CREDIT　SAVER®

MATLAB 版 COX 回帰

テクニカルドキュメント

(株)数理技研

目次

１．はじめに

本文書は、MATLAB 版 COX 回帰プログラムの背景にある理論を、数式を用いて詳細に説明した文書である。2 章では、COX 比例ハザードモデルについて、簡単に説明する。3 章では、比例ハザードモデルの個人の重みを決定する係数の推定方法、4 章では比例ハザードモデルの全個人に共通の要因であるベースラインハザードの推定方法について述べる。5 章で、全変数から必要な変数だけを見つけるステップワイズ変数選択法を説明する。

２．COX 比例ハザードモデルとは

２．１　ハザードと比例ハザードモデル

本プログラムで実装する、COX 比例ハザードモデルは、生存時間解析の手法の 1 つです。生存時間解析は、ある患者がいつ死亡するか、いつ発病するかなど、「1 度しか起こらないような事象がいつどれくらいの確率で起きるか」ということを分析するために、医療分野を中心に発展してきた手法です。また、分析対象の患者が退院や転院などで、観察できなくなるような打ち切りデータを扱えるという特徴もあります。

生存時間解析を、CREDIT　SAVER の分析対象の 1 つである個人無担保融資のケースにあてはめると、死亡や発病をデフォルトと捉え、打ち切りを完済と捉えることで、同じような分析をすることが可能である。これによって、「デフォルトがいつどれくらいの確率で起きるか？」というデフォルト確率の期間構造の分析が可能となる。

この生存時間解析という手法は、ハザードという考え方が中心にある。デフォルトを例にすると、ある月に正常だった債権が、その次の月にデフォルトすることを生存時間解析ではハザードといい、正常債権に対するハザードする割合をハザード率という。例えば、7 月時点で 100 の正常債権があって、その次の月までに、10 の債権がデフォルトしたとすると、(実績の)ハザード率は 0.1 になる。

本システムのベースとなっている COX 比例ハザードモデルは、このハザード率が、全ての人に共通な部分と個人との重みに分けられると仮定するモデルである。全ての人に共通な部分のことをベースラインハザードと呼び、この値は時間をおって変化し、個人の重みは時間変化しないと仮定します。

ここで、個人の共変量を $\mathbf{x}$ とし、個人のハザード率、ベースラインハザードをそれぞれ $\lambda(t_i \mid \mathbf{x}), \lambda_0(t_i)$ とすると、COX 比例ハザードモデルは、

COX 比例ハザードモデルの定義

$$\lambda(\mathbf{x},t) = \lambda_0(t)\cdot\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$$

ハザード率　＝　ベースラインハザード　×　個人の重み

と定義できる。以下の節で、ベースラインハザードと個人の重みを決定する係数の推定方法について述べる。

２．２　ハザードと生存時間確率の関係

まず、連続時間の場合について、考える。ハザードは、死亡時刻を T とすると

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t \le T < t + \Delta t \mid T \ge t\}}{\Delta t}$$

と定義される。極限をとる前の式を考えると、この値は t の直前まで生きていた人が、$t + \Delta t$ までに死ぬ確率を単位時間当たりの量に変換した量である。つまり、ハザードは t まで生き

ていた人が次の瞬間に死ぬ確率を単位時間あたりの量に変換した値である。ハザード率の定義を生存時間関数$S(t) = \Pr\{T \geq t\}$を用いて書き直すと、

$$\begin{aligned}\lambda(t) &= \lim_{\Delta t \to 0} \frac{S(t) - S(t+\Delta t)}{S(t)\Delta t} \\ &= \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t} \\ &= \frac{f(t)}{S(t)}\end{aligned}$$

ここで、$F(t) = 1 - S(t)$は確率変数Tの分布関数で、$f(t) = dF(t)/dt$はTの密度関数である。上式を利用すると、生存時間関数とハザードの関係が以下のように導かれる。

$$\frac{d\log S(t)}{dt} = \frac{1}{S(t)}\frac{dS(t)}{dt} = \frac{-f(t)}{S(t)} = -\lambda(t)$$

この式より

$$\log S(t) = -\int_0^t \lambda(u)du$$

従って

$$S(t) = \exp\left\{-\int_0^t \lambda(u)du\right\} = \exp\{-\Lambda(t)\}$$

を得る。ここで、

$$\Lambda(t) = -\int_0^t \lambda(u)du$$

は累積ハザードと呼ばれる。

３．COX 比例ハザードモデルの回帰係数推定

３．１　最尤法

本プログラムでは、Breslow 法による近似を用いて、COX 比例ハザードモデルの回帰係数を最尤法を用いて推定する。その手順を以下に示す。

観察された死亡時間を$t_i \quad (i=1,...,k)$とする。t_i時点でのリスクセットをR_i、共変量を$\mathbf{x}_j$とすると、部分尤度関数は、

