

تحلیل پارامتری داده‌های بقای سانسور فاصله‌ای با پارامتر شکل غیر ثابت

احمدرضا باغستانی، ابراهیم حاجی‌زاده*

گروه آمار زیستی، دانشگاه تربیت مدرس، تهران، ایران

تاریخ دریافت: ۸۸/۹/۳۰

تاریخ پذیرش: ۸۸/۱۱/۱۴

چکیده

مقدمه: در اکثر مطالعات پارامتری تحلیل بقا از مدل وایبول به دلیل انعطاف‌پذیر بودن استفاده می‌شود. در بیشتر موارد برای مدل وایبول فرض می‌شود که پارامتر شکل ثابت است، در حالیکه که این فرض برای برخی از داده‌ها نادرست است.

هدف: در این مطالعه حالت سانسور فاصله‌ای برای توزیع وایبول، در حالتی که پارامتر شکل ثابت نیست ارائه می‌شود.

روش بررسی: با انجام یک مطالعه شبیه‌سازی و به کارگیری روش مونت کارلویی، مزیت‌های آن نسبت به حالتی که پارامتر شکل ثابت باشد ارائه می‌شود. در پایان با یک مثال کاربردی نحوه محاسبه برآوردها نشان داده می‌شود.

نتایج: برآورد ضرایب مدل، مقدار DIC و مقدار نسبت خطر در مدل با پارامتر شکل ثابت و غیر ثابت محاسبه و در ادامه مقایسه گردید. نتایج نشان داد در صورتی که پارامتر شکل توزیع ثابت نباشد، برآورد ضرایب متغیرهای کمکی دارای اریبی کمتر، برازش مناسب‌تر و نتایج حاصل با حالتی که پارامتر شکل ثابت است متفاوت می‌باشد.

نتیجه‌گیری: بر اساس نتایج حاصل بهتر است مدل اشباع شده به گونه‌ای انتخاب شود که پارامتر شکل ثابت نباشد و فرض ثابت بودن پارامتر شکل مورد آزمون قرار گیرد و مدل مناسب‌تر با استفاده از ملاک‌های برازش مدل انتخاب شود.

واژه‌های کلیدی: سانسور فاصله‌ای، رگرسیون وایبول، پارامتر شکل غیر ثابت، معیار اطلاع آکائیک

مقدمه

یکی از مدل‌های آماری که در دهه‌های اخیر بسیار مورد توجه قرار گرفته و کاربرد بسیاری در مطالعات پزشکی، مهندسی و سایر علوم دارد، تحلیل بقا است. تحلیل بقا شامل کلیه روشهای آماری برای تحلیل داده‌های بقا است. ویژگی اصلی داده‌های بقا در مقایسه با دیگر داده‌های آماری، وجود داده‌های سانسور می‌باشد. یکی از انواع پیچیده سانسور حالتی است که زمان دقیق رخداد پیشامد مشخص نیست و فقط این اطلاع موجود است که پیشامد در یک بازه زمانی اتفاق افتاده است که به آن سانسور فاصله‌ای می‌گویند.^(۱)

