استفاده از مدل بی‌ثبات نرخ خطر متناسب در تحلیل داده‌های واقعی

محمد ملانوری[۱]،حبیب نادری [۲] ، حامد احمدزاده [۳] ، سلمان ایزدخواه [۴]

تاریخ دریافت: ۹۹/۱/۱۳
تاریخ پذیرش: ۹۹/۱۱/۱

چکیده:

جمعیت‌های زیادی در آنالیز بقا اغلب با مشکل ناهمگنی مواجه هستند. افراد در نحوه برخورد با علت مرگ، واکنش به معالجه و تحت تأثیر عوامل خطر قرار گرفتن، انعطاف‌پذیر هستند. نادیده گرفتن این ناهمگنی می‌تواند باعث به دست آمدن نتایج نادرست شود. برای برطرف کردن این مشکلات، مدل بی‌ثبات نرخ خطر متناسب را معرفی می‌کنیم. در این مقاله، ضمن معرفی مدل بی‌ثبات نرخ خطر متناسب، ویژگی‌های آن را موردمطالعه قرار می‌دهیم. برازش مدل بی‌ثبات به داده‌های سانسور شده از راست را در حضور متغیرهای توضیحی (متغیرهای قابل‌مشاهده) بررسی می‌کنیم و در قالب یک مثال کاربردی برای برازش مدل بی‌ثبات به داده‌ها با در نظر گرفتن توزیع پایه وایبول و نمایی در توابع درستنمایی، از آن‌ها برای برآورد پارامترهای مدل استفاده کرده و با معیارهای مختلف به مقایسه مناسبت مدل‌ها می‌پردازیم.

واژه‌های کلیدی: برآورد ماکسیمم درستنمایی، توزیع پایه، تابع بقای غیرشرطی، مدل بی‌ثبات

۱ مقدمه

یک مطالعه تحلیل بقا را در نظر بگیرید که زیرگروه‌های جامعه دارای یک متغیر تصادفی مشترک مشاهده نشده باشند. چنین متغیرهایی می‌توانند یک عامل ژنتیکی، اثر محیطی در سال‌های اولیه زندگی فرزندان یک خانواده یا اثر الگوی زندگی در زوج‌ها باشد. به این متغیرهای مشترک، خطرهای مشترک و به مدل‌هایی که دربرگیرنده این ناهمگنی تصادفی هستند، مدل‌های بی‌ثبات می‌گویند. این مدل‌ها به‌منظور تبیین تغییرات ناشی از عوامل خطر مشاهده نشده به کار می‌روند. مدل‌های بی‌ثبات (مدل اثر تصادفی) قادرند ناهمگنی بین زمان‌های بقا در گروه‌های مختلف و همبستگی ایجادشده بین زمان‌های بقای هر گروه را به‌حساب آورند. مدل‌های بی‌ثبات مدل‌های نسبتاً جدیدی در تحلیل بقا بوده و به‌طور گسترده‌ای در سه دهه‌ی اخیر موردمطالعه قرارگرفته‌اند. مدل‌های بی‌ثبات کلاسیک اولین بار توسط واوپل و همکاران [۱۷]، برای مطالعه برخی ناهمگنی‌هایی که به‌وسیله متغیرهای کمکی قابل‌اندازه‌گیری نیستند، معرفی شدند. کاربرد این مدل‌ها در تحلیل بقا برای ناهمگنی مشاهده نشده (پنهان) بین افراد، با فرض اینکه نرخ خطر افراد متأثر از یک پارامتر مشخص به نام پارامتر شکنندگی (بی‌ثبات) و یک مخاطره پایه در نظر گرفته شود، موردتوجه بسیاری از محققان بوده است. گوپتا و کرمانی [۶] مقایسه‌های تصادفی در مدل‌های بی‌ثبات را از طریق تأثیر مقایسه دو متغیر تصادفی پایه بر روی متغیرهای بی‌ثبات متناظر مطالعه کرده و نتایج مفیدی به دست آوردند. ژو و لی [۱۸] نیز متغیر پایه را مستقل از متغیر بی‌ثبات فرض کردند. در مدل آنان متغیر خروجی (یا متغیر جمعیت) در مدل بی‌ثبات است که با متغیر بی‌ثبات وابستگی دارد. کیاد و همکاران [۱۳] برخی از نتایج ترتیب تصادفی نسبی از دو مدل بی‌ثبات را موردمطالعه قرار دادند. علاوه بر این محققین دیگری در زمینه مدل‌های بی‌ثبات تحقیقاتی را انجام داده‌اند. به‌عنوان‌مثال می‌توان به آلن [۱، ۲، ۳] ، هودگارد [۷، ۸، ۹، ۱۰، ۱۱] ، نلسون [۱۶] و مولر و استویان [۱۵] اشاره کرد.

