
بسم الله الرحمن الرحیم

دانشکده علوم ریاضی

گروه آمار

پایان‌نامه کارشناسی ارشد

تحلیل بیزی مدل‌های نرخ شفایافتگی غیرآمیخته با تابع بقای نیمه‌پارامتری کاکس به کمک P-اسپلاین‌ها

نداشمس

استادان راهنما

دکتر حسین باغیشنی و دکتر نگار اقبال

تیر۱۳۹۷

تقدیم به آن‌ها که سرآغاز تولد من هستند. از یکی زاده می‌شوم و از دیگری جاودانه.

استادی که سپیدی را بر تخته سیاه زندگیم نگاشت و پدر و مادری که تار مویی از آن‌ها به پای من سیاه نماند.

سپاس الطاف بی کران پرودگار متعال راکه آغاز و پایان همه چیز است. تشکر و سپاس از اساتید فرزانه ام سرکار خانم دکتر نگار اقبال و جناب آقای دکتر حسین باغیشنی که از محضر پر فیض تدریسشان بهره های فراوان بردم. از اساتید دلسوز جناب آقای دکتر داوود شاهسونی و جناب آقای دکتر محمدرضا ربیعی که زحمت داوری این پایان نامه را متقبل شدند، کمال تشکر و امتنان را دارم. بر خود وظیفه می دانم از زحمات بی دریغ خانواده عزیزم و دوستان خوبم خانم ها زهره فرهادی و محدثه جامعی قدردانی نمایم که بدون کمک آن ها این کار به نتیجه نمی رسید.

یدا شمس
تیر ۱۳۹۷

تعهدنامه

اینجانب **ندا شمس** دانشجوی کارشناسی ارشد رشته **آمار ریاضی** **دانشکده علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **تحلیل بیزی مدل‌های نرخ شفایافتگی غیرآمیخته با تابع بقای نیمه‌پارامتری کاکس به کمک P-اسپلاین‌ها** ، تحت راهنمایی **دکتر حسین باغیشنی** و **دکتر نگار اقبال** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام " دانشگاه صنعتی شاهرود " یا " Shahrood University of Technology " به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده) شده است، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

ندا شمس
تیر ۱۳۹۷

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان‌نامه بدون ذکر منبع مجاز نمی‌باشد.

چکیده

در تحلیل بقا، معمولا فرض می‌شود که هر واحد، پیشامد مورد نظر را در صورت مشاهده طی مدت طولانی، تجربه می‌کند. با این حال می‌توان به صراحت فرض کرد که نسبت ناشناخته‌ای از جمعیت هرگز پیشامد مورد آزمایش را تجربه نخواهند کرد و به اصطلاح شفایافته محسوب می‌شوند. مدل‌های زمان پیشرفت، زیرده‌ای از مدل‌های بقای شفایافته هستند که این ویژگی را در نظر می‌گیرند. رهیافت‌های متعددی برای تعیین تابع توزیع این مدل ارائه شده‌اند که یکی از آن‌ها رهیافت مدل مخاطرات متناسب کاکس است. در این پایان‌نامه از یک رهیافت نیمه‌پارامتری برای مدل‌بندی تابع بقا استفاده شده و اثرات ناپارامتری در آن با P-اسپلاین‌ها مدل‌بندی شده‌اند. هم‌چنین دیدگاه استنباط بر مبنای مکتب بیزی است. مسأله شناسایی‌پذیری مدل هنگامی‌که طول زمان مطالعه به اندازه کافی طولانی نباشد، مورد بحث قرار گرفته شده است. دقت مدل پیشنهادی با استفاده از شبیه‌سازی مورد ارزیابی قرار گرفته و کاربرد مدل بر اساس داده‌های حاصل از یک آزمایش بالینی نمایش داده شده است.

کلمات کلیدی: تحلیل بقا، مدل زمان پیشرفت، P-اسپلاین، مدل مخاطرات متناسب کاکس، استنباط بیزی.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مقدمات و تعاریف اولیه

۱.۱ مقدمه

تحلیل بقا[۱] مجموعه‌ای از روش‌های آماری متنوع برای تحلیل متغیرهای تصادفی نامنفی است. در واقع، متغیر تصادفی T، متغیر تصادفی بقا نامیده می‌شود، اگر تکیه‌گاه آن در بازه‌ی $[۰, \infty)$ باشد. مشاهده‌های $t_۱, t_۲, \cdots, t_n$، از متغیر تصادفی نامنفی T را داده‌های بقا می‌نامند.

یکی از مهم‌ترین مشخصات داده‌های بقا، زمان از کارافتادگی یک پدیده‌ی فیزیکی، مکانیکی، الکتریکی یا زمان مرگ یک موجود زنده است. در واقع متغیر پاسخ، زمان وقوع (طول عمر) یک پدیده است. در این‌جا به‌طور خلاصه به چند مورد از مسائل تحلیل بقا اشاره می‌کنیم:

- مطالعه چند هفته‌ای بیماران مبتلا به سرطان خون، برای این‌که مشخص شود چه زمانی طول می‌کشد تا آن‌ها بهبود یابند.
- مطالعه چند ساله افراد سالم به صورت گروهی، برای این‌که مشخص شود که آیا این افراد به بیماری قلبی مبتلا می‌شوند یا خیر.
- مطالعه بر روی زندانی‌هایی که تازه از زندان آزاد شده‌اند و می‌خواهیم ببینیم که چند هفته طول می‌کشد تا برای بار دوم دستگیر شوند.
- مطالعه بر روی بیماران قلبی که در گذشته پیوند قلب انجام داده‌اند و می‌خواهیم ببینیم که چند هفته بعد از عمل زنده می‌مانند.

تمامی مسائل ذکرشده، مسائل مربوط به تحلیل بقا می‌باشند زیرا متغیر پاسخ، زمان لازم تا رخداد پیشامد است. در مثال اول با سرطان خون مواجه هستیم و پیشامد مورد نظر، بهبودی است و متغیر پاسخ، زمان لازم (هفته) تا بهبودی فرد بیمار است. در مثال دوم پیشامد مورد نظر، بهتر شدن بیماری قلبی و پاسخ، زمان لازم تا بهبودی کامل است. در مثال سوم مطالعات روان‌شناسی بدون مطالعات پزشکی داریم که

[۱] Survival Analysis

در آن پیشامد مورد نظر، دستگیری مجدد و پاسخ، زمان لازم (هفته) تا بازداشت دوباره فرد است. نهایتا در مثال چهارم همان‌طور که ملاحظه می‌شود پیشامد مورد نظر مرگ و متغیر پاسخ، زمان (هفته) لازم تا وقوع مرگ (تعداد هفته‌های بعد از دریافت پیوند) است. گاهی ممکن است متغیر تصادفی مورد نظر مربوط به زمان نباشد. مانند مقادیر متغیر مربوط به هزینه‌ی پرداختی یک شرکت بیمه به بیمه‌شدگان در یک وضعیت خاص. تحلیل بقا نخستین بار توسط جان گرانت (۱۶۶۲) و ستاره‌شناس معروف، ادموند هالی[۲] به کارهای ابتدایی روی جدول‌های مرگ و میر در قرن هفدهم انجام گرفت، اما دوره‌ی نوین آن از پنجاه سال پیش با کاربردهای مهندسی آغاز شد.

۲.۱ تابع بقا

در ادبیات تحلیل بقا واژگان و عبارت‌های زیادی وجود دارند که دارای معانی یکسانی هستند. این اصطلاحات در عرصه‌های مختلف مهندسی و پزشکی به‌کار می‌روند. به عنوان مثال، طول زمان از شروع یک مشاهده تا هنگامی که پیشامد مرگ یا شکست رخ دهد، ”طول عمر” یا ”زمان بقا” نامیده می‌شود. به پیشامد مورد نظر عنوان ”شکست” داده می‌شود زیرا نوع پیشامد مورد نظر معمولا مرگ یا تجربه‌های منفی افراد است. البته گاهی اوقات آن پیشامد می‌تواند یک تجربه مثبت نیز باشد. مانند بازگشت به کار بعد از انجام یک عمل جراحی قلب.

فرض کنید T دارای تابع چگالی احتمال f و تابع توزیع تجمعی F باشد، در این‌صورت

$$F(t) = P(T \leq t) = \int_{\circ}^{t} f(u)\mathrm{d}u.$$

در تحلیل بقا به تابع $F(t)$، تابع شکست[۳] می‌گویند.

تعریف ۱.۲.۱. تابع بقا (تابع قابلیت اعتماد)، برای تمام مقادیر t، به‌صورت $S(t) = ۱ - F(t)$ تعریف می‌شود، به عبارت دیگر

$$P(T > t) = ۱ - F(t) \equiv \bar{F}(t) \equiv S(t).$$

بنابراین

$$S(t) = P(T > t) = \int_{t}^{\infty} f(u)\mathrm{d}u \qquad (۱.۱)$$

و

$$f(t) = -\frac{\mathrm{d}}{\mathrm{d}t}S(t).$$

[۲] کاشف ستاره‌ی دنباله‌دار هالی

[۳] Failure Function

بررسی توابع بقا از اصلی‌ترین بخش‌های تحلیل بقاست. با استفاده از رابطه‌ی (۱.۱) و هم‌چنین با توجه به این نکته که انتگرال تابع چگالی احتمال برابر یک است، داریم

$$S(۰) = P(T > ۰) = \int_{۰}^{\infty} f(u)\mathrm{d}u = ۱.$$

هم‌چنین

$$S(\infty) = \lim_{t\to\infty} S(t) = \lim_{t\to\infty} \int_{t}^{\infty} f(u)\mathrm{d}u = ۰.$$

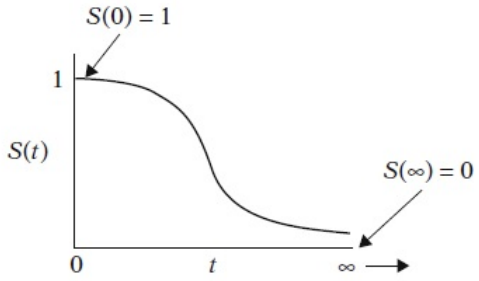
در نهایت، اگر $a \geq b$، در این‌صورت

$$S(b) - S(a) = \int_{b}^{a} f(u)\mathrm{d}u \geq ۰.$$

بنابراین

$$۰ \leq S(t) \leq ۱.$$

در واقع هر تابع یکنوا و نزولی $S(\cdot)$، با تکیه‌گاه $[۰, \infty)$ که $S(۰) = ۱$ و $S(\infty) = ۰$، تابع بقای یک متغیر تصادفی بقا را تعریف می‌کند (شکل ۱.۱). متغیر تصادفی مورد نظر متغیری است که دارای تابع چگالی احتمال $f(t) = -\frac{\mathrm{d}}{\mathrm{d}t}S(t)$ است.



شکل ۱.۱: یک نمودار معمول تابع بقا

ویژگی‌های تابع بقا

برای تابع بقای $S(\cdot)$ می‌توان ویژگی‌های زیر را برشمرد:

۱) $\frac{\mathrm{d}S}{\mathrm{d}t} < ۰$.

۲) $\frac{\mathrm{d}^{۲}S}{\mathrm{d}t^{۲}} \geq ۰$.

۳) $f(t) = \lim_{\Delta t\to ۰} \frac{P(t\leq T\leq t+\Delta t)}{\Delta t} = \frac{\mathrm{d}F(t)}{\mathrm{d}t} = -\frac{\mathrm{d}S(t)}{\mathrm{d}t} \geq ۰$.

۳.۱ تابع نرخ مخاطره

همان‌طور که اشاره شد، تابع بقا $S(t)$، بیانگر احتمال زنده بودن در سن t، یا قابلیت اعتماد یک واحد در زمان t است. در این بخش مفهوم دیگری تحت عنوان تابع مخاطره مورد بررسی قرار می‌گیرد. تابع مخاطره[۴] عبارت است از احتمال شکست یا از کارافتادگی واحد مورد نظر در بازه‌ی $(t, t+\Delta t)$، وقتی که بدانیم تا زمان t زنده بوده است. به این تابع، میزان مرگ و میر آنی، میزان شرطی از کار افتادگی یا شدت مرگ و میر نیز گفته می‌شود که به‌صورت

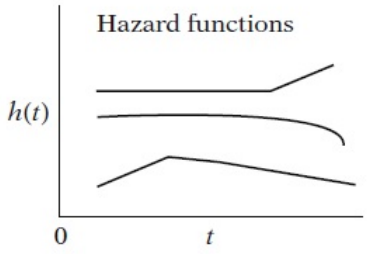
$$\begin{aligned} P(t<T<t+\Delta t|T>t) =& \frac{P(t<T<t+\Delta t, T>t)}{P(T>t)} \\ =& \frac{P(t<T<t+\Delta t)}{P(T>t)} \end{aligned}$$

تعریف می‌شود. کسر فوق را متوسط نرخ شکست در بازه $(t, t+\Delta t)$ گویند. اگر Δt را به سمت صفر میل دهیم، در این‌صورت متوسط نرخ شکست به نرخ شکست آنی تبدیل می‌شود. این تابع می تواند یکنوا یا غیر یکنوا باشد (شکل ۲.۱).

تعریف ۱.۳.۱. متغیر تصادفی بقای T، دارای تابع مخاطره یا نرخ مخاطره یا شدت مرگ و میر $h(t)$ می‌باشد، که برای مقادیر $t>0$، به‌صورت

$$h(t) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{F(t)} = \frac{f(t)}{S(t)}$$

تعریف می‌شود. $h(t)$ نرخ آنی مرگ در زمان t، برای واحدهایی است که تا لحظه t زنده مانده یا از کار نیفتاده‌اند و $f(t)$ تابع چگالی احتمال T است.



شکل ۲.۱: انواع تابع مخاطره

[۴] Hazard Function

روابط بین تابع بقا و تابع مخاطره

طبق تعریف تابع مخاطره می‌توان نوشت

$$h(t)=\frac{f(t)}{۱-F(t)}=\frac{-\frac{\mathrm{d}[۱-F(t)]}{\mathrm{d}t}}{۱-F(t)}=\frac{\mathrm{d}\{-\ln[۱-F(t)]\}}{\mathrm{d}t}=\frac{\mathrm{d}\{-\ln[S(t)]\}}{\mathrm{d}t}.$$

بنابراین

$$\int_{۰}^{t}h(u)\mathrm{d}u=-\ln[۱-F(u)]|_{۰}^{t}.$$

اما $F(۰)=۰$ و $\ln(۱)=۰$، پس

$$\ln[۱-F(t)]=\ln[S(t)]=-\int_{۰}^{t}h(u)\mathrm{d}u$$

و در نتیجه

$$S(t)=\exp\{-\int_{۰}^{t}h(u)\mathrm{d}u\}.$$

پس

$$H(t)=\int_{۰}^{t}h(u)\mathrm{d}u=-\ln[S(t)]$$

که $H(t)$، تابع توزیع تجمعی $h(t)$ است. بنابراین

$$S(t)=\exp[-H(t)].$$

لازم به ذکر است که

$$H(t)\geq ۰$$

و

$$\frac{\mathrm{d}H(t)}{\mathrm{d}t}=h(t).$$

۴.۱ سانسور

داده‌های مربوط به زمان، تا لحظه‌ی رخداد یک پیشامد به انواع مختلفی پدیدار می‌شوند که کار با آن‌ها را سخت می‌سازد. یکی از این ویژگی‌ها، که از ویژگی‌های اصلی داده‌های بقا در مقایسه با دیگر داده‌های

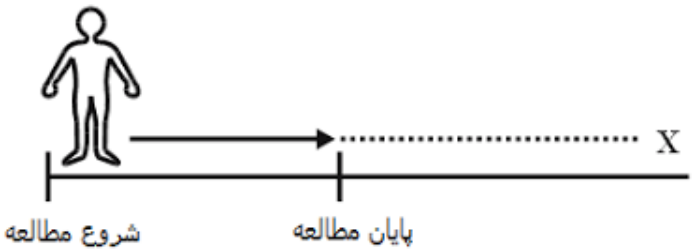
آماری است و گاهی اوقات به وقوع می‌پیوندد ، سانسور[۵] نام دارد. منظور از داده سانسورشده، داده‌ای است که مقدار دقیق آن مجهول بوده و به جای آن از اطلاعات موجود استفاده شده باشد. سانسور به‌طور کلی زمانی اتفاق می‌افتد که تعدادی از طول عمرها، فقط در فاصله‌های مشخصی اتفاق افتاده‌اند و بقیه طول عمرها به‌طور کامل مشاهده می‌شوند. در واقع سانسور به نداشتن دقیق زمان بقا اطلاق می‌شود (امکان عدم رخداد پیشامد مورد نظر در زمان تعیین‌شده برای مطالعه).

به طور کلی سه دلیل عمده برای سانسور شدن وجود دارد:

۱. مطالعه به پایان می‌رسد اما پیشامدی رخ نمی‌دهد (زمان بقا بیشتر از دوران مطالعه). شکل ۳.۱ این نوع از سانسور را نشان می دهد.

۲. فرد مورد مطالعه در زمان مطالعه گم شود و تحت پیگیری قرار نگیرد.

۳. فرد مورد مطالعه بمیرد (اگر مرگ پیشامد مورد نظر نباشد) یا عوارض جانبی دیگر اتفاق بیافتد.



شکل ۳.۱: علامت X به معنی رخداد پیشامد است.