در تحلیل داده‌های بقا به طور کلی از دو مدل رگرسیونی استفاده، مدل مخاطرات متناسب کاکس به عنوان یک مدل نیمه پارامتری و مدل‌های زمان شکست شتابنده به عنوان مدل پارامتری. اگر چه رگرسیون کاکس کاربردی‌ترین مدل در تحلیل بقا است، ولیکن مدل‌های پارامتری در برخی شرایط می‌توانند مناسب‌تر باشند.^(۲و۳) یکی از شرایط انعطاف‌پذیر بودن این مدل‌ها این است که پارامتر شکل و مکان هر دو می‌توانند تابعی از متغیرهای کمکی باشند. بدست آوردن توزیع پارامتری زمان بقا از اهمیت ویژه‌ای برخوردار است، به طوری که می‌توان احتمال بقا و احتمال خطر را در تمام زمان‌ها محاسبه نمود و در برآورد پارامترها بر خلاف مدل کاکس، از روش حداکثر درست‌نمایی استفاده کرد. همچنین نسبت به رگرسیون کاکس پارامترهای کمتری دارد و مدل ساده‌تری نتیجه می‌دهد. اکثر توزیع‌های پارامتری مثل وایبول و لگ‌نرمال دارای دو پارامتر شکل و پارامتر مکان هستند. در حالت استاندارد فرض می‌شود که پارامتر شکل ثابت و پارامتر مکان تابعی از متغیرهای کمکی است. حالتی که پارامتر شکل ثابت است، اولین بار توسط فارول^(۴) ارائه شد. برای برخی از داده‌ها، این فرض که پارامتر شکل ثابت است، مناسب نیست. برای مثال، در بعضی از مطالعات، پارامتر شکل توزیع وایبول را ثابت فرض نکرده‌اند.^(۵-۷)

در این مقاله برآورد توزیع وایبول در حالت سانسور فاصله‌ای، هنگامی که هر دو پارامتر ثابت نیستند ارائه شده و با استفاده از شبیه‌سازی مزایای آن نسبت به حالتی که پارامتر شکل ثابت است مورد بحث قرار می‌گیرد و در ادامه برآورد پارامترها برای یک مثال کاربردی ارائه می‌شود.

مواد و روشها

توزیع وایبول

این توزیع دارای دو پارامتر (α) شکل و مکان (λ) می‌باشد. تابع چگالی توزیع وایبول با پارامترهای (α, λ) به صورت زیر بیان می‌شود:

$$f(t | \alpha, \lambda) = \alpha \times \lambda \times t^{\alpha-1} \times \exp(-\lambda \times t^\alpha); t > 0, \alpha > 0, \lambda > 0$$

همچنین توابع خطر $(h(t))$ و بقا $(s(t))$ برای این توزیع به صورت زیر است:

$$s(t | \alpha, \lambda) = \exp(-\lambda \times t^\alpha)$$

$$h(t | \alpha, \lambda) = \alpha \times \lambda \times t^{\alpha-1}$$

برای این توزیع اگر $\alpha > 1$ باشد خطر افزایشی، اگر $\alpha < 1$ باشد خطر کاهشی و اگر $\alpha = 1$ باشد خطر ثابت است. در حالت اخیر توزیع وایبول تبدیل به توزیع نمایی می‌شود. توزیع وایبول در تمام حالات که تابع خطر یکنوا یا ثابت است به کار می‌رود. از طرفی با تغییر پارامترهای این توزیع، چگالی وایبول می‌تواند برای زمان‌های چوله به راست و زمان‌های چوله به چپ قابل استفاده باشد. یکی از خصوصیات منحصر بفرد برای مدل رگرسیون وایبول این است که اگر خصوصیات زمان شکست شتابنده^۱ برقرار باشد فرض مخاطرات متناسب^۲ نیز برقرار است و برعکس. البته این حالت هنگامی برقرار است که پارامتر شکل ثابت باشد. با توجه به خواص گفته شده دیده می‌شود که توزیع وایبول انعطاف‌پذیری زیادی دارد و به همین دلیل در تحلیل بقا بصورت متداول کاربرد دارد. در تحلیل پارامتری برای استنباط در مورد متغیرهای کمکی و برازش دقیق‌تر، فرض می‌شود پارامتر مکان با متغیرهای کمکی مرتبط است و پارامتر شکل به متغیرهای کمکی وابسته نیست و ثابت در نظر گرفته می‌شود. برای این منظور، پارامتر مکان را به صورت زیر با متغیرهای کمکی مرتبط می‌کنند.^(۱)

$$\lambda = \exp(\beta_0 + \sum_{j=1}^p \beta_j * X_j)$$

در عین حال، این فرض که پارامتر شکل توزیع تحت تأثیر متغیرهای کمکی نیست برای برخی از داده‌ها، نامناسب است. لذا در این حالت پارامتر شکل را بصورت زیر با متغیرهای کمکی مرتبط می‌کنند.

