Cox の比例ハザードモデルについて

土居正明

Q.Cox の比例ハザードモデルとは何なのか、よく分かりません。

1 回帰分析の話

まず、回帰分析とは何かから始めましょう。回帰分析、とは一言で言うと「データにグラフを当てはめる」ことです。

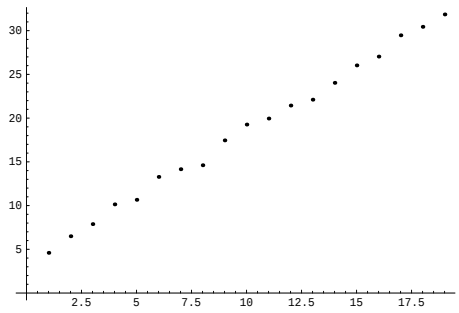

図 1　直線を当てはめたらよさそうなデータ

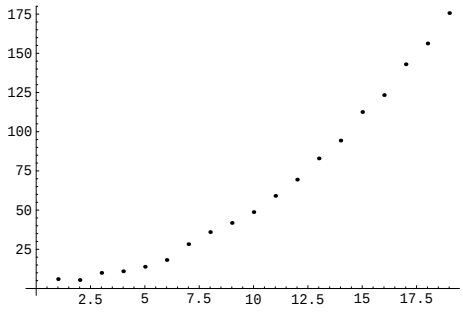

図 2　放物線を当てはめたらよさそうなデータ

上の図 1. は直線を、図 2. は放物線を当てはめたらよさそうに思えます。そうすると、図 1 の場合は $y = ax + b$、図 2 の場合は $y = ax^2 + bx + c$ のようにおいて、データに最も合うような a, b, c を推定していきます。これが回帰分析と呼ばれる手法です。このときの「$y = ax + b$」や「$y = ax^2 + bx + c$」のような式ことを、**「モデル（統計モデル）」**と呼びます。モデルとは要は「データを当てはめるグラフの式」のことです。

2 Cox 回帰とは

「Cox 回帰」とは、Cox の比例ハザードモデルを用いた回帰分析 のことです。では、Cox の比例ハザードモデルとはどのようなモデルなのでしょうか。まず最初に大事なのは、これは**「ハザードをデータとしたモデル」**である、ということです。つまり、**ハザード（正確にはハザードの推定値）をプロットして、グラフ当てはめをするときの、そのグラフのことを Cox の比例ハザードモデルと呼ぶ**のです[*1]。

さて、ではまずモデルを眺めてから、ハザードの定義など詳細に入ります。が、大体のイメージで「ハザード」＝「死亡率」、「共変量」＝「薬も含めて、タバコ、飲酒、カロリーなど、死亡率に影響を与えるもの」くらいと思っておいてください。

2.1 Cox の比例ハザードモデル

時間 t、共変量 $x_1, x_2, x_3, \cdots x_n$ (x_1 を薬の有無とすることが多いです) のときのハザード $\lambda(t|x_1, \cdots, x_n)$ を

$$\lambda(t|x_1, \cdots, x_n) = \lambda_0(t) \exp(\beta_1 x_1 + \cdots + \beta_n x_n) \tag{1}$$

[*1] 厳密には、ハザードのうちの一部にグラフを当てはめます。データそのものにグラフを当てはめる手法を「パラメトリック」、データにグラフを当てはめない手法を「ノンパラメトリック」といいますが、今回は 一部にだけ当てはめる ので**「セミパラメトリック」**といいます。そのため、グラフのあてはめ（推定）にも通常の最尤法は使えず、部分尤度というものを用います。

とおいたモデルを、「Cox の比例ハザードモデル」と呼びます[*2]。

詳しく見ていきます。まず最初に時間 t に依存する部分 $\lambda_0(t)$(ベースラインハザードと呼びます。共変量には依存しません）と $x_1, \cdots, x_n$ に依存する部分 $\exp(\beta_1 x_1 + \cdots + \beta_n x_n)$（時間には依存しません) の掛け算に分けます。これを「変数分離法」と言い、数学では微分方程式を解くときなどによく使います。要は「別々に分けた方が分かりやすくなる」ということです。ここで、$\lambda_0(t)$ については、何も仮定を置きません（厳密には連続性くらいは仮定していると思いますが)。次に、共変量 $x_1, \cdots, x_n$ に対してですが、こちらにはもっと強い仮定がついています。共変量の一次結合 $(\beta_1 x_1 + \cdots + \beta_n x_n)$ の指数関数になっている、というものです。この仮定を「対数を取ったら線形になる」という意味で**対数線形性**と呼びます。

