生存時間モデルの推定精度に関する研究*

A Study on Estimation Efficiency of Hazard-Based Duration Model*

中川　展孝**・山本　俊行***

By Noritaka NAKAGAWA**・Toshiyuki YAMAMOTO***

1. はじめに

生存時間モデル（hazard-based duration model）は，交通行動分析の分野においても自動車保有期間や自由活動時間の分析などに適用されるようになってきた．共変量を含む生存時間モデルには特定の基準ハザード分布を仮定するparametric modelと基準ハザード分布を特定しないsemi-parametric modelの２つのモデルが存在する．従来，semi-parametric modelの推定にはCox回帰モデルが適用されてきた．Cox回帰モデルでは，基準ハザード値を局外母数とし，部分尤度法により共変量の係数を推定する．よって，基準ハザード関数の特定の過誤による影響を回避することが可能である一方で，係数の推定精度はparametric modelより劣るとされてきた．

しかしながらsemi-parametric modelでも，Prentice and Gloeckler[1)]により提案されたordered-response modelを用いると係数の推定精度はparametric modelと同等であるとの知見[2)]もある．この知見に従えば，parametric modelを適用する積極的な理由はなくなり，常にsemi-parametric modelを適用すべきである[3)]．しかしながら，生存時間モデルを用いた分析では，共変量の係数のみに興味があるのではなく，予測モデルとして用いるために基準ハザード値の推定精度も重要な場合が多い．残念ながら上述の知見を導いた分析では，基準ハザード値の推定精度については考慮されていない．

そこで本研究では，シミレーションデータを用いて，各共変量の係数の推定精度のみならず，特に基準ハザード値の推定精度の確認に主眼を置きつつordered-response modelを用いた推定方法によるsemi-parametric modelの有効性を検証することを目的とする．

*キーワーズ：交通行動分析

**学生員，名古屋大学大学院工学研究科
（愛知県名古屋市千種区不老町，TEL:052-789-3565
E-mail: nakagawa@trans.civil.nagoya-u.ac.jp）
**正員，博(工)，名古屋大学大学院工学研究科
（TEL:052-789-4636，E-mail: yamamoto@civil.nagoya-u.ac.jp）

2. 生存時間モデル

（１）生存時間関数とハザード関数

生存時間モデルでは，解析の対象とする事象が生起するまでの時間Tの分布を表現するのに，生存関数（survival function），あるいはハザード関数（hazard function）を用いることが多い．生存関数とは，対象とする事象がある時点tにおいてまだ生起していない確率を表すものである．またハザード関数とは対象とする事象がある時点tまでに生起していないという条件の下で，次の瞬間に事象が生起するという条件付きの確率密度である．

本研究ではハザード関数に以下の式で表される比例ハザードモデル（proportional hazard model）を用いる．

$$h(t \mid Xi) = h_0(t)\exp(-bXi) \tag{1}$$

ここで，$h(t \mid Xi)$：共変量ベクトルXiを持つケースiのハザード関数，$h_0(t)$：基準ハザード値（全ての共変量ベクトルXiが0の時のハザード関数），b：未知パラメータベクトルである．

（２）parametric model

parametric modelは共変量をパラメータとしてモデルに導入し，かつ生存時間の分布にも特定の確率分布を仮定し分析を行うものである．本研究では確率分布にワイブル分布を仮定し分析を行う．ワイブル分布を仮定した場合のハザード関数は以下の式で表される．

$$h(t \mid Xi) = gt^{g-1}l \cdot \exp(-bXi) \tag{2}$$

ここで，g，l：未知パラメータである．また，生存

関数は，以下の式で表される．

$$S(t \mid Xi) = \exp(-l t^{g} \exp(-bXi)) \qquad (3)$$

parametric model は以下の対数尤度関数 LL を用いて g ， l とパラメータベクトル b の推定を行う．

$$LL = \sum_{i=1}^{N} \ln\{h(t \mid Xi) \times S(t \mid Xi)\}$$
$$= \sum_{i=1}^{N} \ln\left\{g t^{g-1} l \cdot \exp(-bXi) \times \exp(-l t^{g} \exp(-bXi))\right\} \qquad (4)$$