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}} S_i)}{\{\sum_{j \in R_i} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j)\}^{d_j}}$$

で与えられる。ここで、d_iはt_i時点で死亡した数、S_iはt_i時点での全死亡例の共変量の和である。以上より、対数尤度は

$$l(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{k} \boldsymbol{\beta}^{\mathrm{T}} S_i - \sum_{i=1}^{k} d_i \ln\left(\sum_{j \in R_i} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j)\right)$$

となるので、その一階微分$U(\boldsymbol{\beta})$のr成分は、

$$\frac{\partial l}{\partial \beta_r} = \sum_{i=1}^{k} \left(S_i^{(r)} - d_i \frac{\sum_{j \in R_i} x_j^{(r)} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j)}{\sum_{j \in R_i} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j)} \right)$$

で与えられるので、COX 比例ハザードモデルの回帰係数$\boldsymbol{\beta}$の最尤推定量は、各要素$r=1,...,p$について、上式を 0 とおいた連立方程式を解くことにより求められる。その連立方程式の解を求めるためには、Newton-Raphson 法を用いる。

対数尤度関数の二階微分のマイマスは、情報行列$\mathbf{I}(\boldsymbol{\beta})$であり、その$(r,s)$成分は、

$$\begin{aligned}\mathbf{I}_{rs} &= -\frac{\partial^2}{\partial \beta_r \partial \beta_s} \ln L \\ &= \sum_{i=1}^{k} d_i \left[\frac{\sum_{j \in R_i} x_j^{(r)} x_j^{(s)} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j)}{\sum_{j \in R_i} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j)} - \frac{\left(\sum_{j \in R_i} x_j^{(r)} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j)\right)\left(\sum_{j \in R_i} x_j^{(r)} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j)\right)}{\left(\sum_{j \in R_i} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j)\right)^2} \right]\end{aligned}$$

となるので、$\boldsymbol{\beta}$の最尤推定量は

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

の繰り返し計算から得られる。

３．２　Newton-Raphson 法の実装

本プログラムでは、Newton-Raphson 法による係数推定値の初期値を$\boldsymbol{\beta}_0 = \mathbf{0}$として、収束判定条件が満たされるまで、

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

を繰返し計算する。収束判定条件は、次の 4 つが指定できる。

①尤度の変化の絶対値

$$\left|l(\boldsymbol{\beta}_m)-l(\boldsymbol{\beta}_{m-1})\right|<\varepsilon$$

②尤度の相対変化の絶対値

$$\frac{\left|l(\boldsymbol{\beta}_m)-l(\boldsymbol{\beta}_{m-1})\right|}{\left|l(\boldsymbol{\beta}_{m-1})\right|+1e-6}<\varepsilon$$

$1e-6$ はゼロ割を防ぐためである。

③係数推定値の相対変化の絶対値(全変数のうち、最も変化が大きいものを基準)

$$\max_r|\delta_m^{(r)}|<\varepsilon$$

$$\delta_m^{(r)}=\begin{cases}\boldsymbol{\beta}_m^{(r)}-\boldsymbol{\beta}_{m-1}^{(r)} & |\boldsymbol{\beta}_{m-1}^{(r)}|<.01\\ (\boldsymbol{\beta}_m^{(r)}-\boldsymbol{\beta}_{m-1}^{(r)})/\boldsymbol{\beta}_{m-1}^{(r)} & \text{otherwise}\end{cases}$$

④対数尤度関数の勾配(対数尤度の 1 次微分)

$$\frac{\left|\mathbf{g}_m\mathbf{I}^{-1}(\boldsymbol{\beta}_m)\mathbf{g}_m\right|}{\left|l(\boldsymbol{\beta}_{m-1})\right|+1e-6}<\varepsilon$$

デフォルトは、○○である。

また、情報行列が特異な行列に近い場合などに

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

に従うと、尤度が下がる場合がある。その場合は、

$$\boldsymbol{\beta}_{m+1}=\boldsymbol{\beta}_m+\mathbf{I}^{-1}(\boldsymbol{\beta}_m)U(\boldsymbol{\beta}_m)/2^p\quad p=1,2,3,...$$