در این مقاله برآورد پارامترها برای حالتی که هر دو پارامتر توزیع ثابت نیستند ارائه می‌شود.

$$\alpha = \exp(\alpha_0 + \sum_{j=1}^p \alpha_j X_j)$$

مدل رگرسیونی وایبول با سانسور فاصله‌ای

فرض کنید که در یک مطالعه n فرد وجود دارد که زمان بقا برای فرد i ام در فاصله $[t_L^i, t_R^i]$ است. در این صورت تابع درست‌نمایی به شکل زیر ارائه می‌شود:

$$L \propto \int_{t_L^i}^{t_R^i} f(t_i) = \prod_{i \in n} (F(t_i^R) - F(t_i^L)) = \prod_{i \in n} (s(t_i^L) - s(t_i^R))$$

حال اگر فرض شود زمان بقا دارای توزیع وایبول با پارامترهای α و λ است، آنگاه تابع درست‌نمایی بصورت زیر خواهد بود: #:

$$L(\alpha, \lambda | n, t) \propto \prod_{i \in n} (\exp(-\lambda \times t_L^{\alpha}) - \exp(-\lambda \times t_R^{\alpha}))$$

برای وارد کردن متغیرهای کمکی در مدل، پارامتر مکان به صورت زیر در نظر گرفته می‌شود:

$$\lambda = \exp(\beta_0 + \sum_{j=1}^p \beta_j * X_j)$$

که در آن $X = (X_1, X_2, \dots, X_p)$ بردار متغیرهای کمکی و β ها ضرایب رگرسیونی می‌باشد. پارامتر شکل (α) ، در این حالت ثابت در نظر گرفته می‌شود. در این حالت شرایط فرض مخاطرات متناسب برقرار است.

برای درک بهتر، در حالتی که یک متغیر کمکی دو حالتی وجود داشته باشد، پارامتر مکان به صورت زیر خواهد بود:

¹- Accelerated Failure Time(AFT)

²- Proportional Hazard(PH)

و نسبت خطر (HR) بصورت

$$\text{Hazard ratio}(X = 1 \text{ vs } X = 0) = HR = \frac{h(t | x_1 = 1)}{h(t | x_1 = 0)} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

$$\lambda = \exp(\beta_0 + \beta_1 X)$$

برای درک بهتر، در حالتی که یک متغیر کمکی دو حالتی وجود داشته باشد، پارامتر مکان به صورت زیر خواهد بود:

و نسبت خطر (HR) بصورت

$$\text{Hazard ratio}(X = 1 \text{ vs } X = 0) = HR = \frac{h(t | x_1 = 1)}{h(t | x_1 = 0)} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

که نشان دهنده این مطلب است که در حالت ثابت بودن α ، شرایط فرض مخاطرات متناسب مدل وایبول برقرار است. همانطور که ملاحظه می‌شود، در این حالت تفسیر نسبت خطر بسیار آسان می‌باشد.

این فرض که شکل پارامتر توزیع وایبول تحت تاثیر متغیرهای کمکی نیست و ثابت می‌باشد برای بعضی از داده‌ها مناسب نیست. در حالتی که پارامتر شکل در توزیع وایبول ثابت نباشد، داریم:

(۲)

$$\alpha = \exp(\alpha_0 + \sum_{j=1}^p \alpha_j X_j)$$

که اگر یک متغیر کمکی دو حالتی در اختیار باشد، نسبت خطر بصورت زیر خواهد بود:

$$\text{Hazard ratio}(X = 1 \text{ vs } X = 0) = HR = \frac{h(t | x = 1)}{h(t | x = 0)}$$

$$\frac{\exp(\alpha_0 + \alpha_1) \times t^{\exp(\alpha_0 + \alpha_1) - 1} \times \exp(\beta_0 + \beta_1)}{\exp(\alpha_0) \times t^{\exp(\alpha_0) - 1} \times \exp(\beta_0)} = \exp(\alpha_1 + \beta_1) \times t^{\exp(\alpha_0 + \alpha_1) - \exp(\alpha_0)}$$

همان طور که دیده می‌شود فرض متناسب بودن مخاطرات برقرار نیست و مدل، وابسته به زمان است.

نتایج و بحث

شبیه‌سازی

در شبیه‌سازی که به منظور بررسی نحوه برآورد پارامترها صورت گرفت، ۲۰۰۰ نمونه با ۱۰۰ بار تکرار به روش زیر تولید شده است.

برای تولید زمان‌های بقا فرض می‌شود که پارامتر مکان و شکل توزیع متأثر از یک متغیر کمکی به صورت زیر

است:

$$\lambda = \exp(\beta_0 + \beta_1 X)$$

که در آن فرض می‌شود X دارای توزیع نرمال با میانگین ۰ و واریانس ۱ است و برای ضرایب، دو حالت زیر در نظر

$$\alpha = \exp(\alpha_0 + \alpha_1 X)$$

گرفته می‌شود:

β_1	β_0	α_1	α_0	حالت
۰/۱	۰/۲	۱	۱	اول
۰/۱	۰/۲	۰	۱	دوم

همانطور که ملاحظه می‌شود در حالت اول، پارامتر شکل ثابت نمی‌باشد ($\alpha \neq 0$) و در حالت دوم پارامتر شکل ثابت است ($\alpha_1 = 0$) پس از تولید مقادیر X برای هر نمونه، برای هر حالت مقادیر λ و α بدست می‌آید و متناظر با λ و α عددی تصادفی از توزیع وایبول با پارامتر λ و α به‌عنوان زمان بقا به دست می‌آید. این عمل آن قدر تکرار می‌شود تا یک نمونه به حجم ۲۰۰۰ از زمان‌های بقا حاصل شود. برای ایجاد سانسور فاصله‌ای، زمان‌های $t'_i = 0, 0.01, 0.02, 0.03, \dots$ در نظر گرفته می‌شود و برای کران پایین سانسور فاصله‌ای، زمان t'_i ای که نزدیکترین مقدار به زمان بقای تولید شده و کوچکتر از آن باشد انتخاب و یک عدد تصادفی از توزیع برنولی با احتمال موفقیت 0.2 تولید می‌شود که اگر عدد مشاهده شده ۱ باشد برای کران پایین مقدار t'_i درج می‌شود، در غیر این صورت t'_i برابر با 0.1 - در نظر گرفته می‌شود. این عمل آنقدر تکرار می‌شود تا عدد تصادفی که از توزیع برنولی تولید می‌شود برابر با ۱ گردد. این عمل برای ایجاد کران بالای سانسور فاصله‌ای انجام می‌شود و فقط زمان ای که نزدیکترین مقدار به زمان بقا تولید شده و بزرگتر از آن باشد انتخاب می‌شود و اگر عدد مشاهده شده با توزیع برنولی صفر باشد مقدار t'_i برابر با 0.1 + در نظر گرفته می‌شود. لازم به ذکر است در کران پایین اگر زمان منفی بدست آمد مقدار کران پایین صفر در نظر گرفته می‌شود (سانسور چپ). پس از تولید داده‌ها، برآورد پارامترهای توزیع وایبول برای دو مدل زیر به دست می‌آید:

مدل ۱: پارامتر مکان تحت تأثیر متغیرهای کمکی است و پارامتر شکل ثابت است.