انتخاب توزیع مناسب برای متغیر بی‌ثبات از دیدگاه بیزی همان انتخاب توزیع پیشین برای پارامتر بی‌ثبات است که نمی‌توان گفت هیچ دیدگاهی برای انتخاب توزیع پیشین وجود ندارد. در عمل بنا به تجربه معمولاً توزیع گاما، وایبول، لگ نرمال و توزیع گوسی وارون را برای متغیر بی‌ثبات انتخاب می‌کنند که دلیل اصلی آن انعطاف‌پذیر بودن این توزیع‌هاست. در واقع می‌توان گفت بیشتر نرم‌افزارها به انتخاب توزیع گاما برای متغیر بی‌ثبات محدود می‌شوند. بنابراین علاقه‌مندیم مدل نرخ خطر متناسب زمانی که عامل بی‌ثبات دارای توزیع گاماست را موردبحث قرار دهیم. در این حالت با

[۱] دانشجوی کارشناسی ارشد دانشگاه سیستان و بلوچستان
[۲] عضو هیئت‌علمی دانشگاه سیستان و بلوچستان h.h.naderi@gmail.com
[۳] عضو هیئت‌علمی دانشگاه سیستان و بلوچستان
[۴] عضو هیئت‌علمی دانشکده علوم پایه و فنی-مهندسی بیجار

انتگرال‌گیری از تابع درستنمایی شرطی نسبت به متغیر بی‌ثبات می‌توان به یک عبارت صریح و ساده برای درستنمایی حاشیه‌ای دست‌یافت. درستنمایی حاشیه‌ای که شامل پارامترهای موردنظر است را ماکسیمم می‌کنیم و برآورد پارامترها و خطای استاندارد آن‌ها را به دست می‌آوریم. در بخش دوم مدل بی‌ثبات نرخ خطر متناسب را معرفی و شاخص‌هایی برای آن را موردمطالعه قرار می‌دهیم. در بخش سوم به برازش مدل بی‌ثبات به داده‌های سانسور شده از راست در حضور متغیرهای توضیحی (قابل‌مشاهده) می‌پردازیم. در بخش چهارم در قالب یک مثال کاربردی مدل بی‌ثبات نرخ خطر متناسب را به داده‌های سانسور شده از راست برازش می‌دهیم. در پایان نتیجه‌گیری را ارائه می‌دهیم.

۲ معرفی مدل بی‌ثبات نرخ خطر متناسب

در این بخش به معرفی مدل بی‌ثبات نرخ خطر متناسب و شرایط معادل با آن می‌پردازیم. قبل از معرفی این مدل لازم است تعریف زیر را بیان کنیم که از طریق آن می‌توانیم از یک توزیع پایه، توزیع دیگری به دست آوریم.

تعریف ۱.۲. برای هر $\alpha > 0$ و تابع توزیع پایه F_0، تابع $F(t|\alpha) = 1 - (1 - F_0(t))^\alpha$ یک تابع توزیع است که یک خانواده از توزیع‌های با نرخ خطر متناسب را تشکیل می‌دهد.

اکنون برای معرفی این مدل فرض کنید متغیر تصادفی نامنفی و پیوسته طول عمر T و متغیر تصادفی نامنفی V، متغیر بی‌ثبات و $F_0(x)$ یک توزیع پایه باشد آنگاه بنا به تعریف (۱.۲) و با توجه به ارتباط بین تابع بقا و تابع نرخ خطر، تابع بقای شرطی شده روی متغیر بی‌ثبات (یعنی به‌شرط $V = v$) برای یک مؤلفه به‌صورت زیر بیان می‌شود:

$$\overline{F}(t \mid v) = (1 - F_0(t))^v = (\overline{F}_0(t)))^v \qquad (1)$$

بنابراین اعضای جمعیت با $v > 1$ در مقایسه اعضای با $v = 1$، دارای تابع بقای نزولی و نرخ خطر صعودی (با توجه به فرمول تابع نرخ خطر، تابع نرخ خطر و تابع بقا رابطه‌ی عکس باهم دارند) و اعضای جمعیت با $v < 1$ در مقایسه با اعضای $v = 1$، دارای تابع بقای صعودی و نرخ خطر نزولی هستند.

با به کار بردن رابطه $\overline{F}(x) = exp\{-H(x)\}$ می‌توان نوشت:

$$\overline{F}(t \mid v) = exp\{-vH_0(t)\}$$

که در آن $H_0(t) = \int_0^t h_0(x)dx = -ln\overline{F}_0(t)$، تابع نرخ خطر تجمعی متناظر با $\overline{F}_0(x)$ می‌باشد. با استفاده از رابطه (۱) مدل بی‌ثبات نرخ خطر متناسب به‌صورت زیر است:

$$h(t \mid v) = -\frac{d}{dt}Ln\overline{F}(t \mid v) = v\left(-\frac{d}{dt}Ln\overline{F}_0(t)\right) = vh_0(t) \qquad (2)$$

که در آن $h_0(t)$ تابع نرخ خطر پایه و مستقل از v است. در واقع اگر T معرف طول عمر یک مؤلفه باشد آنگاه $h(t \mid v)$ معرف نرخ مخاطره در لحظه t برای مؤلفه‌ای با بی‌ثباتی v خواهد بود.

