)	94.21	73.44
ｱﾙﾌﾞﾐﾝ値	56.10(55.14)	55.53	10.99
ｴﾃﾞﾏ･ｽｺｱ	10.36(10.46)	11.86	4.61

更に、患者#9 について Mayo updated モデル、ヨーロッパ new version モデル、および部分ロジスティックモデルによる観測期間の任意時点における次の 6 ヶ月後の条件付き生存率を表 5 に与えておく。同表から、3 つのモデルとも生存率は漸近的に下降傾向を示している。ただし、来院回数 6 のみ、部分ロジスティックモデルに対する値がやや小さい。

表 5.　患者#9(死亡例)する 6 ヶ月後の生存率

ﾓﾃﾞﾙ	来		院		回		数
	1	2	3	4	5	6	7
部分ﾛｼﾞｽﾃｨｯｸ	0.9785	0.9812	0.9766	0.8183	0.9186	0.8221	0.4938
Mayo updated	0.9875	0.9749	0.9822	0.8288	0.9199	0.8453	0.4975
ﾖｰﾛｯﾊﾟ new version	0.9871	0.9765	0.9821	0.8418	0.9278	0.8562	0.4801

3.3 ニューラルネット

部分ロジスティックモデルとニューラルネトとの関係を数値的に示すため、2 章の PBC データのうち、共変量として生存時間の中央値、年齢, プロトロンビン時間、ビリルビン値の 4 個の場合も考える(Chang and Weissfeld, 1999)。この 2 種類のデータに対して、隠れユニットが 1～4 個のニューラルネットおよび部分ロジスティックモデルを当てはめたところ、表 6 ような *EIC* 値が得られた。同表から、*i*) 共変量が (a, x_1, x_2, x_3) と少なくなると、最適な隠れユニットは 2 個となり、隠れユニ

ット数を増やしたニューラルネットモデルの有効性が明白になってくる。*ii*) 共変量が $(a, x_1, x_2, x_3, x_4, x_5)$ の場合、隠れユニットが 1 個のニューラルネットと部分ロジスティックモデルに対する *EIC* 値には、ほとんど差がない。すなわち、隠れユニットが 1 個の場合、(13)式と恒等変換との違いのみである。

表 6.　*EIC* 値

共変量	隠れユニット数				部分ﾛｼﾞｽﾃｨｯｸ
	1	2	3	4	
(a, x_1, x_2, x_3)	723.02	708.28	720.16	732.16	719.34
$(a, x_1, x_2, x_3, x_4, x_5)$	666.04	681.00	686.68	687.20	664.82

*: a : 生存時間の中央値, x_1 : 年齢, x_2 : ﾌﾟﾛﾄﾛﾝﾋﾞﾝ時間, x_3 :ﾋﾞﾘﾙﾋﾞﾝ値, x_4 :ｱﾙﾌﾞﾐﾝ値, x_5 :ｴﾃﾞﾏ・ｽｺｱ

ニューラルネットを生存時間解析に適用したとき、未解決な課題としてモデルの妥当性の検証が残されている(Biganzoli et. al., 1998)。本稿では、2.2 節と同様にして、ブートストラップ法による逸脱度の棄却点を算出することができる。$B = 400$ としたとき、ブートストラップ標本に基づく逸脱度の *Q-Q* プロットは図 3(b)のようになる。逸脱度は $Dev = 655.44$ となり、5%棄却点 719.51 より小さいので、モデルの妥当性が示唆される。

更に、2.2 節と同様に予後予測を行うこともできる。共変量として (a, x_1, x_2, x_3) を用いた場合、患者#9(死亡例)に関する 6 ヶ月後の条件付き生存率は表 7 のようになる。参考までに、共変量 $(a, x_1, x_2, x_3, x_4, x_5)$ について、隠れユニット数 *J*=1,2 に対するニュラルネットによる解析結果も併記しておく。同表から、*i*) 共変量として (a, x_1, x_2, x_3) を用いた場合、ニューラルネットによる条件付き生存率は、他のモデルより低く予測されている。*ii*) 共変量として $(a, x_1, x_2, x_3, x_4, x_5)$ を用いた場合(最適な隠れユニット数は *J*=1 であるが)、*J*=2 に対する来院回数 7 の条件付き生存率が *J*=1 に比べて高く予測される(過学習)。