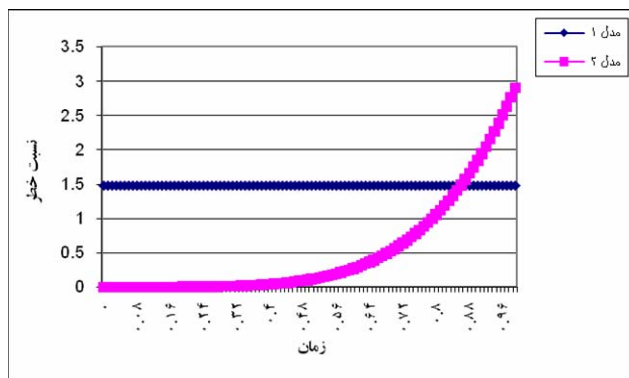
مدل ۲: هر دو پارامتر مکان و شکل، متأثر از متغیرهای کمکی هستند.

برای تولید داده‌ها و برآورد پارامترها برای هر دو مدل مذکور، برنامه‌ای در نرم افزار R نوشته شده است. در این مقاله، برای مقایسه دو مدل مذکور، از ملاک اطلاع آکائیکه^۱ (AIC) استفاده شده است. این معیار هر دو جنبه برازش و پیچیدگی مدل را در برداشته است و بر اساس این معیار، هر مدلی که دارای کمترین AIC باشد، به عنوان بهترین مدل انتخاب می‌شود و از طرفی چون برآوردهای پارامتر بقا، برآوردهای سازگار هستند (با افزایش حجم نمونه، برآوردها به پارامتر واقعی نزدیک می‌شود) لذا در این مقاله با توجه به اینکه اندازه نمونه در نظر گرفته شده ۲۰۰۰ است پس می‌توان علاوه بر معیار AIC، مقدار برآوردهای حاصل از این دو مدل را با مقدار واقعی مقایسه کرد. نتایج حاصل از شبیه‌سازی در جداول و شکل‌های ۱-۲ آمده است. مقادیر برآورد پارامترها، مقادیر AIC و نسبت خطر در دو مدل در جداول ۱-۲ آمده است و همچنین در شکل‌های ۱-۲، نمودار نسبت خطر $X=1$ به $X=0$ آمده است.

جدول ۱ - برآورد پارامترها با ضرایب $\alpha_0 = 1$ و $\alpha_1 = 1$ و $\beta_0 = 0.2$ و $\beta_1 = 0.1$ (حالت اول)

نسبت خطر	AIC	β_1	β_0	α_1	α_0	مدل
۱/۴۷۵	۱۵۲۳۲	۰/۳۸۹	-۰/۱۰۸	۰	-۰/۲۰۷	مدل ۱
4.747×10^{-3}	۱۰۸۴۱	۰/۱۰۱	۰/۲۰۳	۱/۰۰۹	۱/۰۰۲	مدل ۲

با توجه به جدول ۱، مقدار AIC برای مدل ۲ کمتر از مدل ۱ می‌باشد، بنابراین مدل ۲ مناسب‌تر است. برآورد ضرایب مدل ۱، دارای اریبی بیشتری هستند، به عنوان مثال برای $\beta_1 = 0.1$ ، این مقدار در مدل ۱، 0.389 برآورد شده است.



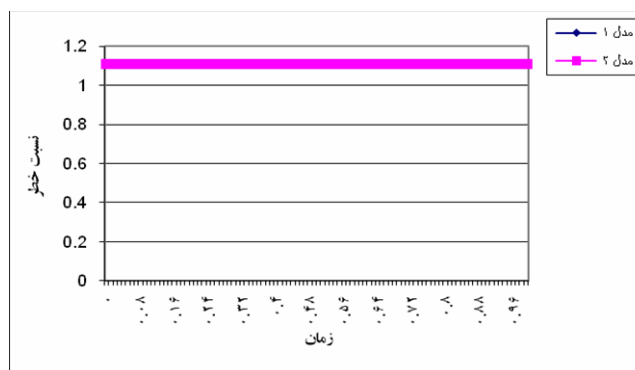
شکل ۱- نسبت خطر بر حسب زمان برای حالت اول