تعریف ۱.۴.۱. نمونه‌های سانسورشده آن‌هایی هستند که ممکن است در طول مطالعه به‌طورکامل شرکت نداشته یا تا پایان مطالعه از کار نیفتاده‌اند. زمان سانسور را با T_{cens} یا با C نشان می‌دهند. به‌طور مشابه زمان پیشامد بقا را با T_{surv} نشان می‌دهند. اگر F تابع توزیع تجمعی باشد، آنگاه

$$F_{cens} = P(T_{cens} \leq t).$$

هم‌چنین اگر $G(t)$ تابع سانسور باشد، آن را به‌صورت زیر تعریف می‌کنیم:

$$G(t) = ۱ - F_{cens}(t) = P(T_{cens} > t).$$

حال با توجه به $F_{surv}(t) = P(T_{surv} \leq t)$ و $S(t) = ۱ - F_{surv}(t) = P(T_{surv} > t)$ و این‌که می‌دانیم T_{surv} و T_{cens} در عمل با هم اتفاق نمی‌افتند، پس

$$T = \min\{T_{surv}, T_{cens}\}.$$

[۵] Censor

نمونه‌های سانسورشده اغلب نتیجه‌ای از آزمایش‌های طول عمر و زمان واکنش هستند. نخستین بار کریچ استفاده از عنوان سانسورشده را برای این نمونه‌ها پیشنهاد کرد. به‌عنوان مثال برای یک بیمار مبتلا به سرطان خون تا زمان بهبود، اگر مطالعه به پایان برسد بدون این‌که بهبودی حاصل شود (پیشامد رخ ندهد)، یا شخص به دلایل مختلف از مطالعه خارج شود، گوییم این پیشامد برای بیمار سانسور شده است. در نتیجه می‌بینیم که زمان بقا برای این شخص طولانی‌تر از مدت زمانی است که شخص تحت مطالعه است. حال اگر شخص بعد از پایان مطالعه بهبود یابد ما زمان دقیق بقا را نخواهیم دانست، که به آن سانسور از راست[۶] می‌گویند. در موارد نادری نیز ممکن است، پیشامد مورد مطالعه برای یک فرد، قبل از ورود به مطالعه، رخ داده باشد. اما زمان دقیق رخداد پیشامد نامعلوم باشد که به این وضعیت سانسور از چپ[۷] می‌گویند. سانسور از چپ برای داده‌ها زمانی اتفاق می‌افتد که زمان بقای فرد، کوچکتر یا مساوی زمان بقای مشاهده‌شده‌ی آن فرد باشد. در واقع زمان بقای فرد مورد مطالعه، کوچکتر از زمان سانسور باشد. به‌عنوان مثال، فرض کنید در یک مطالعه هدف پیگیری افرادی است که به ویروس $H.I.V$ مبتلا هستند. این مطالعه زمانی آغاز می‌شود که برای اولین بار شخص مورد مطالعه دارای آزمایش $H.I.V$ مثبت باشد. به هر حال زمان دقیق در معرض ویروس قرار گرفتن، معلوم نیست. بنابراین، زمان بقا در سمت چپ مطالعه سانسور شده است. هم‌چنین ممکن است در مطالعه‌ای، افراد در نقاطی از زمان که قبلا در نظر گرفته شده‌اند، مورد معاینه قرار گیرند و پیشامد مورد نظر در فاصله زمانی بین دو معاینه رخ دهد. در این صورت زمان ثبت‌شده، با یک بازه عددی مشخص می‌گردد که به آن سانسور فاصله‌ای[۸] گویند.

انواع سانسور از راست

باتوجه به این‌که بیش‌تر داده‌ها از سمت راست سانسور می‌شوند، سانسور از راست دارای اهمیت ویژه‌ای است. سانسور از راست انواع مختلفی دارد.

سانسور نوع یک

در سانسور نوع یک، که ساده‌ترین نوع سانسور از راست می‌باشد، افراد به‌طور هم‌زمان وارد مطالعه شده و مطالعه پس از گذشت زمانی مشخص به پایان می‌رسد.

سانسور نوع دو

در سانسور نوع دو، افراد هم‌زمان وارد مطالعه شده و پس از وقوع پیشامد برای تعداد یا درصد معینی از افراد مطالعه پایان می‌یابد.

[۶] Right Censoring
[۷] Left Censoring
[۸] Interval Censoring

سانسور تصادفی

نوع دیگری از انواع سانسور از راست، سانسور تصادفی[۹] نامیده می‌شود. در این حالت افراد در زمان‌های متفاوت وارد مطالعه می‌شوند و تا زمان وقوع پیشامد مورد نظر تحت پیگیری قرار می‌گیرند. برخی افراد به دلیلی غیر از پیشامد مورد مطالعه از مطالعه خارج می‌شوند. در چنین مواردی، پیشامد مورد نظر قابل مشاهده نیست. برخی عواملی که سبب سانسور شدن تصادفی فرد می‌شوند، عبارتند از: مرگ‌های ناشی از تصادف، مهاجرت جوامع انسانی، مرگ به دلیلی غیر از پیشامد مورد مطالعه و کناره‌گیری فرد از آزمایش بالینی.

سانسور نوع سه

در برخی مطالعات، طرح سانسور، ترکیبی از سانسور تصادفی و سانسور نوع یک است. در چنین مطالعاتی، برخی بیماران به‌طور تصادفی سانسور می‌شوند. به‌عنوان مثال به دلیلی غیر از پیشامد مورد مطالعه، مطالعه را ترک می‌کنند یا می‌میرند. در حالی‌که دیگران هنگامی که دوره ثابت مطالعه پایان می‌یابد، از نوع یک سانسور می‌شوند.

یکی از مفروضات اساسی در تحلیل بقا، این است که سانسور شدن زمان‌های بقا ناآگاهی‌بخش[۱۰] باشد. یعنی علت سانسور شدن زمان‌های بقای یک فرد به پیشامد مورد مطالعه بستگی نداشته باشد. به عبارتی زمان‌های شکست و زمان‌های بقا برای هر فرد از هم مستقل باشند.

۵.۱ مدل‌های رگرسیون در بقا

تحلیل رگرسیونی یکی از روش‌های آماری برای تحلیل داده‌های چندعاملی است، که حوزه کاربرد آن وسعت زیادی دارد. نتایج آن از نظر مفهوم، فرآیند ساده به‌کارگیری یک معادله است، که ارتباط بین مجموعه‌ای از متغیرها را بیان می‌کند. در واقع تحلیل رگرسیونی تکنیکی آماری برای بررسی و ارتباط بین متغیرها است.

ساده‌ترین حالت در مسائل رگرسیونی رابطه بین یک متغیر تبیینی رگرسیونی با متغیر پاسخ است. معمولا X، متغیر تبیینی و Y، متغیر پاسخ نامیده می‌شود.

۱.۵.۱ برازش مدل‌های رگرسیونی برای داده‌های بقا

شناسایی تابع بقا، تابع مخاطره و برآورد پارامترهای آن‌ها از مهم‌ترین اهداف تحلیل بقا است. تابع بقا رفتار احتمالی زمان بقا را نشان می‌دهد و تابع مخاطره بیانگر قدرت آنی خطر در زمان مشخص می‌باشد. این دو تابع اطلاعات هم‌ارزی در مورد زمان‌های بقا در اختیار محقق قرار می‌دهند. از دیگر اهداف تحلیل بقا می‌توان به مقایسه توابع بقا برای گروه‌های مختلف اشاره کرد. با این حال در بسیاری از کاربردهای

[۹] Random Censoring
[۱۰] Noninformative

بقا، هدف مورد نظر بررسی رابطه متغیرهای تبیینی با زمان بقا است. این کار را می‌توان به وسیله یک رابطه رگرسیونی انجام داد. با توجه به ماهیت داده‌های زمان بقا (حضور داده‌های سانسورشده و چولگی مثبت یا منفی زمان‌های بقا)، مدل‌های رگرسیونی معمولی جواب‌گو نیست. در این صورت لازم است از مدل‌های رگرسیونی مخاطره استفاده نمود. به کمک یک مدل مخاطره می‌توان علاوه بر تعیین میزان اثر عوامل خطر بر زمان‌های بقا، میزان مخاطره یا بقا را برای افراد و گروه‌های مختلف پیش‌بینی نمود.

در این بخش رابطه بین زمان بقای T و مقادیر یک متغیر تبیینی (پیش‌بین) $\mathbf{x}$ بررسی خواهد شد. در چارچوب کلی برای رگرسیون T روی $\mathbf{x}$، اصطلاحات دیگری نیز برای $\mathbf{x}$ به‌کار می‌رود که از جمله می‌توان به متغیر کمکی[۱۱] و متغیر رگرسیونی اشاره کرد، که این اصطلاحات می‌توانند به‌جای یکدیگر استفاده شوند.

در حالت کلی‌تر انتظار می‌رود که زمان بقا وابسته به متغیرهای تبیینی مختلفی باشد که در این‌صورت تمام این متغیرهای تبیینی به‌صورت بردار $\mathbf{x}$ نشان داده می‌شوند. در حالت کلی، $\mathbf{x}$ برداری شامل k متغیر تبیینی است که در آن $\mathbf{x}^T = (x_1, \cdots, x_k)$.

۲.۵.۱ مخاطرات نسبی

از آن‌جایی که متغیر پاسخ به‌گونه‌ای مدل‌بندی می‌شود که وابسته به مقدار بردار متغیرهای تبیینی $\mathbf{x}$ باشد، واضح است که تابع مخاطره نیز به $\mathbf{x}$ وابسته خواهد بود. فرض کنید $h(t)$ نشان‌دهنده‌ی تابع مخاطره T باشد، که در آن $\mathbf{x}$ برداری از متغیرهای کمکی است. حال آن‌چه که باید بررسی شود این است که متغیرهای کمکی چه تاثیری بر تابع مخاطره دارند و این تاثیر چگونه باید مدل‌بندی شود. یکی از رایج‌ترین مدل‌های مورد استفاده که در آن متغیرهای کمکی دارای اثر ضربی بر تابع مخاطره پایه[۱۲] می‌باشند، مدل مخاطره پایه نامیده می‌شود. این مدل را مدل مخاطرات نسبی[۱۳] نیز می‌نامند.

تعریف ۱.۵.۱. فرض کنید T نشان‌دهنده متغیر پاسخی باشد که وابسته به بردار متغیرهای کمکی، $\mathbf{x}$ است. مدل مخاطرات نسبی برای T به‌صورت

$$h(t|\mathbf{x}) = h_0(t)g(\mathbf{x})$$

تعریف می‌شود که در آن $g(\cdot)$ تابعی نامنفی از $\mathbf{x}$ و $h_0(t)$ مخاطره پایه نامیده می‌شود.

بر طبق این تعریف، متغیرهای کمکی به‌واسطه حاصل‌ضرب h_0 که به $\mathbf{x}$ وابسته نیست و g که وابسته به $\mathbf{x}$ است، دارای اثر ضربی بر مخاطره هستند. تابع $g(\cdot)$ در تعریف ۱.۵.۱ به‌گونه‌ای است که به ازای هر $\mathbf{x}$ $g(\mathbf{x}) \geq 0$ و فقط $g(\mathbf{0}) = 1$. این تابع با دو بردار $\mathbf{x}^T = (x_1, \cdots, x_k)$ و $\boldsymbol{\beta}^T = (\beta_1, \cdots, \beta_k)$ معمولا به‌صورت

$$g(\mathbf{x}) = \exp(\mathbf{x}^T\boldsymbol{\beta}) \qquad (۲.۱)$$

[۱۱] Covariates
[۱۲] Baseline Hazard Function
[۱۳] Proportional Hazard

ساخته می‌شود. لذا $g(\mathbf{0}) = \exp(\boldsymbol{\beta}^T\mathbf{0}) = \exp(0) = 1$. نکته قابل توجه این است که در رابطه (۲.۱)، $\mathbf{x}^T\boldsymbol{\beta} = \beta_1 x_1 + \cdots + \beta_k x_k$ ترکیبی خطی از مقادیر کمکی بردار $\mathbf{x}$ است، که در آن ضرایب از بردار پارامترهای $\boldsymbol{\beta}$ گرفته شده‌اند. در نتیجه مدل مخاطرات نسبی استاندارد در ۱.۵.۱ به‌صورت

$$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta}) \tag{۳.۱}$$

می‌باشد. بنابراین مخاطره پایه هنگامی رخ می‌دهد که $\mathbf{x} = \mathbf{0}$. هنگامی که از رابطه (۳.۱) استفاده شود، فرآیند برازش یک مدل مخاطرات نسبی شامل برآورد k پارامتر بردار $\boldsymbol{\beta}$ با استفاده از مقادیر متغیر پاسخ و متغیرهای کمکی خواهد بود.

رابطه مخاطرات نسبی و تابع بقا

اگر $S(t|\mathbf{x})$ تابع بقای یک آزمایش با تابع مخاطره $h(t|\mathbf{x})$ باشد، در این‌صورت با توجه به روابط موجود بین تابع بقا و تابع مخاطره، داریم

$$\begin{aligned} S(t|\mathbf{x}) &= \exp\{-\int_0^t h(u|\mathbf{x})\mathrm{d}u\} \\ &= \exp\{-\int_0^t h_0(u)g(\mathbf{x})\mathrm{d}u\} \\ &= \exp\{-g(\mathbf{x})\int_0^t h_0(u)\mathrm{d}u\} \\ &= (\exp\{-\int_0^t h_0(u)\mathrm{d}u\})^{g(\mathbf{x})} \\ &= (S_{\mathbf{0}}(t))^{g(\mathbf{x})} \end{aligned}$$

که در آن

$$S_{\mathbf{0}}(t) = \exp\{-\int_0^t h_0(u)\mathrm{d}u\}$$

تابع بقای پایه است و $S_0(t) = S_{\mathbf{0}}(t)$ هنگامی رخ می‌دهد که $\mathbf{x} = \mathbf{0}$. به‌طور خلاصه، اثر پذیره مخاطرات نسبی بر $S(t|\mathbf{x})$ این‌است که یک تابع مثبت مانند g و یک تابع بقا پایه مانند S_0 را طوری مشخص می‌کند که

$$S(t|\mathbf{x}) = S_{\mathbf{0}}(t)^{g(\mathbf{x})}. \tag{۴.۱}$$

رابطه (۴.۱) نشان می‌دهد که در مدل مخاطرات نسبی، بردار متغیرهای کمکی دارای یک اثر توانی بر تابع بقای پایه است.

۶.۱ مدل‌های مخاطره

در هنگام تحلیل داده‌های بقا، دو هدف عمده مطرح هستند. یکی مدل‌بندی برای پیدا کردن ترکیبی مناسب از متغیرهای کمکی که طول بقای اعضای مورد مطالعه را تحت تاثیر قرار می‌دهند و هدف بعدی پیدا کردن برآوردهای مطمئن برای تابع مخاطره در زیرگروه‌های مورد بررسی است. به‌طور معمول دو راه‌کار اساسی در تحلیل داده‌های آماری، روش‌های نیمه‌پارامتری[۱۴] و پارامتری[۱۵] هستند و هنگامی در تحلیل بقا با این دیدگاه به بررسی مسأله می‌پردازیم، با توجه به هرکدام از اهداف فوق یکی از راه‌های ذکرشده از اهمیت بیشتری برخوردار خواهند بود.

۱.۶.۱ مدل‌های پارامتری

مدل‌های پارامتری، که برروی تبعیت زمان‌های شکست از خانواده خاصی از توزیع‌های پارامتری پایه‌گذاری شده‌اند. توزیع‌های بسیار زیادی مانند وایبل، نمایی، لگ-نرمال، گامپرتز و لگ-لجستیک را می‌توانند شامل شوند. در ادامه برخی از آن‌ها را معرفی می‌کنیم.

مدل مخاطره ثابت

فرض کنید $\beta > ۰$ یک مقدار ثابت و $\omega = \frac{۱}{\beta}$ باشد. یک مدل مخاطره ثابت معمولا زمانی به‌کار می‌رود که خطر حتمی شکست با زمان تغییر نمی‌کند و به‌صورت

$$h(t) = \omega, \qquad t > ۰$$

است. تابع بقا برای $t > ۰$ به‌صورت زیر محاسبه می‌شود:

$$S(t) = \exp\{-\int_۰^t h(u)\mathrm{d}u\} = \exp\{-\int_۰^t \frac{۱}{\beta}\mathrm{d}u\} = \exp\{-\frac{۱}{\beta}\int_۰^t \mathrm{d}u\} = \exp\{-\frac{t}{\beta}\}.$$

هم‌چنین

$$f(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t} = \frac{۱}{\beta}\exp^{-\frac{t}{\beta}}$$

و $E(T) = \beta$، $Var(T) = \beta^۲$.

نتیجه ۱.۶.۱. متغیر طول عمر T، دارای مدل احتمال نمایی با میانگین $\beta > ۰$ است و به‌صورت $T \sim \exp(\beta)$ نشان داده می‌شود، هرگاه تابع مخاطره آن ثابت و به‌صورت $h(t) = \frac{۱}{\beta}$ باشد.

مدل نمایی یکی از رایج‌ترین مدل‌های احتمال مورد استفاده برای مدل‌بندی طول عمر قطعات است. این مدل برای داده‌های بقا نیز مناسب است.

[۱۴] Semi-Parametric
[۱۵] Parametric

مدل مخاطره توانی

فرض کنید $\alpha > ۰$ و $\beta > ۰$ مقادیر ثابتی باشند. یک مدل مخاطره توانی معمولا زمانی به‌کار می‌رود که خطر حتمی شکست به سرعت با زمان افزایش یابد و به‌صورت

$$h(t) = \frac{\alpha}{\beta^{\alpha}} t^{\alpha-۱}, \qquad t > ۰$$

است. تابع بقا برای مدل مخاطره توانی به‌صورت

$$S(t) = \exp\{-\int_{۰}^{t} h(u)\mathrm{d}u\} = \exp\{-\int_{۰}^{t} \frac{\alpha}{\beta^{\alpha}} u^{\alpha-۱}\mathrm{d}u\} = \exp\{-(\frac{t}{\beta})^{\alpha}\}$$

محاسبه می‌شود و تابع چگالی احتمال آن به‌صورت زیر است:

$$f(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t} = \frac{\alpha}{\beta^{\alpha}} t^{\alpha-۱} \exp\{-(\frac{t}{\beta})^{\alpha}\}, \qquad t > ۰.$$

نتیجه ۲.۶.۱. متغیر تصادفی طول عمر T، دارای مدل احتمال وایبل با پارامترهای $\alpha > ۰$ و $\beta > ۰$ است و $T \sim W(\alpha, \beta)$، هرگاه T دارای یک مخاطره توانی به شکل $h(t) = \frac{\alpha}{\beta^{\alpha}} t^{\alpha-۱}$ باشد.