表 7.　患者#9(死亡例)に関する 6 ヶ月後の生存率

ﾓﾃﾞﾙ		来院回数 1	2	3	4	5	6	7
ﾆｭｰﾗﾙﾈｯﾄ (a, x_1, x_2, x_3)		0.9625	0.8970	0.9520	0.6598	0.8930	0.8494	0.6908
部分ﾛｼﾞｽﾃｨｯｸ		0.9830	0.9481	0.9778	0.7734	0.9078	0.8254	0.7627
Mayo updated		0.9870	0.9386	0.9796	0.7667	0.9026	0.8509	0.7634
ﾖｰﾛｯﾊﾟ new version		0.9882	0.9457	0.9986	0.7965	0.9157	0.8719	0.7956
ﾆｭｰﾗﾙﾈｯﾄ $(a, x_1, x_2, x_3, x_4, x_5)$	*J*=1	0.9781	0.9810	0.9769	0.8474	0.9352	0.8589	0.4771
	J=2	0.9737	0.9855	0.9785	0.8485	0.9497	0.8560	0.5386

なお、ニューラルネットのリンク荷重について、部分ロジスティックモデルのような回帰係数の解釈は困難であるが、共変量の係数の有意性は、2.2 節と同様に行うことができる。参考までにその結果を表 8 に示す。すべての共変量が高度に有意である。

表 8.　共変量の係数の有意性検定

共変量	Wald ｶｲ二乗値
年齢	51.18
ﾌﾟﾛﾄﾛﾝﾋﾞﾝ時間	31.04
ﾋﾞﾘﾙﾋﾞﾝ値	138.96

最後に、共変量(a, x_1, x_2, x_3)を用いたとき、全312例中の死亡例(140例)、打切り例(172例)の2群に対して、4手法に基づくΔt後の条件付き生存率を比較する。死亡例群$(g=1)$および打切り例群$(g=2)$について、d番目の患者に対するl番目の来院回数(時間区間)におけるΔt後の条件付き生存率を$\Pr_d^{[g]}(l, l+\Delta t)$とする。直接平均法(Markus et al., 1989; Thomsen et al., 1991)を採用し、群gにおけるl番目の来院回数に対するΔt後の条件付き生存率の平均値

$$S_g(l) = \frac{1}{n_l^{[g]}} \sum_{d=1}^{n_l^{[g]}} \Pr_d^{[g]}(l, l+\Delta t), \quad g = 1, 2; l = 1, 2, \ldots$$

を算出する。ここに$n_l^{[g]}$は、群gにおける来院回数lでの総患者数である。 図4は、全312例中の死亡例、打切り例について、4手法による6ヶ月後の条件付き生存率の比較である。同図から、*i*)死亡例についてはニューラルネットによる条件付き生存率が低く、従来の方法に比べて、より良く予測しており(死亡例の予測が、臨床的には移植の適応を考える上で重要)、*ii*)打切り例に関しては、4手法の差は少ないことが明らかになる。なお表6から、共変量が(a, x_1, x_2, x_3)の場合は、隠れユニット2個のニューラルネットのほうが、部分ロジスティックモデルより*EIC*値が小さいが、参考までに部分ロジスティックモデルの条件付き生存率も図示しておく。