として、尤度が上がるところまで、p を 1 から順に大きくしながら係数推定値の変化を小さくしていく。

４．ベースラインハザードの推定

本章では、ベースラインハザードの推定方法について述べる。比例ハザードモデルと生存時間関数の関係を議論したうえで、ベースラインハザードの推定方法について、説明する。

４．１　比例ハザードモデルと生存関数の関係

T を死亡時間を示す確率変数、$t_1,...,t_i,...$を離散時間とする。共変量$\mathbf{x}$のt_i時点でのハザード率、ベースライン$\mathbf{x}=0$でのt_i時点でのハザード率をそれぞれ、$\lambda(t_i \mid \mathbf{x}), \lambda_0(t_i)$とすると、$t_i$時点における生存関数は、それぞれ

$$S(t_i \mid \mathbf{x}) = \Pr\{T \geq t_i \mid \mathbf{x}\} = \{1-\lambda(t_1 \mid \mathbf{x})\} \Lambda\ \{1-\lambda(t_{i-1} \mid \mathbf{x})\}$$
$$S_0(t_i) = \Pr\{T \geq t_i \mid \mathbf{x}=0\} = \{1-\lambda_0(t_1)\} \Lambda\ \{1-\lambda_0(t_{i-1})\}$$

となる。これらの生存関数が比例ハザード性をもつと仮定すると、ある関数rについて

$$S(t_i \mid \mathbf{x}) = S_0(t_i)^{r(\mathbf{z})}$$

が任意のiについて成立する。そこで、

$$1-\lambda(t_i \mid \mathbf{x}) = \{1-\lambda_0(t_i)\}^{r(\mathbf{x})}$$

を比例ハザードモデルの定義とする。対数線形性を仮定すれば、

$$S(t_i \mid \mathbf{x}) = S_0(t_i)^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}$$

$$1-\lambda(t_i \mid \mathbf{x}) = \{1-\lambda_0(t_i)\}^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}$$

となる。

４．２　COX 比例ハザードモデルにおける生存関数の最尤推定値

ここで、$\alpha_i = 1-\lambda_0(t_i)$（$\alpha_i$はベースラインでの$t_i$から$t_{i+1}$時点までの生存確率となる）と仮定すると、

$$\lambda(t_i \mid \mathbf{x}) = 1-\alpha_i^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}$$

となる。一方、全尤度Lは、D_i, R_iをそれぞれt_i時点の死亡数とリスクセットとすると、

$$L = \prod_{i=1}^{k}\left[\prod_{j\in D_i}\left\{1-\alpha_i^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j)}\right\}\prod_{l\in R_i-D_i}\alpha_i^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j)}\right]$$

となるので、対数をα_iで偏微分して０とおくと、連立方程式から

$$\sum_{j\in D_i}\frac{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j)}{1-\alpha_i^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j)}}=\sum_{l\in R_i}\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_l),\quad i=1,2,...,m$$

を得る。もしタイが無ければ、各 D_i に所属する要素は 1 つなので、

$$\hat{\alpha}_i=\left(1-\frac{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{(i)})}{\sum_{l\in R_i}\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_l)}\right)^{\exp(-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{(i)})},\quad i=1,2,...,m$$

を得る。タイがある場合は、Newton-Raphson 法で求める(詳しく補足)。以上よりベースライン生存関数の最尤推定値

$$S_0(t)=\prod_{t_i<t}\hat{\alpha}_i$$

を得る。

(補足)タイデータが存在する場合
さきの連立方程式から、

$$f(\alpha_i)=\sum_{j\in D_i}\frac{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j)}{1-\alpha_i^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j)}}-\sum_{l\in R_i}\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_l)$$

とおくと、その 1 階微分は

$$f'(\alpha_i)=\sum_{j\in D_i}\frac{\exp^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j)\alpha_i^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j)-1}}{(1-\alpha_i^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j)})^2}$$

となるので、Newton-Raphson 法による繰り返し計算を利用して、$f(\alpha_i)=0$ を満たす(尤度関数を最大にする $\hat{\alpha}_i$ を求める。

$$\alpha_{i(n+1)}=\alpha_{i(n)}-f(\alpha_{i(n)})\Big/f'(\alpha_{i(n)})$$

Newton-Raphson 法の初期値は、

$$\hat{\alpha}_i=\exp\left(\frac{-d_i}{\sum_{l\in R_i}\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_l)}\right)$$