مدل مخاطره نمایی

مخاطره نمایی با پارامترهای مقیاس b و مکان ν به‌صورت

$$h(t) = \frac{۱}{b} \exp(\frac{t-v}{b}), \qquad t > ۰$$

تعریف می‌شود و اغلب در علوم آماری برای مدل‌بندی طول عمر انسان رخ می‌دهند. تابع بقا برای $t > ۰$ عبارت است از

$$\begin{aligned} S(t) &= \exp\{-\int_{۰}^{t} h(u)\mathrm{d}u\} \\ &= \exp\{-\int_{۰}^{t} \frac{۱}{b} \exp(\frac{u-v}{b})\mathrm{d}u\} \\ &= \exp\{-\exp(\frac{u-v}{b})|_{۰}^{t}\} \\ &= \exp\{-\exp(\frac{t-v}{b} + \exp(-\frac{\nu}{b}))\}. \end{aligned}$$

این تابع بقا را می‌توان به‌صورت زیر نیز نشان داد:

$$S(t) = \exp(\theta(۱ - \exp(\alpha t)))$$

که در این صورت تابع بقا، تابع بقای گامپرتز[۱۶] نامیده می‌شود و اولین بار توسط موش‌برگر و کلین (۱۹۹۷) معرفی شد.

[۱۶] Gompertz Survival Function

روابط پارامتری بین (α, θ) و (t, b) عبارتند از

$$\alpha = \frac{۱}{b}, \qquad \theta = \exp(-\frac{\nu}{b}).$$

تابع چگالی احتمال آن نیز برابر است با

$$f(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t} = \frac{۱}{b}\exp(\frac{t-v}{b})S(t).$$

نتیجه ۳.۶.۱. هرگاه متغیر تصادفی طول عمر T دارای یک مخاطره نمایی به شکل $h(t) = \frac{۱}{b}\exp(\frac{t-v}{b})$، $t > ۰$، باشد، در این صورت T دارای مدل احتمال گامپرتز با پارامترهای $b > ۰$ و $-\infty < \nu < +\infty$ است و با نماد $T \sim Gom(v, b)$ نشان داده می‌شود.

مدل‌های مخاطره دیگری نیز وجود دارند که به‌علت عدم استفاده از آن‌ها در این تحقیق، از بیان آن‌ها اجتناب می‌شود.

۲.۶.۱ مدل‌های نیمه‌پارامتری

مدل‌های نیمه‌پارامتری هیچ پذیره پارامتری خاصی بر روی توزیع زمان شکست پایه ندارند ولی در عوض فرض می‌کنند که مخاطرات متناسب هستند. یکی از معمول‌ترین مدل‌های نیمه‌پارامتری، مدل مخاطرات متناسب کاکس[۱۷] است که در ادامه آن‌را معرفی می‌کنیم.

مدل مخاطرات متناسب کاکس

یکی از مدل‌های آماری بسیار مهم در تحلیل بقا و تحقیقات پزشکی مدل مخاطرات متناسب کاکس است. در سال ۱۹۷۲ کاکس آماردان انگلیسی مدل رگرسیون کاکس را که با داده‌های زمان شکست سروکار دارد، ارائه نمود و از آن زمان تا کنون به‌صورت گسترده مورد استفاده قرار می‌گیرد. تحلیل آماری داده‌های زمان شکست عرصه‌ای مهم در تحقیقات است که رشته‌های کاربردی متعددی به آن توجه دارند. تحلیل بقا در آزمایش‌های بالینی و قابلیت اعتماد در سیستم‌های تولیدی دو نمونه از داده‌های زمان شکست هستند. رگرسیون کاکس هنگامی مورد استفاده قرار می‌گیرد که متغیر پاسخ آن زمان بقا یا مدت زمان تا وقوع پیشامد است. معادله رگرسیون با استفاده از یک یا چند متغیر کمکی، احتمال مرگ یا پیشامد مورد نظر را در طول زمان پیش‌بینی می‌کند. از آن‌جا که مدل رگرسیون کاکس می‌تواند هم یک مدل پارامتری و هم یک مدل ناپارامتری محسوب شود، بسیار حائز اهمیت است. ارتباط گسترده‌ای بین تحلیل بقا و رگرسیون کاکس وجود دارد. در واقع این مدل یک نظریه‌ی آماری فرآیندهای شمارشی است که به توسعه‌ی تحلیل بقای ناپارامتری می‌پردازد و مهم‌ترین مدل رگرسیون نیمه‌پارامتری برای تحلیل داده‌های سانسورشده است.

همان‌طور که در قبل گفته شد، مدل مخاطرات نسبی به‌صورت زیر در نظر گرفته می‌شود:

$$h(t|\mathbf{x}) = h_۰(t)\exp(x_۱\beta_۱ + \cdots + x_k\beta_k) = h_۰(t)\exp(\mathbf{x}^T\boldsymbol{\beta}). \qquad (۵.۱)$$

[۱۷] Cox Proportional Hazard Model

نکته قابل توجه در این مدل آن است که عرض از مبدا β_0 وجود ندارد. موفقیت قابل توجه کاکس این بود که با استفاده از روش وی می‌توان در رابطه‌ی (۵.۱) ابتدا $\boldsymbol{\beta}$ و سپس مخاطره پایه را برآورد کرد. بدیهی است که

$$h_{\mathbf{0}}(t) = h_0(t)$$

همان تابع مخاطره پایه می‌باشد. $h_0(t)$ می‌تواند به‌عنوان هر تابعی از t تعریف شود. تنها شرط لازم این‌است که $h_0(t)$ مثبت باشد. این قسمت ناپارامتری مدل و $\mathbf{x}^T\boldsymbol{\beta}$ قسمت پارامتری مدل است. لذا مدل مخاطرات کاکس یک مدل نیمه‌پارامتری است.

حال با توجه به رابطه (۵.۱) داریم

$$S(t|\mathbf{x}) = (S_0(t))^{\exp(\mathbf{x}^T\boldsymbol{\beta})}.$$

بنابراین

$$F(t|\mathbf{x}) = 1 - S_0(t)^{\exp(\mathbf{x}^T\boldsymbol{\beta})}. \qquad (۶.۱)$$

رابطه (۶.۱) را مدل رگرسیون کاکس می‌نامند.

الزامی نبودن یک توزیع احتمالی برای زمان‌های بقا، یکی از مزایای مدل مخاطرات متناسب کاکس است. مدل کاکس را مدل مخاطرات متناسب می‌نامند زیرا نسبت مخاطره‌ها برای دو دسته یا دو گروه از متغیرهای توضیحی $\mathbf{x}_1$ و $\mathbf{x}_2$ به متغیر زمان t بستگی ندارد. یعنی

$$\frac{h_0(t|\mathbf{x}_1)}{h_0(t|\mathbf{x}_2)} = \frac{h_0(t)\exp(\mathbf{x}_1^T\boldsymbol{\beta})}{h_0(t)\exp(\mathbf{x}_2^T\boldsymbol{\beta})} = \frac{\exp(\mathbf{x}_1^T\boldsymbol{\beta})}{\exp(\mathbf{x}_2^T\boldsymbol{\beta})} = \exp((\mathbf{x}_1^T - \mathbf{x}_2^T)\boldsymbol{\beta}).$$

به دلیل وجود پذیره‌های کمتر در روش‌های نیمه‌پارامتری نسبت به روش‌های پارامتری، محققین علوم پزشکی اغلب به استفاده از این روش‌ها راغب‌ترند. یک پذیره مهم و اساسی در مدل کاکس وجود دارد و آن پذیره متناسب بودن خطر برای تمامی متغیرهای کمکی موجود در مدل نهایی است. یعنی نسبت خطر برای یک نفر در زمان t به خطر برای فرد دیگر در همان زمان، مقدار ثابتی دارد. در صورت برقراری این پذیره، تفسیر مدل به‌دست‌آمده، ساده‌تر از مدل‌های پارامتری خواهد بود.

۷.۱ روش‌های برآورد در مدل‌های نیمه‌پارامتری

روش‌های مختلفی به‌منظور برآورد پارامترهای مدل نیمه‌پارامتری ارائه شده‌اند. روش هموارسازی اسپلاین[۱۸] (انگل و همکاران، ۱۹۸۶؛ هکمن، ۱۹۸۶؛ رایس، ۱۹۸۶ و اسپیک من، ۱۹۸۸)، روش هسته[۱۹] (رابینسون، ۱۹۸۸)، روش رگرسیونی تطبیقی چندگانه[۲۰] (فریدمن، ۱۹۹۰)، روش مانده‌های جزئی

[۱۸] Smoothing Spline
[۱۹] Kernel
[۲۰] Multivariate Variable Adaptive Regression

(کیوزیک و همکاران، ۱۹۹۲)، روش درستنمایی موضعی[۲۱] (کارمن و تاتز، ۲۰۰۱) و روش برآورد موجک[۲۲] (لمینک کیو و چانگ، ۲۰۰۴) از جمله روش‌هایی هستند که توسط محققین پیشنهاد شده‌اند. در این پایان‌نامه از رهیافت اسپلاین برای برآورد قسمت ناپارامتری مدل استفاده خواهیم کرد.

۸.۱ دیدگاه بیزی در استنباط آماری

به‌طور کلی، روش‌های استنباط آماری متعلق به یکی از دو مکتب کلاسیک و بیزی هستند. با توجه به مشکلات موجود در دیدگاه استنباط کلاسیک، دیدگاه بیزی، که استنباط آن مبتنی بر توزیع پسین است، می‌تواند جایگزین مناسبی باشد. در روش‌های متعلق به دیدگاه کلاسیک، پارامتر θ را مقداری ثابت اما نامعلوم فرض می‌کنیم و بر اساس یک نمونه تصادفی $X_1, \cdots, X_n$ که دارای توزیع $f(x|\theta)$ هستند، در مورد θ تصمیم‌گیری (برآورد) می‌کنیم. در دیدگاه بیزی، فرض می‌شود که پارامتر θ خود یک متغیر تصادفی است که تغییرات آن توسط یک توزیع احتمالی بیان می‌گردد. این توزیع احتمالی را توزیع پیشین [۲۳] می‌گویند که بیانگر اطلاعات اولیه محقق (قبل از مشاهده داده‌ها) نسبت به پارامتر است. پس از مشاهده نمونه تصادفی، توزیع پیشین تصحیح می‌شود. به عبارت دیگر اطلاعات اولیه و به‌دست آمده ترکیب می‌شوند که منجر به تصحیح توزیع پیشین می‌شود. به این توزیع پیشین تصحیح‌شده، توزیع پسین[۲۴] می‌گویند. نحوه‌ی تشکیل این توزیع برای یک نمونه معمول n تایی، به صورت زیر است:

فرض کنید $X_1, \cdots, X_n$ یک نمونه تصادفی n تایی با تابع چگالی احتمال $f(\cdot|\theta)$ باشد، که در آن θ پارامتر مورد علاقه می‌باشد و مقادیر مختلف خود را در فضای پارامتر Θ اختیار می‌کند. در یک مدل بیزی، فرض می‌شود مقدار واقعی پارامتر θ نیز تحققی از یک توزیع احتمالی به نام توزیع پیشین است که پیشتر معرفی شد. اگر این توزیع را با $\pi(\theta)$ و تابع درستنمایی را با

$$L(\theta|\mathbf{x}) = f_{\mathbf{x}|\theta}(x_1, \cdots, x_n|\theta)$$

نشان دهیم، آنگاه بنابر قاعده بیز، توزیع پسین $\pi(\theta|\mathbf{x})$ معادل است با

$$\pi(\theta|\mathbf{x}) = \frac{L(\theta|\mathbf{x})\pi(\theta)}{\int_\theta L(\theta|\mathbf{x})\pi(\theta)\mathrm{d}\theta} = c(\mathbf{x})L(\theta|\mathbf{x})\pi(\theta)$$

که در آن $c(x)$ ثابت نرمال‌ساز نامیده می‌شود و برابر است با

$$c(\mathbf{x}) = \frac{1}{\int_\Theta L(\theta|\mathbf{x})\pi(\theta)\mathrm{d}\theta}.$$

کمیت‌های مورد علاقه در استنباط بیزی معمولا به‌صورت یک انتگرال بر حسب تابع چگالی توزیع پسین ظاهر می‌شوند. محاسبه این انتگرال‌ها در صورت پیچیده بودن توزیع پسین، مشکل اساسی بر سر راه

[۲۱] Local Likelihood Method
[۲۲] Wavlate
[۲۳] Prior
[۲۴] Posterior

استنباط بیزی است و باید آن‌ها را به صورت تقریبی محاسبه کرد. رهیافت بیزی برای تقریب انتگرال‌های با بعد بالا از روش‌های مبتنی بر نمونه‌گیری (نمونه‌گیری از توزیع پسین) استفاده می‌کند. از جمله این روش‌ها می‌توان به روش مونت کارلوی زنجیر مارکوفی[۲۵] $(MCMC)$ اشاره کرد.

۹.۱ روش‌های نمونه‌گیری MCMC

در روش‌های انتگرال‌گیری مونت کارلویی، هدف، ارائه تقریبی مناسب برای انتگرال‌های به شکل

$$\tau = \int h(x) f(x) \mathrm{d}x$$

می‌باشد.

۱.۹.۱ الگوریتم متروپولیس-هستینگز

الگوریتم متروپولیس-هستینگز[۲۶] مثالی از مجموعه روش‌های نمونه‌گیری MCMC است. این الگوریتم، با شرط داشتن توزیع پسین توام $\pi(\theta|\mathbf{x})$، مجموعه‌ای از متغیرهای تصادفی $\{\theta^{(t)}\}$، برای $t = ۱, ۲, \cdots$، را به‌طور پی‌درپی تولید می‌کند. برای تولید این نمونه‌ها، این الگوریتم، ما را ملزم به تعیین چگالی پیشنهادی $P(\theta^{(t+۱)}, \theta^{(t)})$ می‌کند. این تابع، تابع چگالی احتمال θ در زمان $t+۱$، با توجه به مقدار آن در زمان t است. بنابراین عملکرد الگوریتم متروپولیس-هستینگز به صورت زیر بیان می‌شود:

با شرط داشتن $\theta^{(t)}$،

۱. نمونه $\theta^{(t+۱)} \sim P(\theta^{(t+۱)}, \theta^{(t)})$ را تولید کن.

۲. قرار بده

$$\theta^{(t)} = \begin{cases} \theta^{(t+۱)} & \text{با احتمال } P(\theta^{(t)}, \theta^{(t+۱)}) \\ \theta^{(t)} & \text{با احتمال } ۱ - P(\theta^{(t)}, \theta^{(t+۱)}) \end{cases}$$

به‌طوری که

$$P(\theta^{(t)}, \theta^{(t+۱)}) = \min \left\{ ۱, \frac{f(\mathbf{x}|\theta^{(t+۱)})\pi(\theta^{(t+۱)})}{f(\mathbf{x}|\theta^{(t)})\pi(\theta^{(t)})} \frac{P(\theta^{(t)}, \theta^{(t+۱)})}{P(\theta^{(t+۱)}, \theta^{(t)})} \right\}$$

که به آن احتمال پذیرش متروپولیس-هستینگز می‌گویند.

[۲۵] Markov Chain Monte Carlo
[۲۶] Metropolise-Hastings Algorithm

۲.۹.۱ نمونه‌گیر گیبز

بیش‌ترین کاربرد نمونه‌گیر گیبز[۲۷] در مدل‌های بیزی، زمانی است که توزیع‌های شرطی کامل پارامترهای مورد علاقه به صورت بسته وجود دارند. برای تولید نمونه از توزیع پسین توام $\pi(\boldsymbol{\theta}|\mathbf{x})$، که در آن $\boldsymbol{\theta} = (\theta_۱, \cdots, \theta_B)$ بردار پارامترهای مدل است، ابتدا چگالی‌های شرطی کامل را به‌دست می‌آوریم. این چگالی شرطی عبارت‌اند از

$$\begin{aligned}
&\pi(\theta_۱|\mathbf{x}, \theta_۲, \cdots, \theta_B),\\
&\pi(\theta_۲|\mathbf{x}, \theta_۱, \theta_۳, \cdots, \theta_B),\\
&\vdots\\
&\pi(\theta_B|\mathbf{x}, \theta_۱, \cdots, \theta_{B-۱}).
\end{aligned}$$

نمونه‌گیر گیبز با فرض داشتن $\boldsymbol{\theta^{(t)}} = (\theta_۱^{(t)}, \cdots, \theta_B^{(t)})$ در مرحله tام، $t = ۱, ۲, \cdots$، برای تولید یک زنجیر مارکوف از توزیع پسین توام $\pi(\boldsymbol{\theta}|\mathbf{x})$، دنباله‌ای از نمونه‌ها را به صورت زیر تولید می‌کند:

$$\begin{aligned}
۱ &: \theta_۱^{(t+۱)} \sim \pi(\theta_۱|\mathbf{x}, \theta_۲^{(t)}, \cdots, \theta_B^{(t)}),\\
۲ &: \theta_۲^{(t+۱)} \sim \pi(\theta_۲|\mathbf{x}, \theta_۱^{(t+۱)}, \theta_۳^{(t)}, \cdots, \theta_B^{(t)}),\\
&\vdots\\
B &: \theta_B^{(t+۱)} \sim \pi(\theta_B|\mathbf{x}, \theta_۱^{(t+۱)}, \cdots, \theta_{B-۱}^{(t+۱)}).
\end{aligned}$$

در صورتی‌که برخی از چگالی‌های شرطی کامل صورت پیچیده داشته باشند و نتوان به صورت مستقیم و ساده از آن‌ها نمونه تولید کرد، برای تولید از آن‌ها می‌توان از الگوریتم متروپولیس–هستینگز استفاده کرد که به روش نمونه‌گیری حاصل الگوریتم نمونه‌گیری متروپولیس–هستینگز درون گیبز[۲۸] می‌گویند (رابرت و کسلا، ۲۰۰۴).

[۲۷] Gibbs Sampler

[۲۸] Metropolis-Hastings Within Gibbs

فصل ۲

مدل‌های نرخ شفایافتگی غیرآمیخته

در این فصل به معرفی انواع مدل‌های نرخ شفایافتگی می‌پردازیم. در ادامه ویژگی‌های هر مدل مورد بررسی قرار خواهد گرفت و در انتها با یک لم شرایط شناسایی‌پذیری مدل غیرآمیخته را بیان می‌کنیم:

۱.۲ مدل‌های نرخ شفایافتگی

همان‌طور که در فصل قبل بیان شد تحلیل بقا مجموعه‌ای از روش‌های تحلیل داده‌ها است که متغیر پاسخ در آن‌ها، زمان وقوع (طول عمر) یک پدیده است. این پدیده‌ها می‌توانند ماهیت‌های متفاوتی داشته باشند، اما معمولا مرگ یک انسان به سبب یک بیماری خاص مثل سرطان است. پیشرفت‌های چشم‌گیر پزشکی در درمان بیماری (سرطان) موجب شده است، در صورت درمان به‌موقع، برخی از بیماران از طول عمر طبیعی برخوردار شوند. به عبارت دیگر، برخی از بیماران نسبت به بیماری مورد نظر شفایافته (مصون) هستند. بنابراین یکی از پذیره‌های اساسی تحلیل بقا که احتمال رخداد پیشامد را برای تمامی بیماران یک در نظر می‌گیرد، دچار خدشه می‌شود. در این موارد، یک مدل بقای مطلوب، مدلی است که بیماران شفایافته و سایر را به تفکیک مدل‌بندی کند.