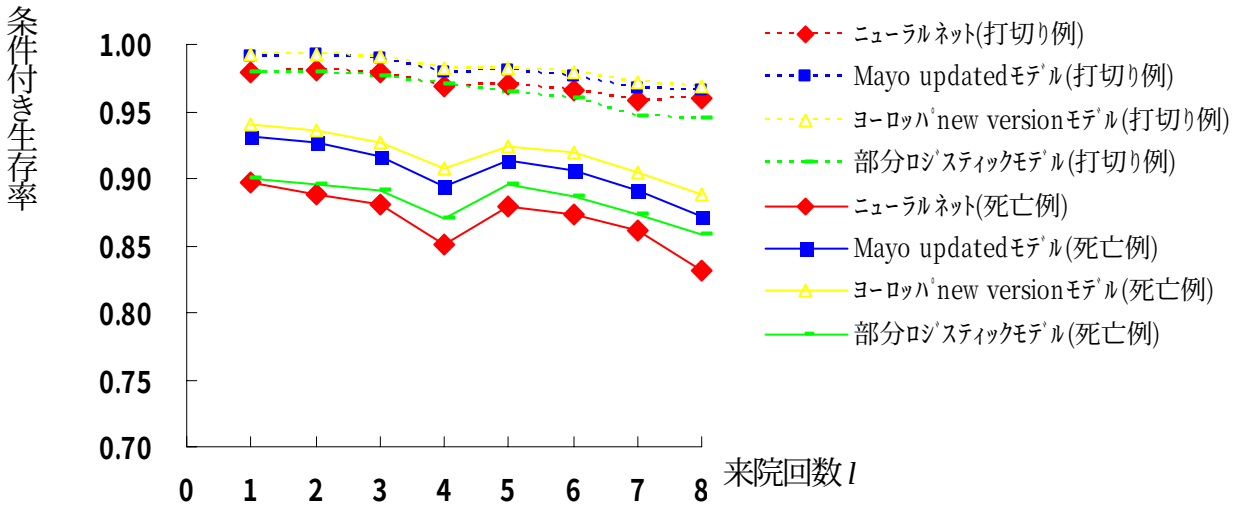


図4.4手法による全312例中の死亡例、打切り例に対する条件付き生存率の比較

4. 結び

共変量の値が時間とともに変化する時間依存型データが含まれる場合、従来のCoxの比例ハザードモデルには、ベースライン生存関数やベースライン累積ハザード関数の推定などに問題点が残されている。この点を踏まえ本稿では、部分ロジスティックモデルおよびニューラルネットモデルを提案し、ブートストラップ法による統計的推測を行った。PBCデータの解析を通じ、従来、広範に利用されてきたMayo updatedモデルやヨーロッパ new versionモデルとの数値的比較も行った。

本手順は、*i*) 時間依存型共変量の取扱いが容易である，*ii*) Mayo updatedモデルでは、観測期間全体を考慮に入れるため、一人の患者が複数個の観測値を生成しているにもかかわらず、互いに独立として従来の時間固定(time-fixed)型の比例ハザードモデルで解析している。そのため、提案法では部分尤度を援用することにより、この独立性を回避している，*iii*) 共変量の非線形性を包含(生存時間値そのものを利用)している、*iv*) モデルの妥当性を検証できる、*v*) ヨーロッパ new version モデルのようなベースラインハザード関数が一定の仮定が不要である，などの利点を有する。欠点としては、ニューラルネットモデルの場合、*i*) 推定値の算出に膨大な時間が必要，*ii*) 最尤解の一意性が

いえない、*iii*) ニューラルネットのリンク荷重について、部分ロジスティックモデルのような回帰係数の解釈は困難である、などが挙げられる。なお、本稿で採用した時間依存型比例ハザードモデルのための統計ソフトは、*http://www.osakac.ac.jp/labs/tujitani/PBC.pdf* に公開している。

謝辞 査読者の方々、編集理事には、大変貴重なご指摘をいただきました。ここに記して厚く御礼申し上げます。

参考文献

安居院猛、長橋宏、高橋裕樹(1993):ニューラルプログラム、昭晃堂.

Akaike, H. (1973): Information theory and an extension of the maximum likelihood principle. *2nd Inter. Symp. on Information Theory* (ed. Petrov, B.N. and Caski, F.), Akademiai Kiado, Budapest. 267-281.

Altman, D.G. and De Stavola, B.L. (1994): Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates. *Stat. Med.* **13**, 301-341.

Anders, V. and Korn, O. (1999): Model selection in neural networks. *Neural Networks* **12**, 309-323.

Arjas, E. (1988): A graphical method for assessing goodness of fit in Cox's proportional hazards model. *J. Amer. Statist. Assoc.* **83**, 204-212.

Biganzoli, E., Boracchi, P., Mariani, L. and Marubini, E. (1998): Feed forward neural networks for the analysis of censored survival data: A partial logistic approach. *Stat. Med.* **17**, 1169-1186.

Biganzoli, E., Boracchi, P., and Marubini, E. (2002): A general framework for neural network models on censored survival data. *IEEE Trans. Neural Networks* **15**, 209-218.

Chang, C.-C. H. and Weissfeld, L.A. (1999): Normal approximation diagnostics for Cox model. *Biometrics* **55**, 1114-1119.

Christensen, E., Schlichting, P., Andersen, P.K., Fauerholdt, L., Schou, G., Pedersen, B.V., Juhl, E., Poulsen, H., Tygstrup, N., Copenhagen Study Group for Liver Disease (1986): Updating prognosis and therapeutic effect evaluation in cirrhosis with Cox's multiple regression model for time-dependent variables. *Scand. J. Gastroenterology* **21**, 163-174.

Christensen, E., Altman, D.G., Neuberger, J., De Stavola, B.L., Tygstrup, N., Williams, R., The PBC1 and PBC2 Trial groups (1993): Updating prognosis in primary biliary cirrhosis using a time-dependent Cox regression model. *Gastroenterogy* **105**, 1865-1876.