この状況で、データとしてハザードを与えてプロットして、そのデータの当てはまるモデル（グラフ）を探します。具体的には、$\beta_1, \cdots, \beta_n$ を推定します。「データがハザードとか言ってるけど、ハザードってどうやったら求まるんですか？」と思われた人はいるでしょうか？[*3]。これについては、もう少し先に述べますので、しばらくお待ちください（いくら何でも、ハザードの定義を述べる前には無理ですので)。

2.2 ハザードとは

ハザードの定義は、以下のようになっています。T をある症例の生存時間とし、時刻 t のハザードは

$$\lambda(t|x_1, \cdots, x_n) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t | t \leq T)}{\Delta t} \tag{2}$$

です。解説しますと、要はハザードとは「人年法で考えた、非常に微小な時間における死亡率」のことです。面倒なので、lim を考えないで、代わりに Δt を「すごく小さい時間」と思いましょう。

具体例で考えていきます。以下の図で縦軸は症例番号、「○」は死亡の発生、「Dt」は Δt と考えてください。

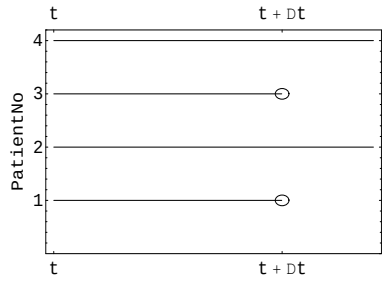

図 3　時点 t において生きている集団のその後

大事な注意が 2 つあります。

(a)死亡は常に時点 $(t + \Delta t)$ で起きたものと考える

(b)時点 t〜時点 $(t + \Delta t)$ の間にセンサーは 1 件もない

このようなことを仮定してよいのでしょうか？よいのです。なぜなら、時点 t の段階では生きていることを仮定されていますし、時間をものすごく小さく ($\lim_{\Delta t \to 0}$ に) してあるからです (死亡とセンサーが本当に同時に起こることがまれにありますが、そういうときは確か SAS では、「一瞬先に死亡が起きた」ことにして解消しているはずです)。

このとき、(時点 t〜時点 $(t + \Delta t)$ の間の総人年)=(時点 t で生きている総人数) $\times$ Δt となります。これより、人年法による死亡率は、

$$\frac{\text{(時点 } t\text{〜時点 } (t + \Delta t) \text{ の間に死亡した人数)}}{\text{(時点 } t\text{〜時点 } (t + \Delta t) \text{ までの総人年)}} = \frac{\text{(時点 } t\text{〜時点 } (t + \Delta t) \text{ の間に死亡した人数)}}{\text{(時点 } t \text{ の段階で生きている人数)} \cdot \Delta t}$$

[*2] この $\lambda_0(t)$ の部分は推定しないでそのまま放っておき、$\exp(\beta_1 x_1 + \cdots + \beta_n x_n)$ の部分だけ推定を行います。このように「推定する部分」と「推定しない部分」に分かれているので、「セミパラメトリック」といいます。

[*3] 統計の数式を理解するときに、個人的に最も大事だと思うのは「何がデータで何を知りたいのか」をはっきりさせることだと思います。このような疑問を常に意識するようにするとよいと思います。

となります。ここで、

$$\frac{\text{(時点 } t\text{〜時点 } (t+\Delta t) \text{ の間に死亡した人数)}}{\text{(時点 } t \text{ の段階で生きている人数)}} = Pr(t \leq T < t+\Delta t | t \leq T)$$

という風に考えることができます。従って、

$$\begin{aligned}\text{(時点 } t\text{〜時点 } (t+\Delta t) \text{ における人年法による死亡率)} &= \frac{\text{(時点 } t\text{〜時点 } (t+\Delta t) \text{ の間に死亡した人数)}}{\text{(時点 } t\text{〜時点 } (t+\Delta t) \text{ までの総人年)}} \\ &= \frac{Pr(t \leq T < t+\Delta t | t \leq T)}{\Delta t}\end{aligned}$$

となり、時間 Δt が非常に小さいとき ($\lim_{\Delta t \to 0}$) に、(2) と合わせて考えると

$$\begin{aligned}\lambda(t|x_1, \cdots, x_n) &= \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t+\Delta t | t \leq T)}{\Delta t} \\ &= \lim_{\Delta t \to 0} (\text{時点 } t\text{〜時点 } (t+\Delta t) \text{ における人年法による死亡率})\end{aligned}$$

となります。

ハザードとは「人年法で考えた、非常に微小な時間における死亡率」であることがお分かりいただけたでしょうか。

2.3 共変量について

共変量とは、一言で言うと 薬も含めた、ハザード (または生存関数) に影響を与える因子 のことです。**共変量が存在するのにモデルに組み込むことを忘れると、検定において検出力の低下を招く**ので、存在する共変量はモデルに組み込むようにしましょう。共変量かどうか分からない場合は、「とりあえず入れておいてあとから本当にハザードに影響を与えるか判断する」こともできますので、気になるものは入れておくとよいと思います。