در تحلیل داده‌های بقا طولانی مدت، با توجه به پیشرفت‌های حاصل در درمان بیمارها، در انتهای مطالعه با تعداد زیادی از موارد سانسورشده مواجه خواهیم شد. اگر طول مدت مطالعه به اندازه کافی طولانی باشد، استفاده از مدل‌های نرخ شفایافتگی [۱] نتایج بهتری خواهد داشت. به‌طورکلی مدل‌های نرخ شفایافتگی به دو دسته‌ی کلی مدل‌های نرخ شفایافتگی آمیخته[۲] و مدل‌های نرخ شفایافتگی غیرآمیخته[۳] تقسیم می‌شوند. در کاربردهای مختلف، مدل نرخ شفایافتگی آمیخته از استقبال بیشتری برخوردار بوده است. اما مدل نرخ شفایافتگی غیرآمیخته بر اساس مکانیسم زیستی توسعه بیماری‌های معمول این نوع از داده‌های بقا (مثل برخی از سرطان‌ها)، تعریف شده است. بنابراین از تفسیرپذیری بالاتری برخوردار است.

---

[۱] Cure Rate Models

[۲] Mixture Cure Models

[۳] Non-Mixture Cure Models

۲.۲ مدل نرخ شفایافتگی آمیخته

اولین خانواده از مدل‌های نرخ شفایافتگی، نخستین بار توسط بوگ (۱۹۴۹) مطرح شد و برکسون و گیج در سال (۱۹۵۲) آن را بسط و توسعه دادند. این رده از مدل‌ها، مدل‌های نرخ شفایافتگی آمیخته نام دارد. در این مدل‌ها فرض بر این است که جامعه از دو دسته بیمار ناهمگن تشکیل شده است. برای مثال، در بیماری سرطان پستان، به‌عنوان شایع‌ترین بیماری در میان زنان که گاهی اوقات در میان مردان نیز دیده می‌شود، دسته اول بیمارانی هستند که در معرض رخداد مرگ قرار دارند و در مدت زمان معقولی پس از شروع مطالعه مرگ را تجربه خواهند کرد. دسته دوم بیمارانی هستند که در معرض رخداد حادثه قرار ندارند. البته این به معنای تجربه نکردن مرگ تا زمان بی‌نهایت نیست. بلکه تعبیری ریاضی است، به این مفهوم که در زمان معقولی رخداد مرگ را تجربه نمی‌کنند، یا مرگ آن‌ها به علت دیگری جز سرطان پستان است. دسته اول را غیرمصون[۴] (شفانیافته) و دسته دوم را مصون (شفایافته) می‌نامند.

اگر T، زمان رخداد پیشامد مورد نظر باشد، تابع مشخصه Y را به‌صورت زیر در نظر بگیرید:

$$Y = \begin{cases} ۱ & \text{شفانیافته} \\ ۰ & \text{شفایافته} \end{cases}$$

هم‌چنین اگر $۱ - P(\mathbf{x})$ نسبت شفایافتگی افراد تحت متغیر کمکی‌های $\mathbf{x}$ باشد، آن‌گاه

$$P(Y = ۱) = P(\mathbf{x}).$$

در نتیجه

$$P(Y = ۰) = ۱ - P(\mathbf{x}).$$

تابع بقای افراد غیرمصون به صورت

$$S_u(t|\mathbf{z}) = P(T > t|Y = ۱, \mathbf{z})$$

تعریف می‌شود که در آن $\mathbf{z}$ بردار متغیرهای کمکی مختص تابع بقای افراد شفایافته است. تابع بقای افراد مصون نیز با $S_c(t|\mathbf{z})$ نشان داده شده و به صورت

$$S_c(t|\mathbf{z}) = P(T > t|Y = ۰, \mathbf{z})$$

تعریف می‌شود و همواره برابربا ۱است. بنابراین تابع بقای جامعه که آن‌را با $S_p(t|\mathbf{x}, \mathbf{z})$ نشان می‌دهند به‌صورت

$$S_p(t|\mathbf{x}, \mathbf{z}) = P(\mathbf{x})S_u(t|\mathbf{z}) + (۱ - P(\mathbf{x})) \qquad (۱.۲)$$

[۴] Uncured

تعریف می‌شود. برکسون و گیج، $P(\mathbf{x})$ را به‌عنوان یک ثابت نامعلوم و $S_u(t|\mathbf{z})$ را مدلی پارامتری در نظر گرفتند.

تابع احتمال و تابع نرخ مخاطره در این مدل به‌صورت زیر است:

$$f(t|\mathbf{z}) = P(\mathbf{x}) f_u(t|\mathbf{z})$$

و

$$h(t|\mathbf{z}) = \frac{P(\mathbf{x}) f_u(t|\mathbf{z})}{(۱ - P(\mathbf{x})) + P(\mathbf{x}) f_u(t|\mathbf{z})}.$$

فیرول (۱۹۸۲, ۱۹۸۶) نسبت شفانیافتگی را با یک مدل رگرسیون لجستیک به‌صورت زیر بسط داد:

$$P(\mathbf{x}) = \frac{\exp(\mathbf{x}^T \boldsymbol{\beta})}{۱ + \exp(\mathbf{x}^T \boldsymbol{\beta})}.$$

رگرسیون لجستیک، یکی از رایج‌ترین مدل‌های مورد استفاده برای طبقه‌بندی و پیش‌بینی در داده‌های بقا است.

تابع درستنمایی مدل شفایافتگی آمیخته

فرض کنید t_i زمان بقای مشاهده‌شده برای i -امین بیمار باشد، و تابع مشخصه y_i را به‌صورت زیر در نظر بگیرید:

$$y_i = \begin{cases} ۱ & \text{اگر } t_i \text{ سانسور شده است} \qquad i = ۱, \cdots, n \\ ۰ & \text{در غیر این صورت} \end{cases}$$

در این صورت تابع درستنمایی مدل به‌صورت زیر است:

$$L = \prod_{i=۱}^{n} [P(\mathbf{x}) f_u(t_i|\mathbf{z})]^{y_i} [(۱ - P(\mathbf{x})) + P(\mathbf{x}) S_u(t_i|\mathbf{z})]^{۱-y_i}.$$

بسیاری از محققین این عرصه از جمله کوک و چن (۱۹۹۲)، تیلور (۱۹۹۵)، پنگ و دیر (۲۰۰۰)، سای و تیلور (۲۰۰۰)، لی و تیلور (۲۰۰۰)، پنگ (۲۰۰۳) و ژانگ، پنگ و لی (۲۰۱۳)، مدل نرخ شفایافتگی آمیخته را در قالب یک مدل نیمه‌پارامتری ارائه کردند و برای اولین بار وانگ و لیانگ (۲۰۱۲) این مدل را به‌صورت کاملا ناپارامتری معرفی کردند.

۳.۲ مدل نرخ شفایافتگی غیرآمیخته

دومین خانواده از مدل‌های نرخ شفایافتگی، مدل‌های نرخ شفایافتگی غیرآمیخته یا مدل‌های زمان پیشرفت[۵] هستند. این مدل‌ها برای جوامعی تعریف شده‌اند که، تابع توزیع زمان بقای کل افراد جامعه

[۵] Promotion Time Cure Model

در آن‌ها کامل نیست. به این معنی که مقدار تابع توزیع تجمعی به یک نمی‌رسد. در این جوامع، میزان اختلاف تابع توزیع تجمعی با مقدار یک، نسبت افراد شفایافته را نشان می‌دهد. اولین مدل نرخ شفایافتگی غیرآمیخته، توسط یاکولف و زودایکف (۱۹۹۶) و زودایکف (۱۹۹۸) ارائه شد و سپس چن و همکاران (۱۹۹۹) مکانیسمی بر اساس توزیع متغیر پنهان[۶] برای آن ارائه کردند. این فرآیند توسط کونر و همکاران (۲۰۰۷) بسط و توسعه یافت.

تابع بقا برای متغیر تصادفی زمان T، در این رده از مدل‌ها، به صورت زیر به‌دست می‌آید:

- فرض کنید N بیانگر تعداد سلول‌های سرطانی از توزیع پواسون با پارامتر θ باشد. یعنی

$$N \sim Pois(\theta)$$

- فرض کنید $\{Y_1, \cdots, Y_N\}$ دنباله‌ای از متغیرهای تصادفی مستقل و هم‌توزیع با تابع توزیع تجمعی $F(t) = 1 - S(t)$ هستند که به N نیز وابسته نیستند.

- متغیر تصادفی زمان بقا به‌صورت زیر تعریف می‌شود:

$$T = \begin{cases} \infty & N = 0 \text{ اگر} \\ Y_{(i)} & N \geq 1 \text{ اگر} \end{cases}$$

که در آن $Y_{(1)} \leq \cdots \leq Y_{(i)} \leq \cdots \leq Y_{(N)}$. در واقع زمان بازگشت مجدد بیماری به‌صورت

$$T = \min\{Y_i, \quad 1 \leq i \leq N\}$$

به‌دست می‌آید.

بنابراین تابع بقای جامعه به صورت زیر قابل محاسبه است:

$$\begin{aligned} S_p(t) &= P(N = 0) + P(Y_1 > t, \cdots, Y_N > t, N \geq 1) \\ &= P(N = 0) + P(Y_1 > t, \cdots, Y_N > t | N \geq 1) P(N \geq 1) \\ &= \exp(-\theta) + \Sigma_{k=1}^{\infty} S(t)^k \frac{\theta^k \exp(-\theta)}{k!} \\ &= \sum_{k=0}^{+\infty} \frac{(S(t)\theta^k e^{-\theta})}{k!} = \frac{e^{-\theta}}{e^{-\theta S(t)}} = \exp[-\theta(1 - S(t)]. \end{aligned} \tag{۲.۲}$$

در این مدل، نسبت افراد شفایافته در داخل مدل بوده و می‌توان آن‌را به‌طور مستقیم با گرفتن حد بقا در

[۶] Latent Variable

بینهایت، محاسبه کرد:

$$\begin{aligned} P(N=۰) &= \lim_{t\to\infty} S_p(t) \\ &= \lim_{t\to\infty} \exp(-\theta(۱ - S(t))) \\ &= \lim_{t\to\infty} \exp(-\theta + \theta S(t)) \\ &= \exp(-\theta) \lim_{t\to\infty} \exp(\theta S(t)) \\ &= \exp(-\theta) \exp(\theta \lim_{t\to\infty} S(t)) \end{aligned}$$

و با توجه به این‌که $S(\infty) = ۰$، آن‌گاه

$$P(N=۰) = \exp(-\theta) \exp(۰) = \exp(-\theta).$$

با در نظر گرفتن مدل مخاطره کاکس به‌صورت

$$F(t|\mathbf{z}) = ۱ - S_۰(t)^{\exp(\mathbf{z}^T\gamma)}$$

که در آن $S_۰(t)$ تابع بقا پایه است، رابطه (۲.۲) با دو بردار کمکی $\mathbf{x}$ و $\mathbf{z}$ و متغیر پنهان با میانگین $\exp(\beta_۰ + \mathbf{x}^T\boldsymbol{\beta})$ به‌صورت زیر درمی‌آید:

$$\begin{aligned} S_p(t|\mathbf{x},\mathbf{z}) &= \exp(-\theta(\mathbf{x})F(t|\mathbf{z})) \\ &= \exp(-\exp(\beta_۰ + \mathbf{x}^T\boldsymbol{\beta})(۱ - S_۰(t)^{\exp(\mathbf{z}^T\gamma)})). \end{aligned} \qquad (۳.۲)$$

در سال‌های اخیر توزیع‌های متفاوتی برای متغیر پنهان N، در نظر گرفته شده‌اند. برای مثال، اگر توزیع متغیر پنهان N برنولی در نظر گرفته شود، مدل شفایافته آمیخته حاصل می‌شود. هم‌چنین کونر و همکاران (۲۰۰۷) توزیع هندسی، برنولی و دوجمله‌ای، رودریگرز و همکاران (۲۰۰۹) توزیع پواسون کانوی-ماکسول[۷]، بورگر و همکاران (۲۰۱۲) توزیع سری‌های توانی، باغستانی و همکاران (۲۰۱۳) توزیع تعمیم‌یافته پواسون لیندلی و رحیم‌زاده و همکاران (۲۰۱۳) توزیع فوق هندسی تعمیم‌یافته و دوجمله‌ای منفی را در مدل‌های شفایافته زمان پیشرفت ارائه کردند. شایان ذکر است توزیع متغیر پنهان، به منظور تبیین احتمال شفایافتگی، می‌تواند هر یک از توزیع‌های گسسته که دارای جرم احتمال در نقطه صفر هستند را شامل شود. به نظر می‌رسد رابطه نزدیکی بین توزیع‌هایی که در تحلیل داده‌های شمارشی و مدل‌های شفایافته زمان پیشرفت مورد استفاده قرار می‌گیرد، وجود دارد. مشکل دیگری که علاوه بر بیش‌پراکنشی در تحلیل داده‌های بقا وجود دارد، چوله بودن داده‌ها است که دلیلی دیگر برای ارائه توزیع‌های متفاوت برای حل این مشکل است.

[۷] Conway-Maxwell-Poisson Distribution

۴.۲ ویژگی‌های مدل نرخ شفایافتگی غیرآمیخته

مدل نرخ شفایافتگی غیرآمیخته با تابع بقا جامعه در رابطه (۳.۲) معرفی شد را در نظر بگیرید. تابع نرخ مخاطره جامعه با $h_p(t|\mathbf{x},\mathbf{z})$ نشان داده می‌شود و به‌صورت زیر تعریف می‌شود:

$$\begin{aligned} h_p(t|\mathbf{x},\mathbf{z}) &= \frac{-\mathrm{d}(\ln(S_p(t|\mathbf{x},\mathbf{z})))}{\mathrm{d}t} \\ &= \frac{-\mathrm{d}(\ln(\exp(-\exp(\beta_\circ + \mathbf{x}^T\boldsymbol{\beta})(1 - S_\circ(t)^{\exp(\mathbf{z}^T\boldsymbol{\gamma})}))))}{\mathrm{d}t} \\ &= \frac{-\mathrm{d}(-\exp(\beta_\circ + \mathbf{x}^T\boldsymbol{\beta})(1 - S_\circ(t)^{\exp(\mathbf{z}^T\boldsymbol{\gamma})}))}{\mathrm{d}t}. \end{aligned}$$

با توجه به این نکته که $\frac{\mathrm{d}(g(x)^n)}{\mathrm{d}x} = ng'(x)g(x)^{n-1}$ و $f(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t}$، داریم

$$\begin{aligned} h_p(t|\mathbf{x},\mathbf{z}) &= -(-\exp(\beta_\circ + \mathbf{x}^T\boldsymbol{\beta})(\exp(\mathbf{z}^T\boldsymbol{\gamma})f_\circ(t)S_\circ(t)^{\exp(\mathbf{z}^T\boldsymbol{\gamma})-1})) \\ &= \frac{\exp(\beta_\circ + \mathbf{x}^T\boldsymbol{\beta})\exp(\mathbf{z}^T\boldsymbol{\gamma})f_\circ(t)S_\circ(t)^{\exp(\mathbf{z}^T\boldsymbol{\gamma})}}{S_\circ(t)} \\ &= \theta(\mathbf{x})f(t|\mathbf{z}). \end{aligned}$$

نسبت مخاطره جامعه برای دو گروه ۱ و ۲ به‌صورت زیر به‌دست می‌آید:

$$\begin{aligned} HR_p &= \frac{h_p(t|\mathbf{x}_1,\mathbf{z}_1)}{h_p(t|\mathbf{x}_2,\mathbf{z}_2)} \\ &= \frac{\frac{\exp(\beta_\circ+\mathbf{x}_1^T\boldsymbol{\beta})\exp(\mathbf{z}_1^T\boldsymbol{\gamma})f_\circ(t)S_\circ(t)^{\exp(\mathbf{z}_1^T\boldsymbol{\gamma})}}{S_\circ(t)}}{\frac{\exp(\beta_\circ+\mathbf{x}_2^T\boldsymbol{\beta})\exp(\mathbf{z}_2^T\boldsymbol{\gamma})f_\circ(t)S_\circ(t)^{\exp(\mathbf{z}_2^T\boldsymbol{\gamma})}}{S_\circ(t)}} \\ &= \frac{\exp(\beta_\circ + \mathbf{x}_1^T\boldsymbol{\beta})\exp(\mathbf{z}_1^T\boldsymbol{\gamma})S_\circ(t)^{\exp(\mathbf{z}_1^T\boldsymbol{\gamma})}}{\exp(\beta_\circ + \mathbf{x}_2^T\boldsymbol{\beta})\exp(\mathbf{z}_2^T\boldsymbol{\gamma})S_\circ(t)^{\exp(\mathbf{z}_2^T\boldsymbol{\gamma})}} \\ &= \exp((\mathbf{x}_1^T - \mathbf{x}_2^T)\boldsymbol{\beta})\exp((\mathbf{z}_1^T - \mathbf{z}_2^T)\boldsymbol{\gamma})S_\circ(t)^{(\exp(\mathbf{z}_1^T\boldsymbol{\gamma})-(\exp(\mathbf{z}_2^T\boldsymbol{\gamma}))}. \end{aligned}$$

اگر $\mathbf{z}_2 = \mathbf{z}_1$، آن‌گاه

$$HR_p = \frac{h_p(t|\mathbf{x}_1,\mathbf{z}_1)}{h_p(t|\mathbf{x}_2,\mathbf{z}_2)} = \exp((\mathbf{x}_1^T - \mathbf{x}_2^T)\boldsymbol{\beta})$$

و این مسأله نشان می‌دهد که نسبت مخاطره‌ها به زمان t بستگی ندارد.