と与えればよいことが知られている(Lawless, 1982)。

４．３　Newton-Raphson 法の実装

初期値は、前節の最後に与えたとおりで、収束判定値は、

前節の関数f及びその1階微分は$1-\alpha_i^{\exp(\boldsymbol{\beta}^T\mathbf{x}_j)}$の値が0になってしまうと、ゼロ割が発生するので、もし$1-\alpha_i^{\exp(\boldsymbol{\beta}^T\mathbf{x}_j)}$が0になった場合は、1e-6に置き換えて計算する。一方、f'が0になった場合も同様である。また、α_iは確率なので[0,1]の値になるはずであるが、もし推定値が0以下になった場合は0に1以上になった場合は1にする。

５．変数選択法

本プログラムでは、尤度比によるスッテプワイズ変数選択法を採用している。本章では、変数選択法の手順と変数選択に使用する統計量について説明する。

５．１　スッテプワイズ変数選択法の手順

本プログラムのステップワイズ選択法は、以下のような手順により行なわれている。

変数選択法の手順

①全ての変数のうち、モデルに無い変数のスコア統計量を計算する。そのうち、もっとも p 値が低い変数の p 値が投入基準以下だったら、その変数をモデルに追加する。モデルに追加する変数が無ければ、そこで終了する。

②全ての変数のうち、モデルにある変数の尤度比統計量、または Wald 統計量を計算する。最も p 値の大きい変数の p 地が除去基準以上だったら、その変数を除外する。①へ。

５．２　変数選択に使用する統計量

５．２．１　スコア統計量

変数の投入基準は、SAS、SPSS と同様に SCORE 統計量を用いる。スコア統計量は、次のように定義される。

まず､現在モデルにある変数の COX 回帰係数とモデルに無い変数は回帰係数を 0 として、情報行列 $\mathbf{I}$ を計算する。次に、情報行列 $\mathbf{I}$ を以下のように 4 つの部分行列に分解する。

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

ここで、$\mathbf{A}_{11}$ と $\mathbf{A}_{22}$ は、それぞれ、モデルに含まれる変数、含まれない変数についての正方行列である。$\mathbf{A}_{12}$ は、モデル内にある変数と無い変数の交互積行列である。そして、ある 1 つの変数 $\mathbf{x}_i$ についての、スコア統計量は、

$$\mathbf{D}'_{x_i}\,\mathbf{B}_{22}\mathbf{D}_{x_i}$$

$$\mathbf{B}_{22,i} = (\mathbf{A}_{22,i} - \mathbf{A}_{21,i}\mathbf{A}_{11}^{-1}\mathbf{A}_{12,i})^{-1}$$

と定義される。ここで、$\mathbf{D}_{x_i}$ は対数尤度関数の 1 階微分の値である。対数尤度の、変数 $\mathbf{x}_i$ に対応する成分である。$\mathbf{A}_{22,i}$ と $\mathbf{A}_{12,i}$ は、それぞれ、変数 $\mathbf{x}_i$ に対応する $\mathbf{A}_{22}$ と $\mathbf{A}_{12}$ の部分行列である。このとき、このスコア統計量は、自由度 1 の χ^2 分布に従う。

５．２．２　Wald 統計量と尤度比統計量

変数除去の基準は、Wald 統計量と尤度比の 2 つから選択可能である。それぞれ、以下のように定義される。

Wald 統計量

そのステップで、モデルにある変数で COX 回帰を行い、モデル内にある変数 $\mathbf{x}_j$ に対する COX 回帰係数を $\hat{\beta}_j$ とする。この時、Wald 統計量は、

$$\hat{\beta}'_j \mathbf{B}_{11,j} \hat{\beta}_j$$

と定義される。ここで、$\mathbf{B}_{11,j}$ は $\mathbf{A}_{11}^{-1}$ の $\mathbf{x}_j$ に対応する部分行列である。このとき、Wald 統計量は自由度 1 の χ^2 分布に従う。

尤度比統計量

そのステップでモデル内にある全ての変数を使用したときの対数尤度を $l(full)$、現在のモデルから $\mathbf{x}_j$ を取り除いたときの対数尤度を $l(reduced)$ とすると、$\mathbf{x}_j$ に対する尤度比統計量は、次のように定義される。

$$-2(l(reduced)-l(full))$$

この統計量も、自由度 1 の χ^2 分布に従う。

＊）以上のスコア統計量、Wald 統計量、尤度比統計量の議論は、対象となる変数が $k(>2)$ 個の場合も同様に成り立つ。その際は、自由度 k の χ^2 分布に従う。