2.4 ハザードの求め方

さて、最初に「データはハザードだ」と言いました。そのハザードの求め方 (推定方法) を見ていきましょう。式の計算は省略しますが、以下の関係式が成り立ちます。$S(t|x_1, \cdots, x_n)$ を共変量 $x_1, \cdots, x_n$ を持つ人の生存関数としたときに、

$$S(t|x_1, \cdots, x_n) = \exp\Big(-\int_0^t \lambda(t|x_1, \cdots, x_n) dt\Big) \tag{3}$$

さて、この式からハザードを推定することができます。何故なら、今「ハザードがいくつか」は分かりませんが「生存関数」は Kaplan-Meier 推定量で求まるからです。そして、今登場人物は「ハザード」、「生存関数」の 2 人なのですから、(3) の式に Kaplan-Meier 法で求めた推定値を代入すれば、ハザードの推定値を得ることができます。

3 比例ハザードモデルを使ってよいとき、ダメなとき

3.1 2 群比較の際の仮定について

さて、今までは「1 つの群におけるハザードの推定方法」について見ていきました。これからは、**「2 群（実薬・プラセボ）を比較するとき、どのような場合に比例ハザードモデルを使ってよいのか」**ということを考えていきます。

x_1 を、薬を表す共変量とし、実薬群は $x_1 = 1$、プラセボ群は $x_1 = 0$ とします。ここで、以下の仮定をします。

仮定：両群のハザードは、共変量 x_1 の値 (=1 か=0 か) を除いて等しい[*4]

[*4] この仮定が満たされるかどうかは、あとできちんとチェックします

どういうことかと言いますと、薬の影響以外では、両群の生存関数 (生存時間) が等しい ということです。薬の影響を比較したいのですから、至極妥当な仮定であることが理解していただけるかと思います。

さて、この仮定が満たされるとき (1) より、両群ともハザードは

$$\lambda(t|x_1,\cdots,x_n)=\lambda_0(t)\exp(\beta_1x_1+\cdots+\beta_nx_n)$$

の形に書けます（特に、時間に依存するベースラインハザード $\lambda_0(t)$ が両群で同じになっています)。さて、話を簡単にするため、以下、薬以外の共変量は考えないことにします。つまり、$x_2=\cdots=x_n=0$ とします。
こうすると、ハザードは $\lambda(t|x_1)=\lambda_0(t)\exp(\beta_1x_1)$ となり、実薬群のハザードは

$$\lambda(t|x_1=1)=\lambda_0(t)\exp(\beta_1\cdot1)=\lambda_0(t)\exp(\beta_1)$$

プラセボ群のハザードは

$$\lambda(t|x_1=0)=\lambda_0(t)\exp(\beta_1\cdot0)=\lambda_0(t)$$

となります。

3.2 仮定から導かれる性質

さて、上の仮定から 目で見て分りやすい性質が導かれます。そのために、少し計算をしましょう。目的は、β_1 をもっと分りやすい場所に連れてくることです。各群のハザードを (3) の式に代入していきます。
まずは実薬群から

$$S(t|x_1=1)=\exp\Big(-\int_0^t\lambda(t|x_1=1)dt\Big) \tag{4}$$

$$\log(S(t|x_1=1))=-\int_0^t\lambda_1(t|x_1=1)dt \tag{5}$$

$$=-\int_0^t\lambda_0(t)\exp(\beta_1)dt \tag{6}$$

$$=-\Big(\int_0^t\lambda_0(t)dt\Big)\exp(\beta_1) \tag{7}$$

$$\log(-\log(S(t|x_1=1)))=\log\Big(\int_0^t\lambda_0(t)dt\Big)+\beta_1 \tag{8}$$

同様にプラセボ群では

$$S(t|x_1=0)=\exp\Big(-\int_0^t\lambda(t|x_1=0)dt\Big) \tag{9}$$

$$\log(S(t|x_1=0))=-\int_0^t\lambda_0(t|x_1=0)dt \tag{10}$$

$$=-\int_0^t\lambda_0(t)dt \tag{11}$$

$$=-\Big(\int_0^t\lambda_0(t)dt\Big)) \tag{12}$$

$$\log(-\log(S(t|x_1=0)))=\log\Big(\int_0^t\lambda_0(t)dt\Big) \tag{13}$$

となります。β_1 がとても分りやすい場所にでてきました。これをもとに、「仮定の妥当性」＝「Cox の比例ハザードモデルを使ってよいかどうか」と考えましょう。

3.3 使ってよいかどうかの判定法

さて､先の式より何が分かるでしょうか。積分の部分は全く同じ形をしていますね。(8) と (13) は、β_1(薬の効き目) 分だけ平行移動させた関係にある、ということです。重要な点は**時間 t によらず、常に β_1 だけ等間隔にずれている**ということです。これは、先の「仮定」を満たしているから ですね。