تابع بقا برای افراد شفانیافته به‌صورت زیر به‌دست می‌آید:

$$\begin{aligned} S_u(t|\mathbf{x},\mathbf{z}) &= P(T > t|N \geq 1,\mathbf{x},\mathbf{z}) \\ &= \frac{S_p(t|\mathbf{x},\mathbf{z}) - P(N=\circ)}{1 - P(N=\circ)} \\ &= \frac{S_p(t|\mathbf{x},\mathbf{z}) - \exp(-\theta(\mathbf{x}))}{1 - \exp(-\theta(\mathbf{x}))}. \end{aligned}$$

هم‌چنین تابع مخاطره برای افراد شفانیافته نیز به‌صورت زیر است:

$$\begin{aligned} h_u(t|\mathbf{X},\mathbf{Z}) &= \frac{-\mathrm{d}(\ln(\mathrm{S_u}(\mathrm{t}|\mathbf{X},\mathbf{Z})))}{\mathrm{d}t} \\ &= \frac{S_p(t|\mathbf{X},\mathbf{Z})}{S_p(t|\mathbf{X},\mathbf{Z}) - \exp(-\theta(\mathbf{X}))} h_p(t|\mathbf{X},\mathbf{Z}) \\ &= \frac{h_p(t|\mathbf{X},\mathbf{Z})}{P(T<+\infty|T>t,\mathbf{X},\mathbf{Z})}. \end{aligned} \tag{۴.۲}$$

همان‌طور که ملاحظه می‌شود رابطه (۴.۲)، ساختار مخاطرات متناسب را ندارد. به عبارت دیگر ویژگی مخاطرات متناسب برای کل افراد تحت مطالعه برقرار است نه فقط افراد شفانیافته.

۵.۲ شناسایی‌پذیری مدل

تعریف ۱.۵.۲. (لیو و چن، ۲۰۰۹). در مدل $F(x;\beta)$ پارامتر β را شناسایی‌پذیر گویند، هرگاه به ازای هر β_1 و β_2 از فضای پارامتر که $F(x;\beta_1) = F(x;\beta_2)$، آن‌گاه $\beta_1 = \beta_2$.

در نظریه تحلیل بقا، پیگیری یک مطالعه به‌اندازه کافی طولانی است اگر بزرگترین زمان سانسور بیشتر از بزرگترین زمان شکست باشد؛ یعنی در صورتی که زمان پیگیری هر واحد غیرمصون (شفانیافته) به‌اندازه کافی طولانی باشد تا شکست آن را ببیند. اما در عمل، اگر منحنی بقای جامعه در دم راست خود صاف باشد، فرض پیگیری کافی منطقی به‌نظر می‌رسد.

شرط‌های لازم برای شناسایی‌پذیر بودن مدل نرخ شفایافتگی غیرآمیخته در لم زیر ارائه شده‌اند.

لم ۲.۵.۲. (برمهورست و لامبرت، ۲۰۱۶). فرض کنید

A_1. بردار متغیرهای کمکی $\mathbf{z}$ شامل هیچ‌گونه داده‌ی سانسورشده‌ای نیست.

A_2. $\mathbf{X}^T\mathbf{X}$ و $\mathbf{Z}^T\mathbf{Z}$ دو ماتریس با رتبه کامل هستند که $\mathbf{x}$ و $\mathbf{z}$ به‌ترتیب متغیرهای کمکی آن‌ها هستند.

A_3. $F_0(t) = 1 - S_0(t)$ یک تابع توزیع تجمعی سره[۸] است. یعنی

$$\lim_{t\to\infty} S_0(t) = 0.$$

آن‌گاه تحت شرط‌های $A_1 - A_3$ می‌توان گفت

(۱) اگر زمان مطالعه به‌طور کافی طولانی باشد، مدل (۳.۲) شناسایی‌پذیر است.

(۲) اگر زمان مطالعه به‌قدر کافی طولانی نباشد و بردارهای $\mathbf{x}$ و $\mathbf{z}$ مستقل از هم باشند، آن‌گاه برآوردهای پارامترهای موجود در احتمال درمان و زمان شکست برای یک سلول سرطانی، شناسایی‌پذیر است.

اثبات. **برهان قسمت** (۱)

[۸] Proper

(a) فرض کنید $(\beta_\circ, \boldsymbol{\beta}, \gamma, S_\circ)$ و $(\tilde{\beta}_\circ, \tilde{\boldsymbol{\beta}}, \tilde{\gamma}, \tilde{S}_\circ)$ دو مجموعه از پارامترها هستند که در رابطه (۳.۲) صدق می‌کنند و $\hat{\chi}$ مجموعه‌ای از تمام مقادیر بردار $\mathbf{x}$ است. اگر

$$\begin{aligned} S_p(t|\mathbf{x}) &= \exp\left[-\exp(\beta_\circ + \mathbf{x}^T\boldsymbol{\beta})\left(1 - S_\circ(t)^{\exp(\mathbf{z}^T\gamma)}\right)\right] \\ &= \exp\left[-\exp(\tilde{\beta}_\circ + \mathbf{x}^T\tilde{\boldsymbol{\beta}})\left(1 - \tilde{S}_\circ(t)^{\exp(\mathbf{z}^T\tilde{\gamma})}\right)\right] \\ &= \tilde{S}_p(t|\mathbf{x}) \qquad \forall \mathbf{x} \in \chi \quad ; \quad \forall t \in [\circ, \infty] \end{aligned}$$

آن‌گاه

$$\beta_\circ = \tilde{\beta}_\circ \quad , \quad \boldsymbol{\beta} = \tilde{\boldsymbol{\beta}} \quad , \quad \gamma = \tilde{\gamma} \quad , S_\circ(t) = \tilde{S}_\circ(t) \quad \forall t \in [\circ, \infty]$$

(b) از آن‌جا که تابع نمایی یکنوا است،

$$\forall \mathbf{x} \in \chi \quad \forall t \in [\circ, \infty]$$

اگر

$$\exp(\beta_\circ + \mathbf{x}^T\boldsymbol{\beta})\left(1 - S_\circ(t)^{\exp(\mathbf{z}^T\gamma)}\right) = \exp(\tilde{\beta}_\circ + \mathbf{x}^T\tilde{\boldsymbol{\beta}})\left(1 - \tilde{S}_\circ(t)^{\exp(\mathbf{z}^T\tilde{\gamma})}\right)$$

آن‌گاه

$$\beta_\circ = \tilde{\beta}_\circ \quad , \quad \boldsymbol{\beta} = \tilde{\boldsymbol{\beta}} \quad , \quad \gamma = \tilde{\gamma} \quad , S_\circ(t) = \tilde{S}_\circ(t) \quad \forall t \in [\circ, \infty].$$

(c) از آن‌جاکه $F_\circ(t)$ تابع توزیع تجمعی سره است، با جایگذاری در روابط قبل داریم

$$F(t|\mathbf{z}) = 1 - S_\circ(t)^{\exp(\mathbf{z}^T\gamma)}$$

که یک تابع توزیع تجمعی سره است. ویژگی (b) به‌ازای $t = \infty$ نیز برقرار است. حال با توجه به این‌که $F(t|\mathbf{z})$ و $F(\tilde{t}|\mathbf{z})$ هر دو تابع توزیع تجمعی سره هستند، اگر

$$\beta_\circ + \mathbf{x}^T\boldsymbol{\beta} = \tilde{\beta}_\circ + \mathbf{x}^T\tilde{\boldsymbol{\beta}} \qquad \forall \mathbf{x} \in \chi$$

آن‌گاه

$$\beta_\circ = \tilde{\beta}_\circ, \qquad \boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}.$$

تحت A_2 مستقیما نتیجه می‌شود

$$P_\chi(\mathbf{x}^T\tilde{\boldsymbol{\beta}} = \alpha) = 1 \Rightarrow \alpha = \circ, \boldsymbol{\beta} = \circ.$$

(d) تا زمانی که بردار $\mathbf{z}$ شامل هیچ سانسوری نیست، واضح است که تحت رابطه $A_۲$ مدل مخاطره متناسب کاکس شناسایی‌پذیر است.

(e) با ادغام نتایج (c)و (d) قسمت (۱) اثبات می‌شود.

برهان قسمت (۲)

(a) اگر تومور در حال رشد از سلول سرطانی نتواند با اتمام مطالعه در زمان t تشخیص داده شود، پیگیری به‌اندازه کافی طولانی نبوده است. این حالت زمانی‌که $F_۰(t)$ خیلی کوچک (نزدیک به ۰) باشد، اتفاق می‌افتد. در این صورت

$$\begin{aligned} F(t|\mathbf{z}) &= ۱ - S_۰(t)^{\exp(\mathbf{z}^T\boldsymbol{\gamma})} \\ &= ۱ - \left[(۱ - F_۰(t))^{\exp(\mathbf{z}^T\boldsymbol{\gamma})}\right] \\ &= ۱ - \left[(۱ - \exp(\mathbf{z}^T\boldsymbol{\gamma})F_۰(t))\right] + ۰ * (F_۰(t)) \\ &\approx ۱ - \left[(۱ - \exp(\mathbf{z}^T\boldsymbol{\gamma})F_۰(t))\right] \\ &= \exp(\mathbf{z}^T\boldsymbol{\gamma})F_۰(t) \end{aligned}$$

بنابراین با استفاده از این تقریب، رابطه (۳.۲) به صورت زیر قابل بازنویسی است:

$$\begin{aligned} S_p(t|\mathbf{x},\mathbf{z}) &\approx \exp\left[-\theta(\mathbf{x})\exp(\mathbf{z}^T\boldsymbol{\gamma})F_۰(t)\right] \\ &= \exp\left[-\exp(\beta_۰ + \mathbf{x}^T\boldsymbol{\beta})\exp(\mathbf{z}^T\boldsymbol{\gamma})F_۰(t)\right] \\ &= \exp\left[-\exp(\beta_۰ + \mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T\boldsymbol{\gamma})F_۰(t)\right]. \end{aligned}$$

بنابراین نتیجه می‌گیریم اگر بردارهای $\mathbf{x}$ و $\mathbf{z}$ با هم اشتراک داشته باشند، تقریب تاثیرات متغیرهای کمکی قابل شناسایی نیست، زیرا پیگیری به‌اندازه کافی طولانی نبوده است.

(c) برآورد اثرات متغیرهای کمکی شناسایی‌پذیر است، هرگاه پیگیری مطالعه به‌اندازه کافی طولانی نیست ولی بردارهای کمکی واحد مشترک ندارند. در واقع می‌توان نوشت

$$\begin{aligned} \theta(\mathbf{x})F(t|\mathbf{z})_{|t_{max}} &\approx \exp(\beta_۰ + \mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T\boldsymbol{\gamma})F_۰(t)_{t_{max}} \\ &= \exp(\beta_۰ + \mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T\boldsymbol{\gamma} + Log(۱ - \epsilon)) \\ &= \exp(\tilde{\beta}_۰ + \mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T\boldsymbol{\gamma}) \end{aligned}$$

که در آن

$$\epsilon = F_۰(+\infty) - F_۰(t_{max}).$$

در نتیجه برآوردی از تاثیر متغیرها در احتمال درمان و زمان شکست برای یک سلول سرطانی، تحت تاثیر قرار نمی‌گیرد. بنابراین اگر بردارهای $\mathbf{x}$ و $\mathbf{z}$ اشتراک نداشته باشند، پارامترهای مذکور می‌توانند قابل شناسایی باشند.

□

فصل ۳

تحلیل بیزی مدل نرخ شفایافتگی غیرآمیخته

۱.۳ مقدمه

تحلیل بیزی مدل‌های نرخ شفایافتگی غیرآمیخته با تابع بقای نیمه‌پارامتری از سال ۱۹۸۰ بسط و توسعه یافت. روش درستنمایی جزئی که روش معمول برآورد پارامترها در مدل مخاطره متناسب کاکس است، در این مدل کاربرد ندارد. برای برآورد پارامترها باید از تابع درستنمایی کامل استفاده شود که نیازمند برآورد تابع مخاطره پایه است. این رهیافت چالش‌های خاص خود را دارد. بنابراین استفاده از دیدگاه استنباط بیزی در این مدل‌ها می‌تواند جانشین مناسب‌تری باشد. در این فصل دیدگاه مدل‌بندی بیزی در مدل‌های نرخ شفایافتگی غیرآمیخته را بیان می‌کنیم. برای مدل‌بندی ناپارامتری تابع بقای پایه از P-اسپلاین‌ها استفاده خواهیم کرد. بنابراین ابتدا مقدمه‌ای بر اسپلاین‌ها را مطرح می‌کنیم.

۲.۳ اسپلاین

در آمار، هموار کردن[۱] یک نمودار پراکنش، استفاده از الگوهای مهم داده‌ها در ایجاد یک تابع تقریبی است. یک هموارکننده تابعی برای خلاصه‌سازی متغیر پاسخ به‌عنوان تابعی از متغیرهای پیش‌بین $X_1, \cdots, X_k$ است. یکی از خصوصیات مهم هموارکننده‌ها، ماهیت ناپارامتری آن‌ها است. هموار کننده‌ها همواره صورت تابعی ساده‌ای از وابستگی Y به $X_1, X_2, \cdots, X_k$ را ارائه می‌دهند. به همین دلیل اغلب در مدل رگرسیون، ساختار ناپارامتری دارند. برآورد تولیدشده توسط یک هموارکننده را یک هموار می‌نامیم.

فرض کنید $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ مقادیر متغیر پاسخ Y و $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$ مقادیر متغیر پیش‌بین باشند. یک هموارکننده، تابعی از $\mathbf{x}$ و $\mathbf{y}$ است که برازش مناسبی در راستای خلاصه‌سازی داده‌ها ایجاد می‌کند. یکی از انواع هموارکننده‌های نمودار پراکنش، اسپلاین‌ها هستند. اسپلاین توسط شخصی به نام شوئنبرگ در سال ۱۹۴۶ معرفی شد.

---

[۱] Smoothing

اسپلاین یا منحنی منعطف، از یک نوار بلند تشکیل شده است که شامل نقاط ثابت و یک منحنی هموار است که هدف آن عبور از همه‌ی نقاط برای انتقال از یک نقطه به نقطه دیگر است. ریشه‌ی کلمه اسپلاین در کارهای چوبی بوده و از لحاظ واژه‌شناسی به‌خاطر ارتباط کلمه‌ی اسپلاین با کلمه‌ی باریکه‌ی چوب است که بعدها صنعتگران و کارگران ماهر، اسپلاین‌ها را از لاستیک، استیل و هر ماده کشپار ساختند. برای کشتی‌سازی اغلب از چوب‌های باریک، بلند و انعطاف‌پذیر استفاده می‌کردند که به آن‌ها اسپلاین گفته می‌شد. از اسپلاین‌ها برای خم کردن چوب برای پیانو و ویولون استفاده شده است. به دلایل فیزیکی، منحنی‌های اسپلاین تقریباً چند ضابطه‌ای‌هایی مرتبه‌ی سوم با مشتق دوم پیوسته‌اند.

تابع توانی بریده

تابع توانی بریده[۲] از مرتبه‌ی p را که در نقطه‌ی t بریده شده باشد، به‌صورت زیر تعریف می‌کنیم:

$$(x-t)_+^p = (x-t)^p I(x>t)$$

که مقدار حقیقی t را گره[۳] می‌نامند و I تابع نشانگر به‌صورت

$$I(x>t) = \begin{cases} ۱ & x>t \\ ۰ & \text{سایر نقاط} \end{cases}$$

است.

اسپلاین مرتبه‌ی اول

فاصله‌ی $[a,b]$ را به‌صورت اجتماع دو بازه‌ی $[a,t]\cup[t,b]$ در نظر بگیرید. یک تابع اسپلاین از مرتبه‌ی یک، تابعی است که در هر یک از بازه‌های مذکور، یک چندجمله‌ای از مرتبه‌ی یک (خط راست) باشد و در نقطه‌ی t نیز به‌هم پیوسته باشند. این تابع را می‌توان به‌صورت رابطه‌ی زیر نوشت:

$$f(x) = \begin{cases} \phi_۰ + \phi_۱ x & x \le t \\ \phi'_۰ + \phi'_۱ x & x > t \end{cases} \qquad (۱.۳)$$

با اعمال شرط پیوستگی تابع f در نقطه‌ی $x=t$ داریم

$$\phi_۰ + \phi_۱ t = \phi'_۰ + \phi'_۱ t$$
$$\phi'_۰ = \phi_۰ + \phi_۱ t - \phi'_۱ t.$$

با جایگذاری $\phi'_۰$ در ضابطه‌ی دوم f داریم

$$\phi'_۰ + \phi'_۱ x = \phi_۰ + \phi_۱ t - \phi'_۱ t + \phi'_۱ x = \phi_۰ + \phi_۱ t + \phi'_۱ (x-t)$$

[۲] Truncated Power Function
[۳] Knot

بنابراین

$$f(x) = \begin{cases} \phi_0 + \phi_1 x & x \le t \\ \phi_0 + \phi_1 t + \phi'_1 (x - t) & x > t \end{cases}$$

که می‌توان این تابع دو ضابطه‌ای را به‌صورت رابطه‌ی زیر نوشت:

$$f(x) = \phi_0 + \phi_1 x + \phi_2 (x - t)_+ \qquad (۲.۳)$$

که $\phi_2 = \phi'_1 - \phi_1$ تفاوت بین شیب خطوط اول و دوم رابطه‌ی (۱.۳) را بیان می‌کند. در واقع می‌توان گفت که تابع f، ترکیبی خطی از توابع پایه‌ای ۱، x و $(x - t)_+$ است.

به همین ترتیب اگر در فاصله‌ی $[a, b]$ تعداد دو گره $t_1 < t_2$ وجود داشته باشند، می‌توان فاصله‌ی $[a, b]$ را به‌صورت اجتماع $[a, t_1] \cup [t_1, t_2] \cup [t_2, b]$ در نظر گرفت و تابع اسپلاین مرتبه‌ی یک را به طریق مشابه رابطه‌ی (۲.۳) به‌صورت زیر نوشت:

$$f(x) = \phi_0 + \phi_1 x + \phi_2 (x - t_1)_+ + \phi_3 (x - t_2)_+.$$

نهایتا فرض کنید که بتوان فاصله‌ی $[a, b]$ را به‌صورت اجتماع فاصله‌های

$$[a, t_1], [t_1, t_2], \cdots, [t_k, b]$$

نوشت که برای گره‌های ذکرشده داریم $t_1 < \cdots < t_k$. در این‌صورت تابع اسپلاین مرتبه‌ی یک را می‌توان به‌صورت رابطه‌ی

$$f(x) = \phi_0 + \phi_1 x + \sum_{j=1}^{k} \phi_{j+1} (x - t_j)_+$$

بیان کرد که ترکیبی خطی از توابع پایه ۱، x، $(x - t_1)_+$، $\cdots$ و $(x - t_k)_+$ به شمار می‌آید.