ここで， N ：サンプル数である．

（3）semi-parametric model

semi-parametric model は共変量をパラメータとしてモデルに導入するが，生存時間の分布には特定の確率分布を仮定せずに分析を行うものである．Ordered-response model を適用した推定方法では，対象となる生存時間を $[0<t\le t_1],[t_1<t\le t_2],....[t_{K-1}<t\le\infty]$ の K 個の区間としてとらえる．さらに，ガンベル分布 G に従う誤差項 e_i を用いると，$[t_{k-1}<t\le t_k]$ の区間において事象が生起する確率は以下の式で表される．

$$\Pr(t_{k-1}<t\le t_k) = G(d_k - bXi) - G(d_{k-1} - bXi) \qquad (5)$$

ここで， $d_k = \ln \int_0^k h_0(u)du$ である．

semi-parametric model は以下の対数尤度関数 LL を用いることによりパラメータベクトル b ，および未知パラメータ $d_1, d_2, \ldots, d_{K-1}$ を推定することが可能である．

$$LL = \sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \ln\{G(d_k - bXi) - G(d_{k-1} - bXi)\} \qquad (6)$$

ここで，N：サンプル数，y_{ik}：ケース i の事象が区間 $[t_{k-1}<t\le t_k]$ で生起した場合 1，それ以外の時は 0 のダミー変数，$d_0 = -8$ ，$d_K = +8$ である．

またこのとき，各区間内で基準ハザード値が一定であると仮定すると，基準ハザード値 $h_0(t)$ は得られた未知パラメータ推定値 $\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_{K-1}$ を用いて，

$$h_0(t) = \frac{\exp(\hat{d}_k) - \exp(\hat{d}_{k-1})}{t_k - t_{k-1}}$$
$$for\ t_{k-1} < t \le t_k \quad k = 1,2,\cdots K-1 \qquad (7)$$

により求められる．

3. シミュレーションデータの概要

（1）生存時間にワイブル分布を仮定したデータ

本研究では，乱数により X_1（0，1 のダミー変数），X_2（N（0，1）の標準正規分布に従う連続変数）の 2 つの共変量データを生成し，以下の式で表されるワイブル分布に従う生存時間 t を生成した．

$$F(t) = 1 - \exp\left\{-t^{g} \exp(-bX)\right\} \qquad (8)$$

ここで，$F(t)$：累積分布関数，$bX = a_0 + b_1 X_1 + b_2 X_2$（$g$ ，a_0 ，b_1 ，b_2 ：パラメータ）である．本研究では，パラメータを $g = 1.5$，$a_0 = 5.0$，$b_1 = 2.0$，$b_2 = 1.0$ と設定し，ワイブル分布に従う生存時間 t を得た．またサンプル数を 100, 1000 としたものをそれぞれ 10 組ずつ作成し，各サンプルを用いたパラメータの推定を行い，parametric model と semi-parametric model の推定精度の比較を行うこととした．

（2）時間依存性共変量を考慮したデータ

本研究では，時間依存性共変量の導入の有無による推定精度の違いを検討するため，前節で作成したデータのうち $t \geqq 8$ の場合のみに影響を及ぼす時間依存性共変量として，式（8）の bX の中に新たに X_3（（0，1）のダミー変数）を導入した．また X_3 の導入に伴い，

$$b_2 X = a_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 \qquad (9)$$

とした $b_2 X$ を新たに導入した．またパラメータは $b_3 = -1.0$ と設定した．ここで式（9）より， X_3 の値が 1 である場合，$t \geqq 8$ の生存時間の分布が左側に移動する（事象の生起が早くなる）ことになる．

4. シミュレーションの推定結果と考察

本研究のシミュレーションデータは乱数を用いて作成しているため，同一モデルを用いて推定した場合でも相異なるパラメータ推定値を得る．そこで「推定値の平均」「推定値の標準偏差の平均」「Mean Square Error(=MSE)」の指標により推定精度を検証した．なお今回は紙面の都合上，グラフや表などの結果の詳細については時間依存性共変量を含まない場合のサンプル数 1000 の場合の結果のみを掲載する．