با استفاده از رابطه (۲) تابع بقای شرطی متغیر T به‌شرط V و با توجه به اینکه $\overline{F}_0(t) = exp\left(-v\int_0^t h_0(x)dx\right)$، به‌صورت زیر به دست می‌آید:

$$\overline{F}(t \mid v) = [\overline{F}_0(t)]^v = exp\left(-v\int_0^t h_0(x)dx\right) \qquad (3)$$

که در آن $\overline{F}_0(t)$ تابع بقای پایه است. تابع چگالی احتمال شرطی متناظر با رابطه (۳) به‌صورت زیر است:

$$\begin{aligned} f(t \mid v) = -\frac{\partial\overline{F}(t \mid v)}{\partial t} &= -v(\overline{F}_0(t))^{v-1} \times (-f_0(t)) \qquad (4)\\ &= v\frac{f_0(t)}{\overline{F}_0(t)}.(\overline{F}_0(t))^v \\ &= vh_0(t)exp\left(-v\int_0^t h_0(x)dx\right) \end{aligned}$$

با استفاده از رابطه (۳) تابع توزیع غیرشرطی (جمعیت) متغیر T پس از متوسط گیری نسبت به تأثیر متغیر پنهان V به‌صورت زیر به دست می‌آید:

$$\begin{aligned} \overline{F}(t) &= \int_0^\infty \overline{F}(t \mid v)g(v)dv \qquad (5)\\ &= \int_0^{+\infty} (\overline{F}_0(t))^v g(v)dv \\ &= E[(\overline{F}_0(t))^V]; t > 0 \end{aligned}$$

که در آن $g(.)$ تابع چگالی متغیر تصادفی V است. با توجه به رابطه (۵) تابع چگالی غیرشرطی T به‌صورت زیر به دست می‌آید:

$$\begin{aligned} f(t) &= \int_0^\infty f(t \mid v)g(v)dv \qquad (6)\\ &= \int_0^\infty vh_0(t)exp\left(-v\int_0^t h_0(x)dx\right)g(v)dv \\ &= h_0(t)\int_0^\infty vexp\left(-v\int_0^t h_0(x)dx\right)g(v)dv \end{aligned}$$

۳ مدل بی‌ثبات نرخ خطر متناسب در برآورد پارامترهای مدل

قبل از شروع مباحث اصلی نیاز به بیان مفاهیمی مقدماتی است که در ابتدا به آن اشاره می‌کنیم.

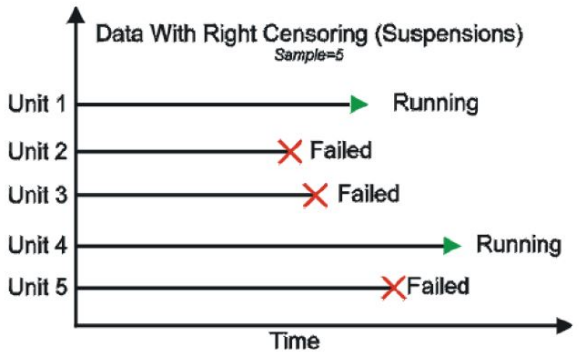
۱.۳ داده‌های سانسور شده از راست

رایج‌ترین نوع سانسور، متعلق به داده‌های سانسور شده از راست [۵] می‌باشد. در داده‌های مربوط به طول عمر، این مجموعه داده‌ها شامل واحدهایی هستند که دچار شکست نمی‌شوند. اگر $C_1, C_2, ..., C_n$ را متغیرهای تصادفی مستقل و هم‌توزیع با تابع توزیع G در نظر بگیریم. در اینجا C_i زمان سانسور با توجه به زمان شکست t_i (زمان شکست i امین واحد) می‌باشد. در این حالت سانسوری، داده‌ها به‌صورت زیر نمایش داده می‌شود

$$(Y_1, \delta_1), (Y_2, \delta_2), ..., (Y_n, \delta_n)$$

که در آن $Y_i = min(T_i, C_i)$ و $\delta_i = I(T_i \leq C_i)$ می‌باشد. $\delta_1, ..., \delta_n$ حاوی اطلاعات مربوط به سانسور شدن زمان‌های شکست واحدهای موردنظر می‌باشند. یعنی اگر $\delta_i = 0$ شد، مشاهده موردنظر (Y_i) مربوط به زمان شکست سانسور شده می‌باشد و اگر $\delta_i = 1$ شد، (Y_i) برابر زمان دقیق شکست است.