اسپلاین مرتبه‌ی p

توابع اسپلاین را برای مراتب بالاتر نیز می‌توان به‌طور مشابه تعریف کرد. تابع $f(\cdot)$ را یک اسپلاین مرتبه‌ی p نامند اگر در هر یک از زیر بازه‌های $[a, t_1], [t_1, t_2], \cdots, [t_k, b]$ یک چندجمله‌ای از مرتبه‌ی p باشد به‌طوری‌که $f(\cdot)$ و مشتق‌های آن از مراتب ۱ تا $p - 1$، در گره‌ها پیوسته باشند. تابع اسپلاین مرتبه‌ی p را می‌توان به‌صورت زیر نوشت:

$$f(x) = \phi_0 + \phi_1 x + \phi_2 x^2 + \cdots + \phi_p x^p + \sum_{j=1}^{k} \phi_{p+j} (x - t_j)_+^p. \qquad (۳.۳)$$

تابع اسپلاین مرتبه‌ی p، ترکیبی خطی از یک چندجمله‌ای مرتبه‌ی p وتوابع توانی بریده‌شده است. به‌عبارت دیگر مجموعه‌ی $\{1, x, x^2, \cdots, x^p, (x - t_1)_+^p, \cdots, (x - t_k)_+^p\}$ یک پایه برای فضای اسپلاین‌های مرتبه‌ی p با گره‌های t_1، t_2، $\cdots$ و t_k است.

اسپلاین مکعبی

اگر $p = ۳$، آن‌گاه تابع ذکرشده در رابطه‌ی (۳.۳) را یک اسپلاین مکعبی[۴] نامند که به‌صورت

$$f(x) = \phi_۰ + \phi_۱ x + \phi_۲ x^۲ + \phi_۳ x^۳ + \sum_{j=۱}^{k} \phi_{۳+j}(x - t_j)_+^۳$$

بیان می‌شود.

خواص اسپلاین‌ها

۱. تابع $f(\cdot)$ در رابطه‌ی (۳.۳)، یک چندجمله‌ای تکه‌ای از مرتبه‌ی p، با تعداد $p + k + ۱$ ضریب است.

۲. "همواری" یکی از خواص مهم اسپلاین‌ها است. تابع اسپلاین (۳.۳) از جنس چندجمله‌ای تکه‌ای است و همان‌طور که گفته شد، این تابع دارای مشتقات پیوسته تا مرتبه‌ی $p - ۱$ است. برای سهولت، اسپلاین مرتبه‌ی p را با یک گره‌ی t به‌صورت زیر در نظر بگیرید:

$$\begin{aligned} f(x) &= \phi_۰ + \phi_۱ x + \phi_۲ x^۲ + \cdots + \phi_p x^p + \phi_{p+۱}(x - t)_+^p \\ &= P(x) + \phi_{p+۱}(x - t)_+^p. \end{aligned}$$

الف- تابع $f(\cdot)$ در گره‌ی t پیوسته است، زیرا

$$\lim_{x \to t} f(x) = P(t), \qquad f(t) = P(t).$$

بنابراین در سرتاسر بازه‌ی تعریف‌شده پیوسته خواهد بود.

ب- مشتقات تابع $f(\cdot)$ تا مرتبه‌ی $p - ۱$ نیز توابعی پیوسته هستند، زیرا

$$f^{(j)}(t) = P^{(j)}(t) \qquad j = ۱, ۲, \cdots, p - ۱$$

و از آن‌جا که $\lim_{x \to t} \phi_{p+۱} p(p - ۱) \cdots (p - j + ۱)(x - t)^{p-j} = ۰$، بنابراین

$$\lim_{x \to t} f^{(j)}(x) = P^{(j)}(t).$$

ج- مشتق مرتبه‌ی p-ام تابع $f(\cdot)$ دارای رفتار متفاوتی است. در واقع

$$f^{(p)}(t) = p! \phi_p$$

$$\lim_{x \to t} f^{(p)}(x) = p! \phi_p + p! \phi_{p+۱}.$$

[۴] Cubic Spline

رگرسیون و کمترین توان‌های دوم اسپلاین

در مدل رگرسیونی

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}$$

اگر $f(\cdot)$ را با رابطه‌ی (۳.۳) جایگزین کنیم، نتیجه‌ی حاصل را رگرسیون اسپلاین[۵] می‌نامند. یک برتری رگرسیون اسپلاین، برای روش‌های ناپارامتری و نیمه‌پارامتری این است که متعلق به رده‌ی رگرسیون خطی است. به عبارت دیگر، ضابطه تابع $f(\cdot)$ نسبت به پارامترها خطی است.

برازش یک اسپلاین به داده‌ها می‌تواند با استفاده از روش کمترین توان‌های دوم خطا انجام پذیرد. استفاده از توابع پایه‌ای بریده‌شده، منجر به برازش مدلی به صورت زیر می‌گردد:

$$\begin{aligned} y_j = {} & \phi_0 + \phi_1 x_j + \cdots + \phi_p x_j^p \\ & + \phi_{p+1}(x_j - t_1)_+^p + \cdots + \phi_{p+k}(x_j - t_k)_+^p + \epsilon_j, \qquad j = 1, \cdots, n \end{aligned}$$

که در آن ϵ_j ها در پذیره‌های کلاسیک رگرسیون صدق می‌کنند. شکل ماتریسی این عبارت را می‌توان به‌صورت زیر بیان نمود:

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{\phi} + \boldsymbol{\epsilon}$$

به‌طوری که $\boldsymbol{\phi}^T = (\phi_0, \phi_1, \cdots, \phi_{p+k})^T$ و D یک ماتریس طرح $n \times (p+k+1)$ است که درآیه‌های ستون اول آن برابر ۱، درآیه‌های p ستون بعدی، مربوط به مشاهدات متغیرهای x، x^2، $\cdots$ و x^p و درآیه‌های ستون‌های $p+2$ تا $(p+k+1)$ مربوط به مشاهده‌های حاصل از متغیرهای $(x-t_1)_+^p$، $\cdots$ و $(x-t_k)_+^p$ به ازای گره‌های از قبل تعیین شده t_1، t_2، $\cdots$ و t_k هستند. به عبارت دیگر

$$\boldsymbol{D} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p & (x_1 - t_1)_+^p & (x_1 - t_2)_+^p \cdots & (x_1 - t_k)_+^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p & (x_2 - t_1)_+^p & (x_2 - t_2)_+^p \cdots & (x_2 - t_k)_+^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p & (x_n - t_1)_+^p & (x_n - t_2)_+^p \cdots & (x_n - t_k)_+^p \end{bmatrix}.$$

با به‌کار بردن روش کمترین توان‌های دوم، برآوردگر زیر حاصل می‌شود

$$\hat{\boldsymbol{\phi}} = (\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{y}. \tag{۴.۳}$$

مشکلی که در برآوردگر (۴.۳) وجود دارد، هم‌خطی متغیرهای

$$(x - t_1)_+^p, (x - t_2)_+^p, \cdots, (x - t_k)_+^p$$

است که باعث عدم وارون‌پذیری یا بدشرطی[۶] ماتریس $\boldsymbol{D}^T\boldsymbol{D}$ خواهد شد.

[۵] Spline Regression
[۶] Ill-Condition

۱.۲.۳ B-اسپلاین

بازه‌ی $[a,b]$ و تعداد q گره

$$a = t_۰ < t_۱ < \cdots < t_q < t_{q+۱} = b$$

را در نظر بگیرید. توابع B-اسپلاین، چندجمله‌ای‌های قطعه‌ای در بازه‌ی $[a,b]$ هستند که در گره‌ها به یکدیگر متصل بوده و دارای تکیه‌گاه کوچکی می‌باشند. متداول‌ترین روش محاسبه توابع B-اسپلاین مرتبه‌ی p، به‌وسیله‌ی توابع B-اسپلاین مرتبه‌ی $p-۱$ و بر اساس رابطه‌ی بازگشتی زیر به‌دست می‌آید:

$$b_{i,p}(x) = (\frac{x - t_i}{t_{i+p} - t_i})b_{i,p-۱}(x) + (\frac{t_{i+p+۱} - x}{t_{i+p+۱} - t_{i+۱}})b_{i+۱,p-۱}(x), \quad i = -p, -p+۱, \cdots, q \tag{۵.۳}$$

که $t_{-p} = t_{-p+۱} = \cdots = t_{-۲} = t_{-۱} = t_۰ = a$ و $t_{q+۱} = b$ است. بنابراین برای یک B-اسپلاین مرتبه‌ی p می‌توان تعداد $p+q+۱$ تابع B-اسپلاین تعریف کرد.

کمترین توان‌های دوم B-اسپلاین

ابتدا تعداد $p+q+۱$ تابع B-اسپلاین مرتبه‌ی p رابطه‌ی (۵.۳) را با اندیس‌گذاری مجدد، به‌صورت

$$b_{۰,p}, b_{۱,p}, \cdots, b_{p+q,p}$$

در نظر می‌گیریم. مدل رگرسیونی B-اسپلاین به‌صورت زیر است:

$$s(x) = \sum_{j=۰}^{p+q} \phi_j b_{j,p}(x) + \boldsymbol{\epsilon}. \tag{۶.۳}$$

از آن‌جا که صورت ماتریسی این مدل به‌صورت

$$s = \boldsymbol{b}\boldsymbol{\phi} + \boldsymbol{\epsilon}$$

است، لذا برآورد کمترین توان‌های دوم $\boldsymbol{\phi}$ عبارت است از

$$\hat{\boldsymbol{\phi}} = (\boldsymbol{b}^T\boldsymbol{b})^{-۱}\boldsymbol{b}^T\boldsymbol{y}.$$

متعامد بودن توابع B-اسپلاین موجب می‌شود که ماتریس $\boldsymbol{b}^T\boldsymbol{b}$ وارون‌پذیر باشد و لذا مشکل موجود در رابطه‌ی (۴.۳) مرتفع می‌شود. لازم به ذکر است که علیرغم به‌دست آوردن این مزیت، درک توابع B-اسپلاین به آسانی درک توابع بریده توانی نیست و لذا ضرایب مدل مذکور غیر قابل تفسیر است.

۲.۲.۳ P-اسپلاین

ایلرز و مارکس (۱۹۹۶) اسپلاین‌ها را به عنوان ترکیبات خطی از توابع B-اسپلاین روی یک بازه از گره‌های هم‌فاصله نشان دادند. طبق مدل رگرسیونی B-اسپلاین در رابطه‌ی (۶.۳) برآورد کمترین توان‌های دوم $\boldsymbol{\beta}$ از مینیمم کردن رابطه‌ی زیر به‌دست می‌آید:

$$SSE = \sum_{i=۱}^{n} \left\{ y_i - \sum_{j=۰}^{p+q} \phi_j b_{j,p}(x_i) \right\}^۲ . \qquad (۷.۳)$$

فرض کنید تعداد گره‌ها نسبتا زیاد است، به‌طوری‌که منحنی برازش‌شده بیشترین تغییراتی که توسط داده‌ها نتیجه شده را نشان خواهد داد. برای اجتناب از پیچیدگی بیش از حد، سیلوان (۱۹۸۶) مجموع توان‌های دوم جریمه‌شده[۷] را به‌صورت رابطه‌ی

$$S_\lambda(\phi_j) = \sum_{i=۱}^{n} \left\{ y_i - \sum_{j=۰}^{p+q} \phi_j b_{j,p}(x_i) \right\}^۲ + \lambda \int_{x_{min}}^{x_{max}} \left\{ \sum_{j=۰}^{p+q} \phi_j b''_j(x) \right\}^۲ dx \qquad (۸.۳)$$

معرفی کرد. ایلرز و مارکس (۱۹۹۶) جریمه مشتق دوم برای هموارسازی اسپلاین‌ها را با تفاضل مرتبه‌ی دوم[۸] ضرایب B-اسپلاین جایگذاری کردند و دلیل استفاده آن‌ها از تفاضل مرتبه‌ی دوم، مشتق دوم تابع B-اسپلاین است.

بردار ضرایب $\boldsymbol{\phi} = (\phi_۱, \phi_۲, \ldots, \phi_j)$ را در نظر بگیرید. تفاضل مرتبه‌ی اول به‌صورت زیر تعریف می‌شود:

$$\Delta\phi_j = \phi_j - \phi_{j-۱}, \quad j = ۲, \cdots, p+q.$$

تفاضل مرتبه‌ی دوم، در حالت کلی به شکل زیر است:

$$\begin{aligned} \Delta^۲\phi_j &= \Delta(\Delta\phi_j) \\ &= (\phi_j - \phi_{j-۱}) - (\phi_{j-۱} - \phi_{j-۲}) = \phi_j - ۲\phi_{j-۱} + \phi_{j-۲}, \quad j = ۳, ۴, \cdots, p+q. \end{aligned}$$

به ازای هر j می‌توان این تفاضل را به‌صورت زیر نوشت:

$$\begin{aligned} \Delta(\Delta\phi_۳) &= \phi_۱ - ۲\phi_۲ + \phi_۳ \\ \Delta(\Delta\phi_۴) &= \phi_۲ - ۲\phi_۳ + \phi_۴ \\ \Delta(\Delta\phi_۵) &= \phi_۳ - ۲\phi_۴ + \phi_۵ \\ &\vdots \end{aligned}$$

[۷] Penalized Sum Of Squares
[۸] 2nd Order Differences

با چنین صورت نگارشی که ضرایب مشابه در زیر هم واقع شده‌اند، می‌توان ماتریس $\boldsymbol{D}_{(K-۲)\times K}$ را که $K=p+q$، متناظر با ضرایب $\Delta^۲\phi_j$ به‌صورت ماتریس (۹.۳) در زیر تعریف کرد:

$$D=\begin{bmatrix} ۱ & -۲ & ۱ & ۰ & ۰ & ۰ & ۰ & ۰ & \cdots & ۰ \\ ۰ & ۱ & -۲ & ۱ & ۰ & ۰ & ۰ & ۰ & \cdots & ۰ \\ ۰ & ۰ & ۱ & -۲ & ۱ & ۰ & ۰ & ۰ & \cdots & ۰ \\ ۰ & ۰ & ۰ & ۱ & -۲ & ۱ & ۰ & ۰ & \cdots & ۰ \\ ۰ & ۰ & ۰ & ۰ & ۱ & -۲ & ۱ & ۰ & \cdots & ۰ \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ ۰ & ۰ & ۰ & ۰ & ۰ & ۰ & ۱ & -۲ & ۱ & ۰ \\ ۰ & ۰ & ۰ & ۰ & ۰ & ۰ & ۰ & ۱ & -۲ & ۱ \end{bmatrix}. \tag{۹.۳}$$

حال ماتریس $\boldsymbol{\mathcal{G}}_{K\times K}=\boldsymbol{D}^T\boldsymbol{D}$ عبارت است از

$$\boldsymbol{\mathcal{G}}=\begin{bmatrix} ۱ & -۲ & ۱ & ۰ & ۰ & ۰ & ۰ & ۰ & \cdots & ۰ \\ -۲ & ۵ & -۴ & ۱ & ۰ & ۰ & ۰ & ۰ & \cdots & ۰ \\ ۱ & -۴ & ۶ & -۴ & ۱ & ۰ & ۰ & ۰ & \cdots & ۰ \\ ۰ & ۱ & -۴ & ۶ & -۴ & ۱ & ۰ & ۰ & \cdots & ۰ \\ ۰ & ۰ & ۱ & -۴ & ۶ & -۴ & ۱ & ۰ & \cdots & ۰ \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ ۰ & ۰ & ۰ & ۰ & ۱ & -۴ & ۶ & -۴ & ۱ & ۰ \\ ۰ & ۰ & ۰ & ۰ & ۰ & ۱ & -۴ & ۶ & -۴ & ۱ \\ ۰ & ۰ & ۰ & ۰ & ۰ & ۰ & ۱ & -۴ & ۵ & -۲ \\ ۰ & ۰ & ۰ & ۰ & ۰ & ۰ & ۰ & ۱ & -۲ & ۱ \end{bmatrix}. \tag{۱۰.۳}$$

به آسانی می‌توان تحقیق کرد که تساوی زیر برقرار است (الیرز و مارکس، ۱۹۹۶):

$$\boldsymbol{\phi}^T\boldsymbol{\mathcal{G}}\boldsymbol{\phi}=\sum_{j=۳}^{K}(\Delta^۲\phi_j)^۲. \tag{۱۱.۳}$$

۳.۳ برآورد تابع مخاطره پایه با P-اسپلاین

تعداد K گره را روی بازه $[۰,t_{Rcens}]$ در نظر بگیرید که در آن t_{Rcens} کران بالای بازه تحت مطالعه یا همان بزرگترین زمان سانسور از راست است. این گره‌ها از پیش تعیین شده و به فاصله برابر از هم قرار دارند. برای برآورد تابع بقای پایه در مدل (۳.۲) به کمک اسپلاین‌ها، لگاریتم تابع مخاطره پایه را به‌صورت ترکیبی خطی از B -اسپلاین‌های مکعبی می‌نویسیم. فرض کنید $b_۱(t),\cdots,b_K(t)$، توابع پایه‌ای B -اسپلاین‌های مکعبی باشند. در این صورت می‌توان تابع بقای پایه را به‌صورت زیر نوشت:

$$h_۰(t)=\exp(\sum_{k=۱}^{K}b_k(t)\phi_k) \tag{۱۲.۳}$$

که در آن $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_K)$ ضرایب مجهول B -اسپلاین هستند و باید برآورد شوند. برای اطمینان از انعطاف لازم همراه با عدم بیش‌برازش[۹]، ایلرز و مارکس (۱۹۹۶) انتخاب تعداد بزرگی از B -اسپلاین‌ها را با افزودن یک جمله جریمه به لگاریتم تابع درستنمایی مدل به شکل

$$\tau \sum_k (\Delta^r \phi_k)^2 = \tau \boldsymbol{\phi}^{\boldsymbol{T}} \mathbf{D}^{\mathbf{T}} \mathbf{D} \boldsymbol{\phi}$$

پیشنهاد کردند، که در آن τ پارامتر جریمه و $\mathbf{D}$ ماتریس تفاضلی جریمه مرتبه r است. برای مثال، اگر $r = 2$ ماتریس تفاضلی مرتبه دو $\mathbf{D}$ به‌صورت

$$\mathbf{D} = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -2 & 1 \end{bmatrix}$$

و اگر $r = 3$ به‌صورت

$$\mathbf{D} = \begin{bmatrix} 1 & -3 & 3 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -3 & 3 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -3 & 3 & -1 \end{bmatrix}$$

تعریف می‌شود.