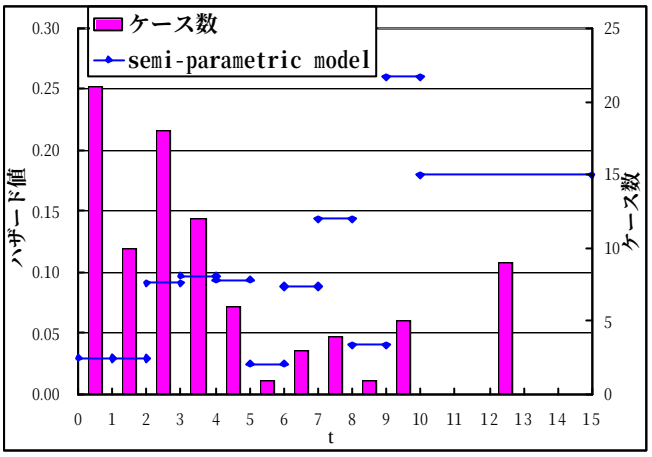

図 1　サンプル分布と ordered-response model による推定値

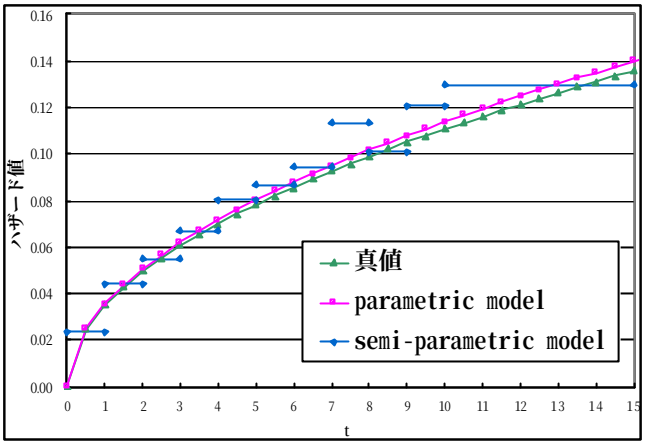

図 2　基準ハザード値の推定結果

（1）共変量のパラメータベクトルの推定結果

まず時間依存性を持たない共変量パラメータベクトルは，サンプル数に関わらず parametric model の場合も semi-parametric model の場合も不偏的に推定されており，推定精度は変わらないことが示された．

次に，時間依存性を持つ共変量のパラメータベクトルの推定結果に関しては，時間依存性を持たない共変量に比べて若干精度が劣ることが確認された．よって semi-parametric model は時間依存性共変量を含むモデルの推定には適さないと考えられる．この結果は Meyer[2] の結果と一致するものである．

（2）基準ハザード値の推定結果

Ordered-response model による semi-parametric model の推定に際しては，まず基準ハザード値が一定となる区間を外生的に設定する必要がある．そこで本研究では対象となる生存時間を，$[0<t\le1],....,[9<t\le10]$, $[10<t\le15],[15<t\le\infty]$ の 12 個の区間にわけて推定を行った．

まず基準ハザード値のパラメータ推定の結果に関してだが，parametric model，semi-parametric model の両モデルとも，時間依存性共変量を含まない場合，含む場合のいずれの場合でもパラメータ推定値は不偏性を持つことが確認された．

次に，求められた基準ハザード値のパラメータから実際に基準ハザード値を求めて両モデル間で比較した．その結果，parametric model はサンプル数に関わらずかなり高い精度で推定できていることが確認できた．しかし semi-parametric model の場合は，サンプル数が少ないと推定結果のばらつきが大きく必ずしも推定精度が高いとは言えない結果になった．この原因として，本研究では生存時間の分布を 12 個の区間に区切って推定を行ったが，サンプル数が小さいと各区間に該当するサンプル数が十分でなく，該当するサンプル数が小さい区間では推定精度が低くなることが考えられる．また，基準ハザード値がサンプルの分布に過度に適合する結果，オーバーフィッティングが発生している傾向が見られた．例として，サンプル数が 100 の場合の各ケースの生存時間とハザード値の関係を図 1 に示す．図 1 より，サンプルはワイブル分布に従っているため，全体的には t が大きいほどケース数が少なくなっているが，個々のケースは乱数に基づくため t の値によってケース数がばらついている．基準ハザード値の推定結果はケース数のばらつきに沿った形で推定されており，ワイブル分布からは逸脱していることが分かる．

このように semi-parametric model で推定を行う場合は，推定区間の設定が大きな問題となってくる．しかしサンプル数が十分に大きい場合は，semi-parametric model の基準ハザード値についても時間依存性共変量を含まない場合，含む場合とも一見すると十分に推定精度が高いように見受けられた．サンプル数 1000 の結果を図 2 に示す．

ただし，図 2 では各サンプル数ごとに 10 回ずつ推定し，その結果の平均を示しており，両モデルの推定精度の比較を行うためには，各サンプルごとに生じているばらつきを考慮する必要がある．そこで parametric model と semi-parametric model の各基準ハザード値の MSE を求めて比較することにより，さらに詳しく両モデル間の推定精度を比較した．その結果，時間依存性共変量を含まない場合，含む場合とも semi-parametric