مثال ۱.۳. اگر ۵ واحد را موردبررسی قرار دهیم و فقط ۳ تا از آن‌ها تا پایان بررسی به زمان شکست برسند. با توجه به ۲ واحدی که به زمان شکستشان نرسیده‌اند، ما داده‌های سانسور شده از راست خواهیم داشت. واژه "از راست سانسور شده" بدین مفهوم است که زمان رخداد پیشامد موردعلاقه (زمان شکست) در سمت راست نقاط داده‌ها (برای ۳ واحد مشاهده) قرار دارد. به‌عبارت‌دیگر اگر واحدها به کارکرد خود ادامه دهند، زمان شکست این دو واحد، زمانی بعد از سایر واحدهای مشاهده‌شده و یا سمت راست مقیاس زمانی آن‌ها قرار خواهد گرفت.



شکل ۱: نمودار سانسور از راست

۲.۳ تابع درستنمایی بقا برای داده‌های سانسور شده از راست

در حالت کلی، فرض می‌کنیم که زمان سانسور و زمان بقا از نظر آماری، متغیرهای تصادفی مستقل هستند. برای داده‌های سانسور شده‌ی راست، اطلاعات حقیقی برای مؤلفه‌ی i ام، به ازای $i = 1, 2, ..., n$ شامل زوج (y_i, δ_i) است به‌طوری‌که y_i مینیمم زمان پیشامد t_i و زمان سانسور c_i است. یعنی $y_i = min(t_i, c_i)$ و δ_i نشانگر سانسور است و اگر پیشامد موردنظر اتفاق افتد، مقدار یک را می‌گیرد و در غیر این صورت، مقدار صفر را خواهد گرفت.

$$\delta_i = \begin{cases} 1 & , t_i \leq c_i \\ 0 & , t_i > c_i \end{cases}$$

برای داده‌های سانسور شده، تابع درستنمایی بقا کاملاً از تابع درستنمایی کلاسیک برای داده‌های مستقل بدون سانسور متفاوت است. فرم عبارت درستنمایی با نوع داده‌هایی که در دسترس هستند تعیین می‌شود. به‌طورکلی در اینجا، داده‌هایی که در نظر می‌گیریم یا داده‌های کامل هستند (همه‌ی زمان‌های پیشامد مشاهده‌شده‌اند) یا داده‌های سانسور شده از راست هستند (زمان پیشامد مؤلفه‌ها برای سانسور راست) و یا داده‌های سانسور شده‌ی بازه‌ای هستند (زمان‌های پیشامد معلوم هستند و در یک بازه مشاهده می‌شوند).

برای داده‌های سانسور شده از راست با سانسور تصادفی، سهم درستنمایی زمان پیشامد $(y_i = t_i, \delta_i = 1)$ به‌صورت زیر داده می‌شود:

$$\begin{aligned} &lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} P(y_i - \varepsilon_i < Y_i < y_i + \varepsilon_i, \delta_i = 1) \\ &= lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} P(y_i - \varepsilon_i < Y_i < y_i + \varepsilon_i, T_i \leq C_i) \\ &= lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_{y_i - \varepsilon_i}^{y_i + \varepsilon_i} \int_t^{\infty} dG(t) dF(t) \\ &= lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_{y_i - \varepsilon_i}^{y_i + \varepsilon_i} (1 - G(t) dF(t) = (1 - G(y_i)) f(y_i) \end{aligned}$$

از طرف دیگر برای مشاهدات سانسور شده‌ی راست $(y_i = c_i, \delta_i = 0)$ با سانسور تصادفی سهم درستنمایی به‌صورت زیر داده می‌شود:

$$\begin{aligned} &lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} P(y_i - \varepsilon_i < Y_i < y_i + \varepsilon_i, \delta_i = 0) \\ &= lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} P(y_i - \varepsilon_i < Y_i < y_i + \varepsilon_i, T_i > C_i) \\ &= (1 - F(y_i)) g(y_i) \end{aligned}$$

تحت سانسور تصادفی راست، داده بقا شامل ترکیب زمان‌های پیشامد و مشاهدات سانسور شده‌ی راست است.

بنابراین طبق توضیحات بالا درستنمایی برای یک نمونه به‌اندازه‌ی n به‌صورت زیر به دست می‌آید:

$$L = \prod_{i=1}^{n} [(1 - G(y_i)) f(y_i)]^{\delta_i} [(1 - F(y_i)) g(y_i)]^{1 - \delta_i}$$

اگر فرض کنیم که توزیع زمان‌های سانسور به پارامترهای مرتبط با تابع بقا وابسته نیستند، آنگاه گوییم سانسور حاوی اطلاعات مفید نمی‌باشد. (لیانگ [۱۴]، فلمینگ و هرینگتون [۵]) و عوامل $(g(y_i))^{\delta_i}$ و $(1 - G(y_i))^{\delta_i}$ برای

[5] Right Censoring

استنباط درباره‌ی تابع بقا مفید نیستند و می‌توان آن‌ها را از درستنمایی حذف کرد، لذا داریم:

$$L = \prod_{i=1}^{n} (f(y_i))^{\delta_i} \left(\overline{F}(y_i)\right)^{1-\delta_i} = \prod_{i=1}^{n} (h(y_i))^{\delta_i} \overline{F}(y_i) \quad (۷)$$