Collett,D.(1994): *Modelling Survival Data in Medical Research.* Chapman and Hall, London.

Collett,D.(2003): *Modelling Binary Data*, 2nd ed., Chapman and Hall, London.

Cox, D.R. (1975): Partial likelihood. *Biometrika* **62**, 269-276.

Efron, B.(1988): Logistic regression, survival analysis, and Kaplan-Meier curve. *J. Amer. Stat. Assoc.* **83**, 414-425.

平野勝也, 惣田隆生, 田崎武信(1998):生存時間解析におけるニューラルネットワークの魅力と限界. 癌臨床研究・生物統計研誌 **19**, 53-61.

本田圭一, 寺西孝司, 田崎武信(2000):MARS による生存時間解析. 癌臨床研究・生物統計研誌 **20**, 34-46.

市田文弘, 谷川久一(1991): 肝移植適応基準, 国際医書出版.

井上恭一(1994):原発性胆汁性肝硬変：病態・治療・予後, へるす出版.

Ishiguro, M., Sakamoto, Y. and Kitagawa, G. (1997): Bootstrapping log likelihood and EIC, An extension of AIC. *Ann. Inst. Stat. Math.* **49**, 411-434.

Kalbfleisch, J.D. and Prentice, R.L.(2002): *The Statistical Analysis of Failure Time Data*, 2nd ed., John Wiley, New York.

Klein, J.P. and Moeschberger, M.L. (2003): *Survival Analysis*, 2nd ed., Springer, New York.

Konishi, S. and Kitagawa, G.(1996): Generalized information criteria in model selection. *Biometika* **83**, 875-890.

Landwehr, J.M, Pregibon,D. and Shoemaker,A.C.(1984): Graphical methods for assessing logistic regression models. *J. Amer. Stat. Assoc.* **79**, 61-71.

Lawless, J.F.(2003): *Statistical Models and Methods for Lifetime Data*, 2nd ed., John Wiley, New York.

Markus, B.H., Dickson, E.R. Grambsch, P.M., Fleming, T.R., Mazzaferro, V., Klintmalm,G.B., Wiener, R.H., Van Thiel, D.H. and Starzl, T.E. (1989): Efficacy of liver transplantation in patient with primary biliary cirrhosis. *New Engl. J. Med*. **320**, 1709-1713.

Marubini, E. and Valsecchi, M.G.(1995):*Analysing Survival Data from Clinical Trials and Observational Studies*, John Wiley, New York.

Murtaugh, P.A., Dickson, E.R., Van Dam, G.M., Malinchoc, M, Grambsch, P.M., Langworthy, A.L., Gips, C.H.(1994): Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits. *Hapatology* **20**, 126-134.

中村剛(2001):Cox 比例ハザードモデル，朝倉書店.

大橋靖雄，浜田知久馬(1995):生存時間解析，東大出版.

Preston,D.L., Lubin, J.H. and Piece, D.A.(1991): *EPICURE User's Guide*, Hirosoft International Corporation, Seattle.

Shibata, R. (1997): Bootstrap estimate of Kullback-Leibler information for model selection. *Statist. Sinica* **7**, 375-394.

Therneau, T. M. and Grambsch P.M. (2000): *Modeling Survival Data: Extending the Cox Model*. Springer: New York.

Thomsen, B.L., Keiding, N., and Altman, D.G.(1991):A note on the calculation of expected survival. *Statist. Med.* **10**, 733-738.

Tsujitani, M. and Koshimizu,T.(2000)：Neural Discriminant Analysis. *IEEE Trans. Neural Networks* **11**, 1394-1401.

Markus, B.H., Dickson, E.R. Grambsch, P.M., Fleming, T.R., Mazzaferro, V., Klintmalm,G.B., Wiener, R.H., Van Thiel, D.H. and Starzl, T.E. (1989): Efficacy of liver transplantation in patient with primary biliary cirrhosis. *New Engl. J. Med*. **320**, 1709-1713.