P- اسپلاین‌ها در موقعیت‌های مختلفی مورد استفاده قرار می‌گیرند. محققین زیادی از جمله ایلرز و مارکس (۱۹۹۶)، لانگ و برزگر (۲۰۰۴) و لامبرت (۲۰۰۷, ۲۰۱۳) از آن‌ها در چارچوب بیزی استفاده کرده‌اند. نکته قابل توجه در همه این مراجع آن است که اگر K به‌اندازه کافی بزرگ (بین ۱۰ تا ۲۰) انتخاب شود، نیازی به معرفی معیارهای انتخاب مدل نیست؛ زیرا تمام K ها تقریبا نتایج مشابهی را نشان می‌دهند.

۱.۳.۳ برآورد تابع بقای پایه

اگر f تابعی پیوسته (یا حتی ناپیوسته) و نامنفی بر بازه $[a, b]$ باشد، برای محاسبه مساحت محصور به نمودار f و محور x ها و خطوط $x = a$ و $x = b$، دنبال محاسبه مساحت مستطیل‌هایی هستیم که این محدوده (یعنی از a تا b) را پوشش دهند. اگر ارتفاع مستطیل‌ها بالاتر از نمودار تابع باشد، جمع همه مساحت را تقریب اضافی[۱۰] و اگر ارتفاع مستطیل‌ها پایین‌تر از نمودار تابع باشد به آن تقریب نقصانی[۱۱] می‌گویند. در نهایت سطح زیر نمودار تقریب برابر است با مجموع همه مستطیل‌ها.

[۹] Over Fitting
[۱۰] Additional Approximation
[۱۱] Defective Approximation

همان‌طور که در فصل ۲ بیان شد

$$S_{\circ}(t) = \exp\{-\int_{\circ}^{t} h_{\circ}(u)\mathrm{d}u\}.$$

لذا با دانستن این رابطه بین تابع مخاطره پایه و تابع بقای پایه و جایگذاری رابطه (۱۲.۳) در آن، داریم

$$S_{\circ}(t) = \exp(-\int_{\circ}^{t} \exp(\sum_{k=۱}^{K} b_k(u)\phi_k)\mathrm{d}u). \qquad (۱۳.۳)$$

واضح است که رابطه‌ی (۱۳.۳) به‌صورت عددی غیرقابل بررسی است. در نتیجه باید آن‌را به‌صورت تقریبی به‌دست آوریم. بازه $[\circ, t_{Rcens}]$ را در نظر بگیرید. این بازه را به یک دنباله متناهی از زیربازه‌ها با کران‌های $J_j = [\tau_{j-۱}, \tau_j]$ تقسیم می‌کنیم که هر زیربازه به‌صورت $\circ = \tau_{\circ} < \tau_۱ < \cdots < \tau_j = t_{Rcens}$ قابل نمایش است. انتخاب $J = ۳۰۰$ می‌تواند یک انتخاب معقول باشد. فرض کنید u_j و δ_j به ترتیب میانه و عرض J_j باشند. با استفاده از محاسبه انتگرال به روش ریمان رابطه (۱۳.۳) را می‌توان به‌صورت زیر تقریب زد:

$$S_{\circ}(t) \approx \exp(-\sum_{j=۱}^{j(t)} \exp(\sum_{k=۱}^{K} b_k(u_j)\phi_k)\delta_j) \qquad (۱۴.۳)$$

به‌طوری که $j(t)$ نشان‌دهنده فاصله شامل t است. برای شناسایی‌پذیر شدن مدل، آخرین پارامتر اسپلاین، یعنی ϕ_K، را یک مقدار به‌اندازه کافی بزرگ مثل $\phi_K = ۱۰$ در نظر می‌گیرند. این انتخاب باعث می‌شود تابع بقای پایه برآوردشده $\hat{S}_{\circ}(\cdot)$ در پایان زمان مطالعه صفر شود.

۴.۳ تابع درستنمایی مدل

فرض کنید $\mathbb{D}_i = (t_i, \nu_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ مجموعه مشاهدات آزمودنی iام باشد به‌طوری که t_i نشان‌دهنده زمان شکست یا زمان سانسور مشاهده‌شده، ν_i تابع مشخصه پیشامد مورد مطالعه و $\boldsymbol{x}_i$ و $\boldsymbol{z}_i$ دو بردار متغیرهای تبیینی در مدل نرخ شفایافتگی زمان پیشرفت هستند. اگر $\mathbf{D} = \{\mathbb{D}_i; i = ۱, \cdots, n\}$ مجموعه همه مشاهدات و $\boldsymbol{\Phi}$ بردار تمام پارامترهای مدل باشد، تابع درستنمایی مدل به‌صورت زیر خواهد بود:

$$L(\boldsymbol{\Phi}|\mathbf{D}) = \prod_{i=۱}^{n} h_p(t_i)^{\nu_i} S_p(t_i)$$

که در آن $S_p(t_i)$ در (۳.۲) نوشته شده است.

۵.۳ تحلیل بیزی

استنباط داده‌های بقا نسبت به سایر مدل‌های آماری پیچیده است. بنابراین همان‌طور که ذکر شد، رهیافت بیزی یک انتخاب معمول برای این داده‌ها است. در روش بیزی اگر اطلاعات اضافی در مورد پارامترها

در دست باشد، می‌توان از توزیع‌های پیشین برای وارد کردن آن‌ها در استنباط استفاده کرد. تابع چگالی پسین پارامترهای مدل از حاصل‌ضرب تابع درستنمایی و توابع چگالی پیشین نتیجه می‌شود. برای معرفی مدل بیزی نرخ شفایافتگی زمان پیشرفت پیشنهادی، ابتدا باید توزیع‌های پیشین پارامترها معرفی شوند.

مدل بیزی

بر اساس لانگ و برزگر (۲۰۰۴)، تابع چگالی توزیع پیشین بردار ضرایب B-اسپلاین، یعنی $\mathbf{\Phi}$، را به‌صورت زیر در نظر می‌گیریم:

$$\pi(\boldsymbol{\phi}|\tau) \propto \tau^{\frac{K}{2}} \exp(-\frac{\tau}{2}\boldsymbol{\phi}^T \mathbf{P}\, \boldsymbol{\phi})$$

که در آن $\mathbf{P} = \mathbf{D}^T\mathbf{D} + \epsilon \mathbf{I}_k$ یک ماتریس رتبه کامل به‌ازای یک مقدار کوچک برای ϵ مانند 10^{-6} است. به‌عبارت دیگر، یک توزیع نرمال چندمتغیره با بردار میانگین $\mathbf{0}$ و ماتریس واریانس $\mathbf{P}^{-1}$ برای پارامترهای اسپلاین در نظر گرفته شده است. بر اساس پیشنهاد جولیون و لامبرت (۲۰۰۷) برای پارامتر جریمه از یک توزیع پیشین سلسله‌مراتبی به‌صورت زیر استفاده شده است:

$$\tau|\phi \sim G(\frac{\nu}{2}, \frac{\nu\delta}{2})$$

$$\delta \sim G(a_\delta, b_\delta)$$

که در آن $G(a,b)$ توزیع گاما با میانگین $\frac{a}{b}$ و واریانس $\frac{a}{b^2}$ را نشان می‌دهد. آن‌ها نشان دادند اگر یک مقدار کوچک برای a_δ و b_δ انتخاب شود (مثل 10^{-4})، انتخاب ν (که اینجا برابر ۲ قرار داده شده است) بر شکل منحنی برآوردشده مخاطره پایه تاثیر نمی‌گذارد. اگر اطلاعات قبلی مثل یکنوا بودن در مورد مخاطره پایه در دسترس باشد، می‌توان آن‌را از طریق توزیع پیشین برای پارامترهای اسپلاین وارد کرد. برای تمام پارامترهای رگرسیونی $\boldsymbol{\beta}$، β_0 و γ از توزیع پیشین نرمال استفاده شده است.

اکنون با استفاده از قضیه بیز، توزیع پسین مدل به صورت

$$\pi(\mathbf{\Phi}|\mathbb{D}) \propto L(\mathbf{\Phi}|\mathbf{D})\pi(\phi|\tau)\pi(\delta)\pi(\beta_0, \boldsymbol{\beta})\pi(\gamma) \qquad (۱۵.۳)$$

نوشته می‌شود. توزیع پسین (۱۵.۳) شکل بسته ندارد و پیچیده است. بنابراین برای تقریب آن از الگوریتم‌های $MCMC$ استفاده می‌کنیم.

با توجه به آن که توزیع‌های پیشین انتخابی ذکر شده همگی سره هستند، توزیع پسین نیز حتما سره است. اما تنها توابع چگالی شرطی کامل[۱۲] پارامترهای τ و δ شکل بسته دارند و برای سایر پارامترها شکل چگالی شناخته‌شده نیست. در واقع می‌توان نشان داد

$$\tau|\boldsymbol{\phi}, \delta, \mathbb{D} \sim G(\frac{\nu + K}{2}, \frac{\nu\delta + \boldsymbol{\phi}^T\mathbf{P}\boldsymbol{\phi}}{2})$$

[۱۲] Full Conditional Densities

و

$$\delta|\tau,\mathbb{D}\sim G(a_\delta+\frac{\nu}{۲},b_\delta+\frac{\nu\tau}{۲}).$$

بنابراین از یک الگوریتم متروپولیس-هستینگز درون گیبز برای تولید نمونه از پسین (۱۵.۳) باید استفاده کرد.

۶.۳ تولید نمونه از توزیع پسین با الگوریتم‌های $MCMC$

زمانی که توزیع پسین به‌دست‌آمده دارای ابعاد زیاد و هم‌چنین پیچیدگی زیادی باشد امکان محاسبه توزی‌های پسین پارامترهای مدل به روش مستقیم وجود ندارد. بنابراین می‌توان با استفاده از روش‌های نمونه‌گیری $MCMC$ برای هر حجم نمونه دلخواه، استنباط قابل قبولی از پارامترها ارائه داد. در این روش با نمونه‌گیری‌های متوالی از توزیع شرطی کامل هر یک از پارامترها، با استفاده از یک الگوریتم مناسب مانند نمونه‌گیری گیبز و متروپولیس-هستینگز. زنجیرهای مارکوفی تشکیل می‌شود که در صورت همگرایی آن‌ها، می‌توان مقادیر شبیه‌سازی‌شده را نمونه‌هایی از توزیع پسین پارامترها دانست و بر اساس آن استنباط در خصوص پارامترهای مدل را انجام داد. الگوریتم تولید نمونه از توزیع پسین (۱۵.۳) به‌صورت زیر است:

فرض کنید $\tilde{\beta}_۰=(\beta_۰,\boldsymbol{\beta})$ و $\boldsymbol{\phi}^{(۰)}$، $\tau^{(۰)}$، $\delta^{(۰)}$، $\tilde{\boldsymbol{\beta}}^{(۰)}$ و $\boldsymbol{\gamma}^{(۰)}$ مقادیر اولیه برای شروع زنجیر باشد. در تکرار mام مراحل الگوریتم عبارتند از:

۱) تولید $\boldsymbol{\phi}^{(m)}$ از $\pi(\boldsymbol{\phi}|\tau^{(m-۱)},\delta^{(m-۱)},\tilde{\boldsymbol{\beta}}^{(m-۱)},\boldsymbol{\gamma}^{(m-۱)})$ با استفاده از یک الگوریتم متروپولیس-هستینگز.

۲) تولید $\tau^{(m)}$ از $G(\frac{\nu+k}{۲},\frac{\nu\delta^{(m-۱)}+\boldsymbol{\phi}^{(m)^T}\mathbf{P}\boldsymbol{\phi}^{(m)}}{۲})$ با استفاده از یک مرحله نمونه‌گیری گیبز.

۳) تولید $\delta^{(m)}$ از $G(a_\delta+\frac{\nu}{۲},b_\delta+\frac{\nu\tau^{(m)}}{۲})$ با استفاده از یک مرحله نمونه‌گیری گیبز.

۴) تولید $\tilde{\boldsymbol{\beta}}^{(m)}$ از چگالی شرطی کامل $\pi(\tilde{\boldsymbol{\beta}}|\boldsymbol{\phi}^{(m)},\tau^{(m)},\delta^{(m)},\boldsymbol{\gamma}^{(m-۱)})$ با استفاده از یک الگوریتم متروپولیس-هستینگز چندمتغیره.

۵) تولید $\boldsymbol{\gamma}^{(m)}$ از چگالی شرطی کامل $\pi(\tilde{\boldsymbol{\gamma}}|\boldsymbol{\phi}^{(m)},\tau^{(m)},\delta^{(m)},\boldsymbol{\beta}^{(m)})$ با استفاده از یک الگوریتم متروپولیس-هستینگز چندمتغیره.

فصل ۴

ارزیابی عملکرد مدل پیشنهادی

در این فصل با استفاده از مثال‌های واقعی و شبیه‌سازی، عملکرد مدل پیشنهادی معرفی‌شده در فصل سوم را ارزیابی می‌کنیم. برای این منظور، در مطالعه شبیه‌سازی دو سناریوی زمان پیگیری طولانی و کوتاه را در نظر می‌گیریم تا ویژگی شناسایی‌پذیر بودن مدل را نیز بررسی کنیم. مثال واقعی نیز یک مجموعه داده از بیماران مبتلا به ملانومای پوستی را شامل می‌شود.

نتایج گزارش شده در این فصل برگرفته از مقاله برمهورست و لامبرت (۲۰۱۶) هستند.

۱.۴ مطالعه شبیه‌سازی

۱.۱.۴ پیگیری طولانی مدت

همان‌طورکه می‌دانیم اگر روند مطالعه به اندازه‌کافی طولانی باشد، مدل پیشنهادی در (۱۲.۳) شناسایی‌پذیر است و در نتیجه بزرگترین زمان سانسور بیشتر از بزرگترین زمان شکست است. در این شبیه‌سازی در هر سناریوی در نظر گرفته شده، تابع بقای پایه در مدل (۱۲.۳) دارای توزیع وایبل با میانگین ۸ و انحراف معیار ۴٫۱۸ است. هم‌چنین متغیرهای تبیینی رگرسیونی، دو متغیر به‌صورت $W_1 \sim N(0, 1)$ و $W_2 \sim Bernoulli(0.5)$ در نظر گرفته شدند. از آن‌جایی‌که در این شبیه‌سازی فرض پیگیری کافی برقرار است، بنابراین هر دو متغیر می‌توانند به‌طور هم‌زمان برای مدل نسبت شفایافتگی و زمان لازم برای تشکیل یک تومور قابل تشخیص، بدون ایجاد یک مشکل شناسایی‌پذیری، مورد استفاده قرارگیرند. بنابراین

$$\mathbf{X} = \{W_1, W_2\} = \mathbf{Z}$$

در نظر گرفته شد. ضرایب رگرسیونی مربوط به W_1 و W_2 در مدل مخاطرات متناسب کاکس، به‌ترتیب ۰٫۴ و -0.4 در نظر گرفته شدند. کران بالا برای زمان شکست مشاهده‌شده نیز عدد ۲۳ قرار داده شد. با این انتخاب بیش از ۹۹٪ پیشامدهای تولید شده تحت توزیع وایبل انتخاب‌شده قبل از $t = 23$ اتفاق

می‌افتند. دو درصد ۲۵ و ۴۰ نیز برای نسبت افراد شفایافته در نظر گرفته شدند. این دو نسبت با تنظیم مقادیر واقعی ضرایب رگرسیونی $(\beta_0, \beta_1, \beta_2)$ در پارامتر نسبت شفایافته به‌دست می‌آیند.

با این توضیحات، هر مجموعه داده با استفاده از مدل زمان پیشرفت پیشنهادی به شرح زیر تولید شد:

(۱) تولید تعداد سلول‌های سرطان‌زا با استفاده از $N \sim Pois(\theta(\mathbf{x}))$ که در آن $\theta(\mathbf{x}) = \exp(\beta_0 + \mathbf{x}^T\beta)$.

(۲) اگر $N \neq 0$، متغیرهای پنهان $Y_1, \cdots, Y_N$ با استفاده از مدل مخاطرات متناسب کاکس تولید می‌شوند. سپس زمان شکست مشاهده‌شده به‌صورت $T = \min\{Y_1, \cdots, Y_N\}$ تعریف می‌شود. این مرحله تا زمانی که $T < 23$ تکرار می‌شود. توجه داشته باشید که این فرآیند در کمتر از ۱٪ موارد باید بیش از یک بار تکرار شود. اگر $N = 0$، آن‌گاه زمان شکست برای فرد شفایافته مقدار دلخواه بزرگی مثل ۹۹۹ قرار داده می‌شود.

(۳) نرخ کلی سانسور از راست تحت یکی از دو توزیع سانسور زیر کنترل می‌شود:

(a) سناریو (۱): یک توزیع یکنواخت در فاصله $[20, 25]$. این توزیع سانسور تضمین می‌کند که تقریبا همه افراد از راست سانسورشده، شفایافته و شناسایی‌پذیر هستند.

(b) سناریو (۲): یک توزیع وایبل با میانگین ۲۲٫۲۸ و انحراف معیار ۸٫۰۸ که در ۲۵ بریده[1] شده‌است. با استفاده از این توزیع سانسور، زمان‌های سانسور تنها ۲۵٪ افراد شفایافته در دم راست برآوردگر کاپلان‌مایر منحنی بقا قرار می‌گیرند و ۴٪ افراد شفانیافته، از راست سانسور می‌شوند.