در این بخش مدل بی‌ثبات $h(t \mid v) = vh_0(t)$ برای داده‌های سانسور تصادفی از راست در حضور متغیرهای توضیحی (متغیرهای قابل‌مشاهده) بررسی می‌شود. به این منظور، فرض کنید t_i، $i = 1,2,...,n$، متغیرهای پیوسته نشان‌دهنده زمان بقای مؤلفه i، مستقل و هم توزیع با تابع توزیع F و تابع چگالی f باشند و همچنین فرض کنید C_i ، $i = 1,2,...,n$، زمان سانسور مؤلفه i ام باشد. در این صورت، متغیرهای تصادفی قابل‌مشاهده عبارت‌اند از: $Y_i = min(T_i, C_i)$ و $\delta_i = I(T_i \le C_i)$ که در آن $I(.)$ نشانگر وقوع حادثه موردنظر است. هدف، برازش مدل بی‌ثبات پارامترهای نرخ خطر متناسب به نمونه تصادفی یادشده است. اکنون فرض کنید مدل بی‌ثبات $h(t \mid v) = vh_0(t)$ به‌صورت زیر در نظر گرفته شود:

$$h(t \mid \underline{\mathbf{x}}, v) = vg(\underline{\mathbf{x}}, \underline{\beta})h_0(t;\theta) \quad (۸)$$

که در آن $\underline{\mathbf{x}}$ مقدار مشاهده‌شده بردار متغیرهای توضیحی $\underline{\mathbf{x}}_{p\times 1}$ و $\beta_{p\times 1}$ بردار پارامترهای رگرسیونی، $g(.)$ تابعی نامنفی از $\square$ و β و همچنین $\underline{\theta}$ بردار پارامترهای توزیع پایه فرض شده است. با فرض $g(\underline{\mathbf{x}}, \underline{\beta}) = exp(\underline{\beta}'\underline{\mathbf{x}})$ مدل (۸) به‌صورت زیر درمی‌آید:

$$h(t \mid \underline{\mathbf{x}}, v) = vexp(\underline{\beta}'\underline{\mathbf{x}})h_0(t;\theta) \quad (۹)$$

در اینجا می‌خواهیم پارامترهای $\underline{\theta}$ و پارامترهای توزیع $g(.)$ (پارامترهای توزیع بی‌ثبات) و نیز بردار ضرایب رگرسیونی $\underline{\theta}$ را برآورد کنیم. با استفاده از رابطه (۹) و نیز با فرض

$$H_0(t;\theta) = \int_0^t h(u;\theta)du = -log\overline{F}_0(t_i;\theta)$$

و همچنین طبق (۷) تابع درستنمایی شرطی بر اساس مشاهدات به‌صورت زیر است:

$$\begin{aligned} L(\beta,\theta \mid \underline{\mathbf{x}}, v) &= \prod_{i=1}^{n} (f(t_i \mid x_i, v_i))^{\delta_i} \left(\overline{F}(t_i, x_i, v_i)\right)^{1-\delta_i} \quad (۱۰)\\ &= \prod_{i=1}^{n} (h(t_i \mid x_i, v_i))^{\delta_i} \overline{F}(t_i, x_i, v_i)\\ &= \prod_{i=1}^{n} \left(v_i exp(\underline{\beta}' x_i) g_0(t_i,\theta)\right)^{\delta_i}\\ &\times \quad exp\left(-v_i exp(\underline{\beta}' x_i) R_0(t_i,\theta)\right) \end{aligned}$$

برای برازش مدل بی‌ثبات گاما فرض می‌کنیم متغیر بی‌ثبات V دارای توزیع گاما با میانگین یک و واریانس σ ($V \sim \Gamma(\frac{1}{\sigma}, \sigma)$) باشد، یعنی،

$$g(v) = \frac{v^{\frac{1}{\sigma}-1}}{\sigma^{\frac{1}{\sigma}}\Gamma(\frac{1}{\sigma})} exp(-\frac{v}{\sigma}) \qquad v > 0, \sigma > 0 \quad (۱۱)$$

بنابراین تابع درستنمایی غیرشرطی بر اساس مشاهدات و روابط (۱۰) و (۱۱) به‌صورت زیر است:

$$\begin{aligned} L(\beta,\theta,\sigma \mid \underline{\mathbf{x}}) &= \int_0^{\infty} L(\beta,\theta \mid \underline{\mathbf{x}}, v)h(v)dv \quad (۱۲)\\ &= \int_0^{\infty} \prod_{i=1}^{n} \left(v_i exp(\underline{\beta}' x_i)h_0(t_i,\theta)\right)^{\delta_i} exp\left(-v_i exp(\underline{\beta}' x_i)R_0(t_i,\theta)\right) \frac{v_i^{\frac{1}{\sigma}-1}}{\sigma^{\frac{1}{\sigma}}\Gamma(\frac{1}{\sigma})} exp(-\frac{v_i}{\sigma})dv_i\\ &= \prod_{i=1}^{n} \int_0^{\infty} \left(v_i exp(\underline{\beta}' x_i)h_0(t_i,\theta)\right)^{\delta_i} exp\left(-v_i exp(\underline{\beta}' x_i)R_0(t_i,\theta)\right) . \frac{v_i^{\frac{1}{\sigma}-1}}{\sigma^{\frac{1}{\sigma}}\Gamma(\frac{1}{\sigma})} exp(-\frac{v_i}{\sigma})dv_i\\ &= \prod_{i=1}^{n} \frac{\left(exp(\underline{\beta}' x_i)h_0(t_i,\theta)\right)^{\delta_i}}{\sigma^{\frac{1}{\sigma}}\Gamma(\frac{1}{\sigma})} \int_0^{\infty} v_i^{\delta_i+\frac{1}{\sigma}-1} exp\left(-\frac{v_i}{\sigma} - v_i exp(\underline{\beta}' x_i)R_0(t_i,\theta)\right) dv_i\\ &= \prod_{i=1}^{n} \frac{\left(exp(\underline{\beta}' x_i)h_0(t_i,\theta)\right)^{\delta_i}}{\sigma^{\frac{1}{\sigma}}\Gamma(\frac{1}{\sigma})} \int_0^{\infty} v_i^{\delta_i+\frac{1}{\sigma}-1} exp\left(-\frac{v_i}{\sigma} - \frac{v_i}{\sigma}\sigma exp(\underline{\beta}' x_i)R_0(t_i,\theta)\right) dv_i\\ &= \prod_{i=1}^{n} \frac{\left(exp(\underline{\beta}' x_i)h_0(t_i,\theta)\right)^{\delta_i}}{\sigma^{\frac{1}{\sigma}}\Gamma(\frac{1}{\sigma})} \int_0^{\infty} v_i^{\delta_i+\frac{1}{\sigma}-1} e^{-\frac{v_i}{\sigma}\left(1+\sigma exp(\underline{\beta}' x_i)R_0(t_i,\theta)\right)} dv_i\\ &= \prod_{i=1}^{n} \frac{\left(exp(\underline{\beta}' x_i)h_0(t_i,\theta)\right)^{\delta_i}}{\sigma^{\frac{1}{\sigma}}\Gamma(\frac{1}{\sigma})} \frac{\Gamma(\delta_i+\frac{1}{\sigma})}{\left(1+\sigma exp(\underline{\beta}' x_i)R_0(t_i,\theta)\right)^{\delta_i+\frac{1}{\sigma}}}\\ &= \prod_{i=1}^{n} \frac{\Gamma(\delta_i+\frac{1}{\sigma})}{\sigma^{\frac{1}{\sigma}}\Gamma(\frac{1}{\sigma})} \left(\frac{exp(\underline{\beta}' x_i)h_0(t_i,\theta)}{1+\sigma exp(\underline{\beta}' x_i)R_0(t_i,\theta)}\right)^{\delta_i} \times \left(1+\sigma exp(\underline{\beta}' x_i)R_0(t_i,\theta)\right)^{-\frac{1}{\sigma}} \end{aligned}$$

$$= \prod_{i=1}^{n} \frac{\Gamma(\delta_i + \frac{1}{\sigma})}{\sigma^{\frac{1}{\sigma}}\Gamma(\frac{1}{\sigma})} \left(\frac{exp(\underline{\beta}' x_i) h_0(t_i, \theta)}{1 - \sigma exp(\underline{\beta}' x_i) log\overline{F}_0(t_i; \theta)} \right)^{\delta_i} \times \left(1 - \sigma exp(\underline{\beta}' x_i) log\overline{F}_0(t_i; \theta)\right)^{-\frac{1}{\sigma}}$$

۴ مثال کاربردی

اکنون برازش مدل‌های (۸) به داده‌های سرطان ریه (کلیفلش و پرنتیس [۱۲]) بررسی می‌شود. در آزمایش مربوط به بیماران سرطان ریه، ۱۳۷ مرد مبتلا به سرطان ریه پیشرفته به‌طور تصادفی تحت درمان معمولی و یا شیمی‌درمانی قرار می‌گیرند. نقطه‌ی پایان آزمایش، زمان فوت بیمار در نظر گرفته‌شده است. در اینجا برای برازش مدل (۹) به داده‌ها با در نظر گرفتن دو حالت زیر برای توزیع پایه در توابع درستنمایی (۱۲) از آن برای برآورد پارامترهای مدل استفاده می‌شود.