Analysis of survival data having time-dependent covariates

Masaaki Tsujitani[1,*] **and Masato Sakon**[2]

[1]Faculty of Information Science & Arts, Osaka Electro-Communication University, Osaka, JAPAN
[2] Department of Surgery, Nishinomiya Municipal Central Hospital, Hyogo, JAPAN

Abstract

Cox's proportional hazards mode has been widely used for the analysis of treatment and prognostic effects with censored survival data. The model was developed based on the relation between survival and the patient characteristic observed when the patient entered the study. When the covariates values change for the duration of the study, however, some theoretical problems to be solved with respect to baseline survival function and baseline cumulative hazard function are involved. Several prognostic models (e.g., Mayo updated model and European new version model) have become widespread using the Cox's proportional hazards modes for the analysis of survival data having time-dependent covariates. In the present study, we propose a partial logistic model and neural network model based on bootstrapping to estimate survival function and predict the survival for the following short-term (say, 6 months) at any time during the course of the disease . The proposed method is illustrated using data from a long-term study of patients with PBC(primary biliary cirrhosis) to aid the decision when to undertake liver transplantation.

Key words: Cox proportional hazards regression model, time-dependent covariates, partial logistic regression models, bootstrapping, neural network model.

[*] Corresponding author.
E-mail address:ekaaf900@ricv.zaq.ne.jp

追加資料:統計ｿﾌﾄ S+の紹介(時間依存型比例ﾊｻﾞｰﾄﾞﾓﾃﾞﾙ)

時間依存型比例ハザードモデルに関する統計ソフト ***S+***を紹介する。例として、肝硬変データ(Collett,1994)を取上げる。同表は、12 例の患者データである。ここに

TIME：生存時間

$$\text{CENS} = \begin{cases} 0: \text{打切り} \\ 1: o.w. \end{cases}$$

LBR：初診時のビリルビン値(対数値)

とする。

表 A.1.　12 例の患者に対する生存時間

患者番号	TIME	CENS	LBR
1	281	1	3.2
2	604	0	3.1
3	457	1	2.2
4	384	1	3.9
5	341	0	2.8
6	842	1	2.4
7	1514	1	2.4
8	182	0	2.4
9	1121	1	2.5
10	1411	0	2.3
11	814	1	3.8
12	1071	1	3.1

表 A.1 において、LBR は時間依存型であり、観測時点ごとに変化する。その値を表 A.2 に示す。

表 A.2.　観測時点ごとの LBR 値

患者番号	観測時点	LBR	
1	47	3.8	
	184	4.9	
	251	5.0	※
2	94	2.9	
	187	3.1	※
	321	3.2	
3	61	2.8	
	97	2.9	
	142	3.2	※
	359	3.4	
	440	3.8	
4	92	4.7	
	194	4.9	※
	372	5.4	
5	87	2.6	
	192	2.9	※
	341	3.4	
6	94	2.3	
	197	2.8	※
	384	3.5	
	795	3.9	
7	74	2.9	
	202	3.0	※
	346	3.0	
	917	3.9	
	1411	5.1	
8	90	2.5	
	182	2.9	
9	101	2.5	※
	410	2.7	
	774	2.8	
	1043	3.4	
10	182	2.2	※
	847	2.8	
	1051	3.3	
	1347	4.9	
11	167	3.9	※
	498	4.3	
12	108	2.8	
	187	3.4	※
	362	3.9	
	694	3.8	

※印：表 A.3 のベースライン累積ハザード関数の算出法で、患者番号(1)のベースラインハザード関数で使用する LBR 値

入力データ

event	LBRT	start	stop	患者番号
0	3.2	0	47	1
0	3.8	47	184	1
0	4.9	184	251	1
1	5	251	281	1
0	3.1	0	94	2
0	2.9	94	187	2
0	3.1	187	321	2
0	3.2	321	604	2
0	2.2	0	61	3
0	2.8	61	97	3
0	2.9	97	142	3
0	3.2	142	359	3
0	3.4	359	440	3
1	3.8	440	457	3
0	3.9	0	92	4
0	4.7	92	194	4
0	4.9	194	372	4
1	5.4	372	384	4
0	2.8	0	87	5
0	2.6	87	192	5
0	2.9	192	341	5
0	3.4	341	341.1	5
0	2.4	0	94	6
0	2.3	94	197	6
0	2.8	197	384	6
0	3.5	384	795	6
1	3.9	795	842	6
0	2.4	0	74	7
0	2.9	74	202	7
0	3	202	346	7
0	3	346	917	7
0	3.9	917	1411	7
1	5.1	1411	1514	7
0	2.4	0	90	8
0	2.5	90	182	8
0	2.9	182	182.1	8
0	2.5	0	101	9
0	2.5	101	410	9
0	2.7	410	774	9
0	2.8	774	1043	9
1	3.4	1043	1121	9
0	2.3	0	182	10
0	2.2	182	847	10
0	2.8	847	1051	10
0	3.3	1051	1347	10
0	4.9	1347	1411	10