表 1　基準ハザード値の MSE の比較

t	parametric	semi -parametric
0～1	9.40E-07	3.51E-06
1～2	2.82E-06	1.59E-05
2～3	5.14E-06	2.52E-05
3～4	8.26E-06	5.93E-05
4～5	1.21E-05	1.52E-04
5～6	1.65E-05	1.58E-04
6～7	2.16E-05	5.92E-04
7～8	2.71E-05	9.21E-04
8～9	3.31E-05	5.79E-04
9～10	3.95E-05	1.59E-03
10～15	6.15E-05	3.67E-04

model の方が parametric model に比べて MSE の値が大きく，推定値にばらつきがあり推定精度が劣ることが確認された．サンプル数 1000 の場合の結果を表 1 に示す．

（3）基準ハザード値の推定結果の信頼区間

次に基準ハザード値の信頼区間を求めることにより，両モデルの推定結果の有効性を検証した．

今回の推定法では，基準ハザード値を直接未知パラメータとして推定しておらず，基準ハザード値の標準偏差は直接推定されず，基準ハザード関数を規定するパラメータに関する標準偏差が推定されている．本研究では，以下の式を用いて 95%有意水準での基準ハザード値の推定値の信頼区間を求めた．

$$\Pr\left(\hat{h}_0(t) - t_{95}\sqrt{Var[\hat{h}_0(t)]} \le h_0(t) \le \hat{h}_0(t) + t_{95}\sqrt{Var[\hat{h}_0(t)]}\right) \quad (10)$$

ここで，$\hat{h}_0(t)$：基準ハザード値の推定値，t_{95}：t 分布の 95 パーセンタイル値，$Var[\hat{h}_0(t)]$：$\hat{h}_0(t)$ の漸近分散共分散行列である．このとき，$Var[\hat{h}_0(t)]$ はデルタ法を用いることにより，以下のように求められる．

$$Var[\hat{h}_0(t)] = [\partial h_0(t)/\partial \hat{q}]' V[\partial h_0(t)/\partial \hat{q}] \quad (11)$$

ここで，$h_0(t)$：基準ハザード関数，$\hat{q}$：基準ハザード関数のパラメータ推定値，V：パラメータ推定値 $\hat{q}$ の漸近分散共分散行列である．

信頼区間を算出した結果，parametric model, semi-parametric model とも，時間依存性共変量を含む場合，含まない場合に関わらず，サンプル数の増加に伴い推定値の推定精度が高くなっていることが示された．

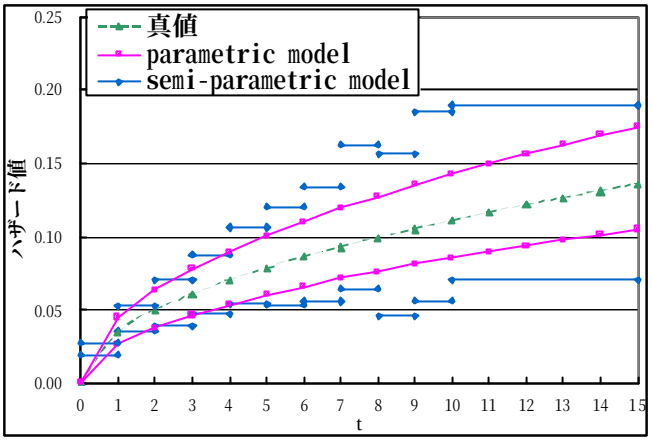

図 3　基準ハザード値の信頼区間

しかしながら，サンプル数に関わらず semi-parametric model は parametric model に比べると常に推定値の信頼度が低いという結果が得られた．サンプル数 1000 での結果を図 3 に示す．

5. おわりに

シミュレーション分析の結果，Prentice and Gloeckler が提案した ordered-response model による semi-parametric model の推定精度を parametric model と比べると，時間依存性を持たない共変量の推定精度は劣らないものの，時間依存性共変量及び基準ハザードの推定精度は劣ること，また基準ハザードに関してはオーバーフィッティングの危険性があることが示された．また推定に要する時間でも，サンプル数や共変量が多くなると parametric model に対して semi-parametric model は計算に要する時間が飛躍的に増大した．これも semi-parametric model の適用に際しての実用的な課題である．

今後の課題としては，ordered-response model における区間分割の最適化法に関する分析や，分布形の仮定に過誤がある場合の parametric model との推定精度の比較分析が考えられる．

参考文献

1) Prentice, R. and Gloeckler, L.: Regression analysis of grouped survival data with application to breast cancer data, *Biometrics*, Vol. 34, pp. 57-67, 1978.
2) Meyer, B.D.: Semiparametric estimation of hazard models, Research Report, Northwestern University, 1995.
3) Bhat, C.R.: Duration modeling, Handbook of Transport Modelling, Hensher, D.A. and Button, K.J. (eds.), Pergamon, pp. 91-111, 2000.