الف) توزیع پایه نمایی باشد یعنی تابع نرخ خطر آن به‌صورت زیر باشد:

$$g(t; c, \gamma) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

ب) توزیع پایه وایبول باشد یعنی تابع نرخ خطر آن به‌صورت زیر باشد:

$$g(t; \lambda) = \frac{c\gamma^c x^{c-1} e^{-(\gamma x)^c}}{e^{-(\gamma x)^c}} = c\gamma^c x^{c-1}$$

که در آن $\underline{\theta} = (c, \gamma)$ است. در این مطالعه فرض شده است بیمارانی که از شروع آزمایش دارای طول عمر بیشتر از ۱۵ ماه بوده‌اند، به‌عنوان داده‌های سانسور شده از راست باشند. همچنین تنها نوع درمان بیمار به‌عنوان متغیر توضیحی در نظر گرفته‌شده است. فرض می‌کنیم یک عامل ژنتیکی که قابل‌مشاهده نیست در درمان بیماری افراد مؤثر باشد، که در اینجا از آن به‌عنوان متغیر بی‌ثبات یاد می‌شود.

تذکر: یکی از شیوه‌ها برای آزمودن فرض

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$$

آزمون نسبت درستنمایی است. آماره این آزمون عبارت است از:

$$\chi^2_{(LR)} = 2\left(logL(\widehat{\beta}) - logL(\beta_0)\right)$$

که تحت فرض H_0 برای n های بزرگ دارای توزیع کای دو با یک درجه آزادی است. بنابراین در سطح α فرض H_0 رد می‌شود اگر

$$\chi^2_{(LR)} > \chi^2_{1,\alpha}.$$

بنابراین برای بررسی تأثیر متغیر توضیحی بر متغیر پاسخ یعنی آزمون

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$$

آماره آزمون نسبت درستنمایی محاسبه و با مقدار توزیع کای دو با یک درجه آزادی مقایسه می‌شود. در واقع کافی است تفاوت دو مقدار $-2logL$ در حالتی که متغیر توضیحی در مدل حضور دارد و نیز در حالتی که متغیر توضیحی حذف‌شده است محاسبه‌شده و با مقدار توزیع کای دو با یک درجه آزادی مقایسه شود که آماره آزمون نسبت درستنمایی و p-مقدار در جدول (۱) آورده شده است. به این نتیجه رسیدیم که نوع درمان تأثیر معنی‌داری بر طول عمر بیماران دارد. در این مثال توزیع وایبول را با توزیع نمایی مقایسه می‌کنیم. نتایج به‌دست‌آمده از برازش مدل‌ها، برآورد ماکسیمم درستنمایی پارامترهای مدل مطرح‌شده، برآورد انحراف استاندارد برآوردگرها و معیار اطلاع آکائیکه و معیار اطلاع بیزی در جدول (۲) ارائه‌شده است.

نکته: در مورد معیار اطلاع آکائیکه (AIC) باید گفت معیاری برای سنجش نیکویی برازش است. این معیار بر اساس مفهوم آنتروپی بناشده است و نشان می‌دهد که استفاده از یک مدل آماری به چه میزان باعث از دست رفتن اطلاعات می‌شود. به‌عبارت‌دیگر، این معیار تعادلی میان دقت مدل و پیچیدگی آن برقرار می‌کند. این معیار توسط هیروتسوگوآکائیک [۴] برای انتخاب بهترین مدل آماری پیشنهاد شد.

جدول ۱: مقادیر آماره آزمون و p-مقدار برای دو مدل نمایی و وایبول

توزیع پایه	آماره آزمون	p-مقدار
توزیع نمایی	۳۸/۷۹۴	۰/۰۰۰۴
توزیع وایبول	۴۹/۲۰۴	۰/۰۰۰۲

با توجه به داده‌ها، چند مدل آماری ممکن است با توجه به مقدار AIC رتبه‌بندی شوند و مدل دارای کمترین AIC، بهتر است. در حالت کلی (AIC) برابر است با:

$$AIC = -2ln(L) + 2k = -2\ell(\widehat{\theta}) + 2k$$

که k تعداد پارامترهای مدل آماری است و L مقدار ماکسیمم تابع درستنمایی برای مدل برآورد شده است. معیار اطلاع بیزی (BIC) به‌صورت زیر است:

$$BIC = ln(n)k - 2ln(L) = ln(n)k - 2\ell(\widehat{\theta})$$

که k تعداد پارامترهای مدل آماری است و L مقدار ماکسیمم تابع درستنمایی برای مدل برآورد شده و n تعداد مشاهدات است. در توضیح معیار BIC باید گفت همانند AIC برای انتخاب مدل از بین چند مدل به کار می‌رود و بر اساس تابع درستنمایی است و لذا بسیار مرتبط با AIC می‌باشد. زمانی که مدل برازش می‌دهیم این امکان وجود دارد که تابع درستنمایی با اضافه کردن پارامتر، افزایش پیدا کند و این امر ممکن است منجر به بیش برآورد شود که در این شرایط AIC و BIC، این مسئله را با اعمال محدودیت روی تعداد پارامترهای مدل حل می‌کنند که محدودیت BIC در تعداد پارامترها نسبت به AIC بیشتر است. هنگامی‌که در جستجوی بهترین برازش برای داده‌ها در بین سایر توزیع‌ها هستیم توزیع با کوچک‌ترین مقدار AIC به‌عنوان بهترین برازش انتخاب می‌شود. باید توجه داشت که در مبحث تشخیص مدل معیار AIC با استفاده از مشاهدات به‌دست‌آمده به انتخاب مدل مناسب می‌پردازد. ممکن است با تغییرات اندکی در مشاهدات، این معیار مدل دیگری را به‌عنوان مدل درست مشخص کند. پس اگر داده‌ها در نقاط مناسبی جمع‌آوری نشده باشند این‌چنین معیارهایی به نتایج گمراه‌کننده منجر شده و مدل را به‌درستی انتخاب نمی‌کنند. بنابراین لازم است مشاهدات به‌گونه‌ای باشند که معیار انتخاب مدل در برابر آن‌ها پایدار بوده و تصمیم این معیارها در تعیین مدل درست با تغییر مشاهدات از ثبات کافی برخوردار باشد.