逆に言うと（「対偶」というやつです)、**「時間によらず** $\log(-\log(S(t|x_1=0))$ **と** $\log(-\log(S(t|x_1=1))$ **が平行になっていない」のなら、「仮定が満たされていない」**ということになります。

これが、Cox の比例ハザードモデルを使ってよいかどうかの判定条件 となります。ここで、$S(t|x_1=0)$ と $S(t|x_1=1)$ には、Kaplan-Meier 推定量 $\hat{S}(t|x_1=0)$ と $\hat{S}(t|x_1=1)$ をあてはめて、$\log(-\log(\hat{S}(t|x_1=0))$ **と** $\log(-\log(\hat{S}(t|x_1=1))$ **をプロットしてみて、「グラフの間隔が時間によらず一定」であるときには Cox の比例ハザードモデルを使ってよい。時間によって間隔が変化するときは使ってはいけない**のです。以下の図はすべて、横軸に時間をとり $\log(-\log(\hat{S}(t|x_1=0))$ や $\log(-\log(\hat{S}(t|x_1=1))$ をプロットしたものです。

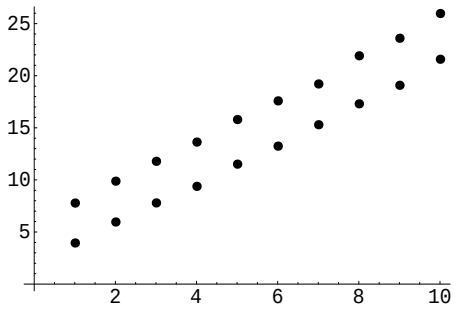

図 4　Cox の比例ハザードモデルを使えそうなデータ 1

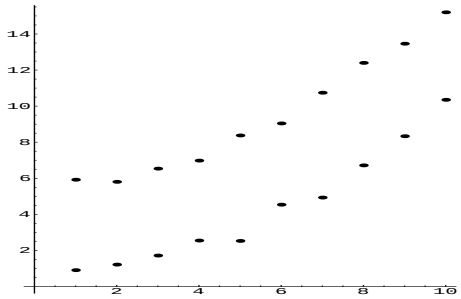

図 5　Cox の比例ハザードモデルを使えそうなデータ 2

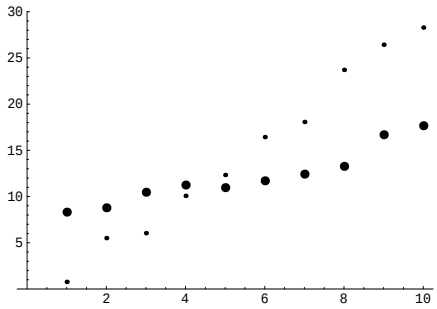

図 6　Cox の比例ハザードモデルを**使えなさそう**なデータ 1

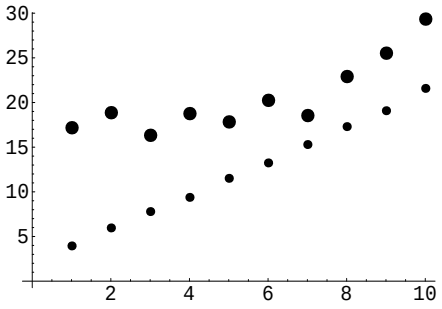

図 7　Cox の比例ハザードモデルを**使えなさそう**なデータ 2

4　パラメータの推定方法

今までは、簡単のため共変量は x_1 しか考えませんでしたが、**これから最後まで少しの間、共変量が** $x_2,\cdots,x_n$ **まで含まれている場合を考えます**。$\beta_1,\cdots\beta_n$ の推定法については、今回はご説明しません。最尤法の子孫である「部分尤度法」を用いる、ということだけ述べておきます。このあたりは式の計算が面倒な方は SAS にお任せするのも 1 つだと思います。

5　モデルに当てはめたあと

パラメータを推定して、$\beta_1,\cdots,\beta_n$ が求まったとします。その後に行うこととしては、主に

(a)$H_0:\beta_1=0$ を検定する（薬の効き目があるかどうか)。

(b)$x_2,\cdots,x_n$ がモデルに含まれるかどうか判断する（薬以外のどの共変量が生存時間に影響を与えているか)。

(c) 別の被験者の予後を予測する。

などがあげられます。