0	3.8	0	167	11
0	3.9	167	498	11
1	4.3	498	814	11
0	3.1	0	108	12
0	2.8	108	187	12
0	3.4	187	362	12
0	3.9	362	694	12
1	3.8	694	1071	12

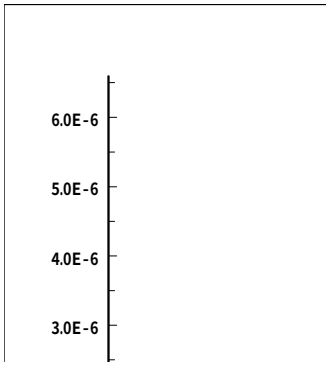
プログラム

```
> attach(COLLETT)
> coxph(Surv(start,stop,event)~LBRT)
```

解析結果

```
        coef  exp(coef) se(coef)    z     p
LBRT    3.3       27.1     1.73   1.9   0.057
Likelihood ratio test=13.1  on 1 df, p=0.0003  n= 54
```

ベースライン累積ハザード関数の計算法



存時間

上記のベースライン累積ハザード関数$\Lambda_0(t)=\int_0^t \lambda_0(u)du$の図は、時間依存型比例ハザードモデルの場合、

$$\hat{\Lambda}_0(t)=\sum_{i:t_i\le t}\left[\frac{\delta_i}{\sum_{l\in R(t_i)}\exp\left\{x_l(t_i)\hat{\beta}\right\}}\right]$$

から計算できる(Lawless,2003.7.1.8 節)。ここに、$x_l(t_i)$は観測時点t_iにおける共変量 LBRT の値で、

$$\delta_i = \begin{cases} 0 : \text{打切り} \\ 1 : o.w. \end{cases}$$

とする。内的時間依存型共変量(internal time-dependent covariates)が含まれる場合、ベースライン生存関数

$$S_0(t,x) = \exp\left\{-\int_0^t \lambda_0(u)du\right\}$$

は、解釈が困難である(Kalbfleish & Prentice,2002,6.4.1 節)。ベースライン累積ハザード関数の具体的な算出法を表 A.3 に与えておく。

表 A.3.　ベースライン累積ハザード関数の算出法

患者番号 (i)	死亡時点 t_i	打切りの有無 δ_i	リスク集合 $R(t_i)$	ベースラインハザード関数 $1 \Big/ \sum_{j\in R(t_i)} \exp\{3.308\times z_j(t_i)\}$	ベースライン累積ハザード関数 $\hat{\Lambda}_0(t)$
(8)		0			
(1)	1	1	(1)(5)(4)(3)(2)(11)(6)(12)(9)(10)(7)	0.0000000373†	0.0000000373
(5)		0			
(4)	2	1	(4)(3)(2)(11)(6)(12)(9)(10)(7)	0.0000000171	0.0000000544
(3)	3	1	(3)(2)(11)(6)(12)(9)(10)(7)	0.000000790	0.000000844
(2)		0			
(11)	4	1	(11)(6)(12)(9)(10)(7)	0.000000449	0.00000129
(6)	5	1	(6)(12)(9)(10)(7)	0.00000139	0.00000268
(12)	6	1	(12)(9)(10)(7)	0.00000123	0.00000390
(9)	7	1	(9)(10)(7)	0.00000188	0.00000577
(10)		0			
(7)	8	1	(7)	0.0000000471	0.00000582

†: $\sum_{j\in R(t_1)} \exp\{3.308\times z_j(t_1)\}$ =exp(3.308×5.0)+exp(3.308×2.9)+exp(3.308×4.9)+exp(3.308×3.2)
+exp(3.308×3.1)+exp(3.308×3.9)+exp(3.308×2.8)+exp(3.308×3.4)
+exp(3.308×2.5)+exp(3.308×2.2)+exp(3.308×3.0)

(1)* (5) (4) (3) (2) (11) (6) (12) (9) (10) (7)

=26809651.47

＜注*＞ ( ) 内の値は、患者番号で、対応する矢印の下の値は患者 1 の生存時間 281 以下で、最も近い観測時点の LBR 値(表 A.3 の※印)

$$\therefore \quad 1\Big/\sum_{j\in R(t_1)} \exp\left\{3.308\times z_j\left(t_1\right)\right\} = 1/26809651.47 = 0.0000000373$$