در شکل ۱ ملاحظه می‌شود که در تمام زمان‌ها، برای مدل ۱ نسبت خطر ثابت است به طوری که این مقدار در حالتی که $X = 1$ است $1/475$ برابر حالتی است که $X = 0$ می‌باشد ($HR = 1/475$). این نسبت در مدل ۲ وابسته به زمان است ($HR = 3/0.33 \times t^{4/47}$) به طوری که در زمان‌های کمتر از $t = 0.86$ نسبت خطر کمتر از مدل ۱ است و برای زمان‌های بیشتر از $t = 0.86$ نسبت خطر مدل ۲ بیشتر است. برای این مدل، تنها در زمان $t = 0.86$ دو نسبت با هم برابرند.

جدول ۲- برآورد پارامترها با ضرایب $\alpha_0 = 1$ و $\alpha_1 = 0$ و $\beta_0 = 0.2$ و $\beta_1 = 0.1$ (حالت اول)

نسبت خطر	AIC	β_1	β_0	α_1	α_0	مدل
$1/1.09$	11149	0.103	0.199	0	$1/0.06$	مدل ۱
$1/1.07 \times t^{-0.0005}$	11150	0.104	0.198	-0.002	$1/0.06$	مدل ۲

در جدول و شکل ۲ به دلیل اینکه مقدار واقعی ضرایب α_1 برابر با صفر است (پارامتر شکل ثابت است)، مقادیر ضرایب برآورد شده در هر دو مدل، با مقادیر واقعی تفاوتی ندارند و نسبت خطر در دو مدل یکسان است. با توجه به این جدول، در مدل ۲ به علت اینکه یک پارامتر (α_1) بیشتر وجود دارد، مدل پیچیده‌تر است و بنابراین DIC آن از مدل ۱ بیشتر است و در نتیجه مدل ۱، مدل بهتری است.



شکل ۲- نسبت خطر بر حسب زمان برای حالت دوم

لازم به ذکر است، به دلیل اینکه متغیر X کمی در نظر گرفته شده است، نمودارهای ۱-۲ فقط برای دو نقطه $X=1$ و $X=0$ ، نسبت خطر گزارش شده است.

مثال کاربردی

در این مقاله از داده‌های مربوط به مبتلایان سرطان معده که در بخش گوارش بیمارستان طالقانی تهران تحت درمان بوده‌اند، استفاده شده است. نوع مطالعه طولی تاریخی و از طریق مراجعه به پرونده بیماران مبتلا به سرطان معده به تاریخ بهمن ۱۳۸۲ لغایت دی ۱۳۸۶ که در بخش گوارش تحت درمان بودند، انجام شد. این داده‌ها مربوط به ۲۱۴ بیمار می‌باشد که از طریق تماس تلفنی، اطلاعات مربوط به بقای آنها جمع‌آوری شده است. متغیرهای کمکی موجود در این داده‌ها، اندازه تومور (کمتر از ۳۵ میلی‌متر و بیشتر از ۳۵ میلی‌متر) و سن هنگام تشخیص بیماری می‌باشد.

جداول ۳ و ۴ نتایج برآورد مدل وایبل برای هر دو مدل ۱ و ۲ را نشان می‌دهد.