جدول ۲: مقادیر MLE و لگاریتم درستنمایی (LL) و BIC و AIC برای دو مدل نمایی و وایبول

توزیع پایه	پارامترها	برآورد	انحراف استاندارد	AIC	BIC	$loglike$
توزیع نمایی	λ	۰/۰۳۷۴	۰/۰۱۰۹	۱۴۵۱/۱۸۸	۱۴۵۹/۹۲۶	-۷۲۲/۵۹۳۹
	β	۰/۰۳۶۶	۰/۰۴۸۶			
	σ	-۰/۰۲۴۵	۰/۰۰۹۱			
توزیع وایبول	c	۱/۱۳۵۸	۰/۱۳۹۴	۱۴۴۲/۷۰۰	۱۴۵۴/۳۵۱	-۷۱۷/۳۵۰۱
	γ	۰/۰۹۲۴	۰/۰۳۸۵			
	σ	۰/۲۵۷۹	۰/۲۰۵۱			
	β	-۰/۰۴۱۱	۰/۰۰۷۱			

همان‌طور که مشاهده می‌شود توزیع وایبول دارای معیار اطلاع آکائیکه و بیزی کمتری نسبت به توزیع نمایی است بنابراین مدل بی‌ثبات نرخ خطر متناسب گاما با توزیع پایه وایبول نسبت به توزیع نمایی، برازش مناسب‌تری به این داده‌ها ارائه کرده است.

۵ نتیجه‌گیری

در این مقاله مدل نرخ خطر متناسب را به‌عنوان مدل بی‌ثبات در نظر گرفتیم و ویژگی‌های توزیعی آن را موردبررسی قراردادیم. همچنین در قالب یک مثال کاربردی نشان دادیم که مدل بی‌ثبات نرخ خطر متناسب گاما با توزیع پایه وایبول با توجه به کمتر بودن معیارهای مقایسه مدل نسبت به توزیع نمایی، برای برازش به این داده‌ها مناسب است.

مراجع

[1] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes. *Ann. Stat.*, **6**, 701–726.

[2] Aalen, O.O. (1998). Heterogeneity in survival analysis. *Statist. Med.*, **7**, 1121-1137.

[3] Aalen, O.O. (1992). Modelling heterogeneity in survival analysis by the compound poisson distribuithion. *Ann. Appl. Probab.*, **2**, 951-992.

[4] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

[5] Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and survival analysis*. Wiley, New York.

[6] Gupta, R.C. and Kirmani, S. (2005). Stochastic comparisons in frailty models. *J Stat Plan Inference.*, **136**, 3647–3658.

[7] Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**, 387–396.

[8] Hougaard, P. (1995). Frailty models for survival analysis. *Lifetime Data Anal.*, **1**, 255-273.

[9] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag, New York.

[10] Hougaard, P. (1984). Life table methods for heterogeneous population: distribution *describing the heterogeneity*. Biometrika, **71**, 75-83.

[11] Hougaard, P. (1991). Modelling heterogeneity in survival analysis. *J. Appl. Prpbab.*, **28**, 695-701.

[12] Kalbfleisch, D. and Prentice, R. L. (2011), *The Statistical Analysis of Failure Time Data*, John Wiley, New York.

[13] Kayid, M., Izadkhah, S., & Zuo, M. J. (2017). Some results on the relative ordering of two frailty models. *Statistical Papers*, **58(2)**, 287-301.

[14] Liang, K. Y., Self, S. G., Bandeen-Roche, K. J. and Zeger, S. L. (1995), Some Recent Developments for Regression Analysis of Multivariate Failure Time Data, *Lifetime Data Analysis*, **1**, 403-415.

[15] Muller, A. and Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*. Wiley, NewYork.

[16] Nelsen, R. B. (2006). *An Introduction to Copulas*. Lectures Notes in Statistics, 139, Springer-Verlag, New York.

[17] Vaupel, J.W., Manton, K.G. and Stallard, E., (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography.*, **16**, 439–454.

[18] Xu, M. and Li, X.,(2008). Negative dependence in frailty models. *J Stat Plan Inference.*, **138**, 1433–1441.