جدول ۳ - برآوردهای میانگین و انحراف معیار برای مدل ۱

پارامتر	میانگین	انحراف معیار
مقدار ثابت *	-۷/۷۲	۰/۸۵
اندازه تومور *	۰/۸۱	۰/۲۸
سن هنگام تشخیص بیماری *	۰/۰۲	۰/۰۱
شکل * (λ)	۱/۲۰	۰/۱۱

جدول ۴ - برآوردهای میانگین و انحراف معیار برای مدل ۲

پارامتر	متغیر	میانگین	انحراف معیار
مکان	مقدار ثابت *	-۹/۶۲	۲/۲۸
مکان	اندازه تومور *	۲/۷۶	۰/۹۷
مکان	سن هنگام تشخیص بیماری	۰/۰۰۰	۰/۰۳
شکل	مقدار ثابت	۰/۵۳	۰/۴۲
شکل	اندازه تومور x	-۰/۴۲	۰/۱۵
شکل	سن هنگام تشخیص بیماری	۰/۰۰۵	۰/۰۰۴

* معنی‌دار در سطح ۰/۰۵

همانطور که مشاهده می‌شود برآورد ضرایب پارامترهای بدست آمده مدل ۱ و ۲ با هم متفاوتند. درعین حال، نتایج استنباط با هم تفاوت دارند به طوری که در مدل ۱، اندازه تومور و سن هنگام تشخیص بیماری به عنوان عوامل تاثیرگذار در زمان بقا می‌باشند ولی در مدل ۲، تنها اندازه تومور بر روی زمان بقا تاثیر دارد. مقدار AIC برای مدل ۱، ۵۹۲ و برای مدل ۲، ۵۸۷ بدست آمد که نشان‌دهنده این مطلب است که مدل ۲ برازش بهتری به داده‌ها داشته

است و فرض ثابت بودن پارامتر شکل رد می‌شود. با توجه به نتایج حاصل برای این داده‌ها، پیشنهاد می‌شود که مدل نهایی، مدلی باشد که در آن پارامتر شکل ثابت نیست (مدل ۲).

نتیجه‌گیری

در سالهای اخیر تحلیل داده‌های بقا به دلیل وجود سانسور، بیشتر مورد توجه قرار گرفته است. در اکثر مطالعات بقا برای توزیع‌های پارامتری مثل وایبول و لگ نرمال فرض می‌شود که پارامتر شکل در طول زمان ثابت است و پارامتر مکان یک تابعی از متغیرهای کمکی است. این فرض برای بعضی از مطالعات درست نیست و در این حالت اگر پارامتر شکل تابعی از متغیرهای کمکی در نظر گرفته شود، برازش بهتری به داده‌ها صورت می‌گیرد. نتایج شبیه‌سازی نشان داد، در صورتی که پارامتر شکل توزیع ثابت نباشد، ممکن است برآوردهای ضرایب متغیرهای کمکی، دارای اریبی کمتر، تفسیر شاخص‌های مهمی در بقا همچون نسبت خطر برای زمان‌های مختلف در آن دقیق‌تر، برازش مدل مناسب‌تر و نتایج استنباط متفاوت است. بر اساس نتایج حاصل، پیشنهاد می‌شود که مدل اشباع شده به گونه‌ای انتخاب شود که پارامتر شکل ثابت نباشد و فرض ثابت بودن پارامتر شکل مورد آزمون قرار گیرد و مدل مناسب‌تر با استفاده از ملاک‌های برازش مدل انتخاب شود.

References:

1. Klein, J.P., and Moeschberger, M.L., *Survival Analysis Techniques for Censored and Truncated Data*, Springer-Verlag, New York (1997).
2. Lawless, J.F., *Parametric Models in Survival Analysis*, In *Encyclopedia of Biostatistics*, Armitage, P., Colton, T., Wiley, New York (1998).
3. Lindsey, J.C., *Lifetime Data Anal.*, **4**, 329 (1998).
4. Farewell, V.T., *Biometr.*, **38**, 1041 (1982).
5. Hirose, H., *IEEE Transactions on Reliability*, **42**, 650 (1993).
6. Meeter, C.A., and Meeker, W.Q., *Technometr.*, **36**, 71 (1994).
7. Hutton, J.L., and Solomon, P.J., *J. R. Statist. Soc. B*, **59**, 125 (1997).