حال با استفاده از رابطه (۱۲.۳) و در نظر گرفتن ۱۲ گره با فاصله‌های برابر در $[0, t_{Rcens}]$ که t_{Rcens} در آن برابر با ۲۵ است و ماتریس جریمه تفاضلی مرتبه سوم اقدامات زیر را انجام می‌دهیم:

- تولید تعداد ۵۰۰ مجموعه داده با حجم‌های نمونه $n = 300$ و $n = 600$.
- تولید یک زنجیر $MCMC$ با حجم مونت‌کارلوی ۲۳۰۰۰ از توزیع پسین مدل. دوره داغیدن[2] نیز برابر ۳۰۰۰ انتخاب می‌شود.
- بررسی همگرایی زنجیر تولید شده توسط معیار تشخیص جی‌وک (جی‌وک، ۱۹۹۲). برای جزییات این معیار پیوست ب را مشاهده کنید.
- محاسبه برآورد پارامترها (میانه توزیع پسین) و سایر کمیت‌های مورد نظر.

نتایج حاصل برای پارامترهای رگرسیونی $\tilde{\boldsymbol{\beta}}$، β_0 و $\tilde{\gamma}$ برای دو حجم نمونه $n = 300$ و $n = 600$ به‌ترتیب در جدول‌های ۱.۴ و ۲.۴ گزارش شده‌اند. در جدول‌ها، ستون اول نسبت افراد شفایافته، ستون دوم

[1] Truncated
[2] Burn-In

سناریوهای کنترل نرخ کلی سانسور، ستون سوم مقادیر واقعی پارامترها، ستون چهارم اریبی برآوردها، دو ستون پنجم وششم احتمال پوشش متناظر فواصل اعتبار ۹۰ و ۹۵ درصد، ستون هفتم خطای استاندارد تجربی[۳] (ESE) و ستون آخر مجذور میانگین توان دوم خطا[۴] $(RMSE)$ را نشان می‌دهند.

هنگامی‌که درصد افراد شفایافته ۲۵٪ و ۴۰٪ باشد، جدول‌های ۱.۴ و ۲.۴ نتایج شبیه‌سازی برای پارامترهای رگرسیون را نشان می‌دهند.

جدول ۱.۴: نتایج شبیه‌سازی با پیگیری طولانی مدت برای $n = ۳۰۰$ و ۵۰۰ مجموعه داده

شفایافته	سناریو	پارامترها	اریبی	$CV_{\%۹۰}$	$CV_{\%۹۵}$	ESE	$RMSE$
		$\beta_۰ = ۰/۷۵$	۰/۰۲۸	۸۸/۴	۹۳/۸	۰/۱۳۰	۰/۰۱۸
		$\beta_۱ = ۰/۸۰$	۰/۰۱۶	۹۰/۸	۹۴/۸	۰/۱۱۲	۰/۰۱۳
	۱	$\beta_۲ = -۰/۵۰$	−۰/۰۱۳	۹۱/۶	۹۶/۶	۰/۱۷۲	۰/۰۳۰
		$\gamma_۱ = ۰/۴۰$	−۰/۰۳۵	۹۱/۰	۹۵/۸	۰/۱۳۸	۰/۱۴۲
٪۲۵		$\gamma_۲ = -۰/۴۰$	۰/۰۰۸	۸۷/۴	۹۴/۲	۰/۲۲۱	۰/۲۲۱
		$\beta_۰ = ۰/۷۵$	۰/۰۵۲	۸۵/۸	۹۲/۶	۰/۱۶۲	۰/۰۳۰
		$\beta_۱ = ۰/۸۰$	−۰/۰۰۳	۹۳/۴	۹۷/۸	۰/۱۲۹	۰/۰۱۷
	۲	$\beta_۲ = -۰/۵۰$	۰/۰۱۷	۸۸/۶	۹۴/۰	۰/۲۰۹	۰/۰۴۴
		$\gamma_۱ = ۰/۴۰$	۰/۰۰۱	۹۲/۸	۹۵/۸	۰/۱۷۱	۰/۱۷۱
		$\gamma_۲ = -۰/۴۰$	−۰/۰۲۹	۸۸/۲	۹۳/۶	۰/۲۸۳	۰/۲۸۲
		$\beta_۰ = ۰/۳۰$	۰/۰۰۶	۹۰/۶	۹۴/۰	۰/۱۲۸	۰/۰۱۷
		$\beta_۱ = ۱/۰۰$	۰/۰۱۳	۹۰/۶	۹۵/۰	۰/۱۱۹	۰/۰۱۴
	۱	$\beta_۲ = -۰/۷۵$	−۰/۰۰۸	۹۰/۶	۹۴/۴	۰/۱۸۲	۰/۰۳۳
		$\gamma_۱ = ۰/۴۰$	−۰/۰۲۱	۹۰/۰	۹۵/۶	۰/۱۵۰	۰/۱۵۱
٪۴۰		$\gamma_۲ = -۰/۴۰$	۰/۰۰۳	۹۲/۰	۹۶/۸	۰/۲۱۵	۰/۲۱۵
		$\beta_۰ = ۰/۳۰$	۰/۰۴۳	۸۶/۲	۹۱/۸	۰/۱۵۵	۰/۰۲۷
		$\beta_۱ = ۱/۰۰$	−۰/۰۱۴	۹۳/۶	۹۷/۰	۰/۱۳۷	۰/۰۱۹
	۲	$\beta_۲ = -۰/۷۵$	۰/۰۲۴	۸۸/۰	۹۳/۴	۰/۲۲۲	۰/۰۴۹
		$\gamma_۱ = ۰/۴۰$	۰/۰۰۵	۹۱/۸	۹۶/۴	۰/۱۸۳	۰/۱۸۳
		$\gamma_۲ = -۰/۴۰$	−۰/۰۲۵	۹۰/۴	۹۴/۴	۰/۲۷۶	۰/۲۷۶

[۳] Emprical Standard Error
[۴] Root Mean Squared Error

جدول ۲.۴: نتایج شبیه‌سازی با پیگیری طولانی مدت برای $n = ۶۰۰$ و ۵۰۰ مجموعه داده

شفایافته	سناریو	پارامترها	اریبی	$CV_{\%۹۰}$	$CV_{\%۹۵}$	ESE	$RMSE$
		$\beta_۰ = ۰/۷۵$	۰/۰۲۰	۹۰/۲	۹۴/۰	۰/۰۸۷	۰/۰۰۸
		$\beta_۱ = ۰/۸۰$	۰/۰۱۷	۹۰/۴	۹۵/۲	۰/۰۷۷	۰/۰۰۶
	۱	$\beta_۲ = -۰/۵۰$	$-۰/۰۲۰$	۹۱/۶	۹۵/۸	۰/۱۱۵	۰/۰۱۴
		$\gamma_۱ = ۰/۴۰$	$-۰/۰۳۸$	۸۷/۶	۹۳/۶	۰/۰۹۷	۰/۱۰۶
٪۲۵		$\gamma_۲ = -۰/۴۰$	۰/۰۲۲	۸۹/۸	۹۶/۴	۰/۱۳۹	۰/۱۴۱
		$\beta_۰ = ۰/۷۵$	۰/۰۳۶	۸۷/۲	۹۲/۸	۰/۱۰۷	۰/۰۱۳
		$\beta_۱ = ۰/۸۰$	۰/۰۰۹	۹۳/۲	۹۷/۲	۰/۰۹۰	۰/۰۰۸
	۲	$\beta_۲ = -۰/۵۰$	$-۰/۰۱۱$	۸۸/۶	۹۳/۶	۰/۱۵۰	۰/۰۲۳
		$\gamma_۱ = ۰/۴۰$	$-۰/۰۲۳$	۹۱/۰	۹۶/۰	۰/۱۱۵	۰/۱۱۸
		$\gamma_۲ = -۰/۴۰$	۰/۰۱۸	۹۱/۶	۹۵/۰	۰/۱۸۲	۰/۱۸۳
		$\beta_۰ = ۰/۳۰$	۰/۰۰۵	۹۱/۲	۹۵/۲	۰/۰۸۷	۰/۰۰۸
		$\beta_۱ = ۱/۰۰$	۰/۰۱۶	۹۱/۲	۹۵/۶	۰/۰۷۹	۰/۰۰۶
	۱	$\beta_۲ = -۰/۷۵$	$-۰/۰۰۳$	۹۱/۶	۹۵/۶	۰/۱۱۹	۰/۰۱۴
		$\gamma_۱ = ۰/۴۰$	$-۰/۰۲۷$	۹۱/۲	۹۵/۰	۰/۰۹۸	۰/۱۰۱
٪۴۰		$\gamma_۲ = -۰/۴۰$	۰/۰۳۶	۹۲/۴	۹۵/۲	۰/۱۴۲	۰/۱۴۶
		$\beta_۰ = ۰/۳۰$	۰/۰۲۱	۸۹/۴	۹۴/۸	۰/۰۹۴	۰/۰۰۹
		$\beta_۱ = ۱/۰۰$	$-۰/۰۰۴$	۹۲/۸	۹۵/۴	۰/۰۹۳	۰/۰۰۹
	۲	$\beta_۲ = -۰/۷۵$	۰/۰۱۰	۹۱/۲	۹۵/۸	۰/۱۴۲	۰/۰۲۰
		$\gamma_۱ = ۰/۴۰$	$-۰/۰۰۲$	۹۳/۶	۹۷/۸	۰/۱۱۵	۰/۱۱۵
		$\gamma_۲ = -۰/۴۰$	۰/۰۲۵	۹۲/۶	۹۶/۴	۰/۱۸۴	۰/۱۸۴

با توجه به اعداد جدول‌ها می‌توان نتایج زیر را خلاصه کرد:

۱) برآوردهای ضرایب رگرسیونی برای هر دو حجم نمونه در نظر گرفته شده اریبی ناچیزی دارند.

۲) خطای تجربی و $RMSE$ برآوردهای پارامترهای رگرسیونی با افزایش حجم نمونه کوچک می‌شوند. یعنی دقت برآوردها افزایش می‌یابند.

برعکس این دو معیار در دو حالت افزایش می‌یابند:

الف) زمانی‌که نسبت افراد شفایافته با زمان سانسور بزرگتر از ماکسیمم زمان شکست مشاهده‌شده کاهش یابد.

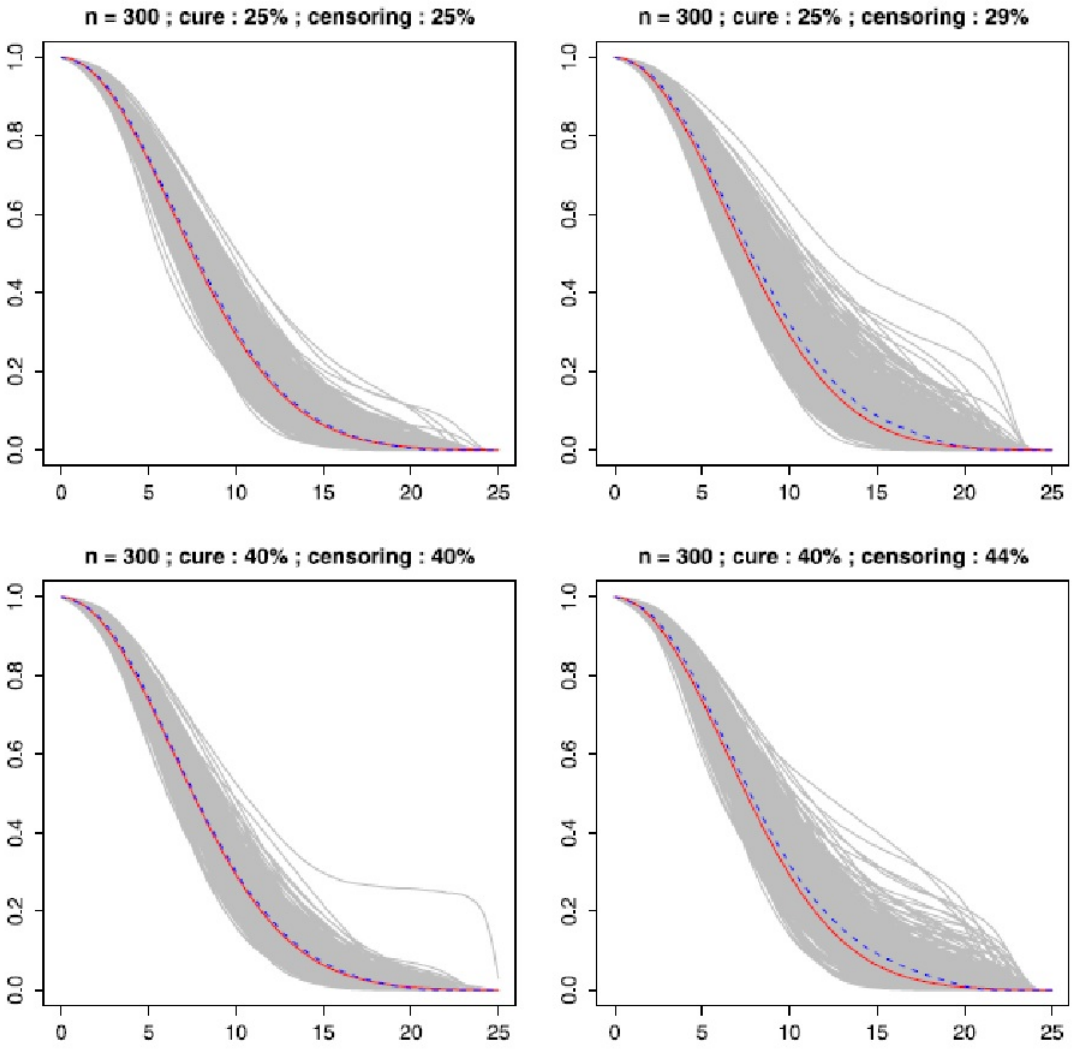
ب) درصد سانسور از راست در بین افراد شفانیافته افزایش یابد.

۳) احتمال‌های پوشش فواصل اعتبار ۹۰ و ۹۵ درصد به مقادیر اسمی خود نزدیک هستند.

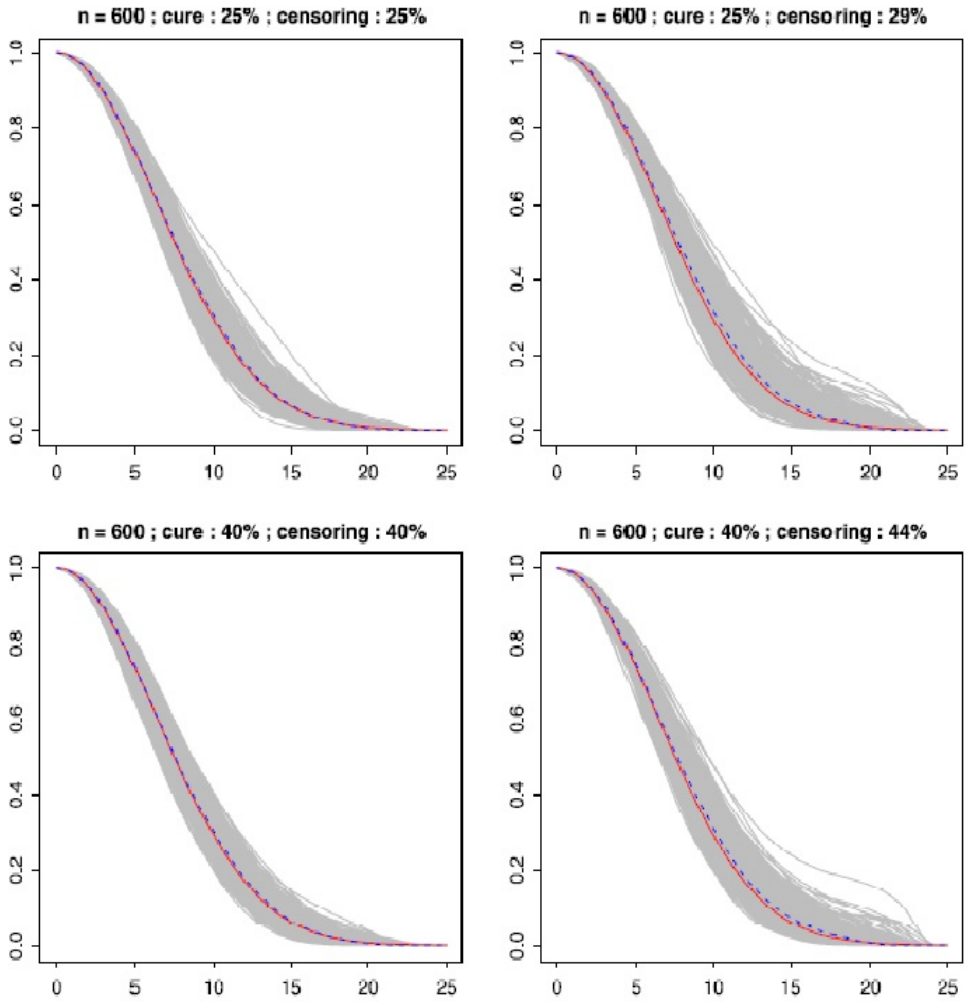
۴) نسبت افراد شفایافته بر دقت برآوردها تاثیر چندانی ندارند.

برآوردهای تابع بقای پایه برای حجم‌های نمونه $n = ۳۰۰$ و $n = ۶۰۰$ به‌ترتیب در شکل‌های ۱.۴ و ۲.۴ نمایش داده شده‌اند. از روی منحنی‌های برآوردشده می‌توان دریافت که زمانی که نسبت افراد شفایافته با زمان سانسور بزرگتر از ماکسیمم زمان شکست مشاهده‌شده، کاهش یابد و هم‌چنین زمانی که درصد سانسور از راست در بین افراد شفانیافته افزایش یابد، تغییرپذیری توزیع پایه برآوردشده تا حدودی افزایش خواهد یافت. از طرفی این تغییرپذیری با افزایش حجم نمونه کاهش می‌یابد.

از طرف دیگر می‌توان گفت تا حدودی اریبی برآورد دم سمت راست $S_۰(t)$ با افزایش حجم نمونه زمانی که نسبت افراد شفایافته با زمان سانسور بزرگتر از ماکسیمم زمان شکست مشاهده‌شده کاهش یابد و نسبت افراد از راست سانسور شده غیرمصون افزایش یابد، کاهش یافته است.

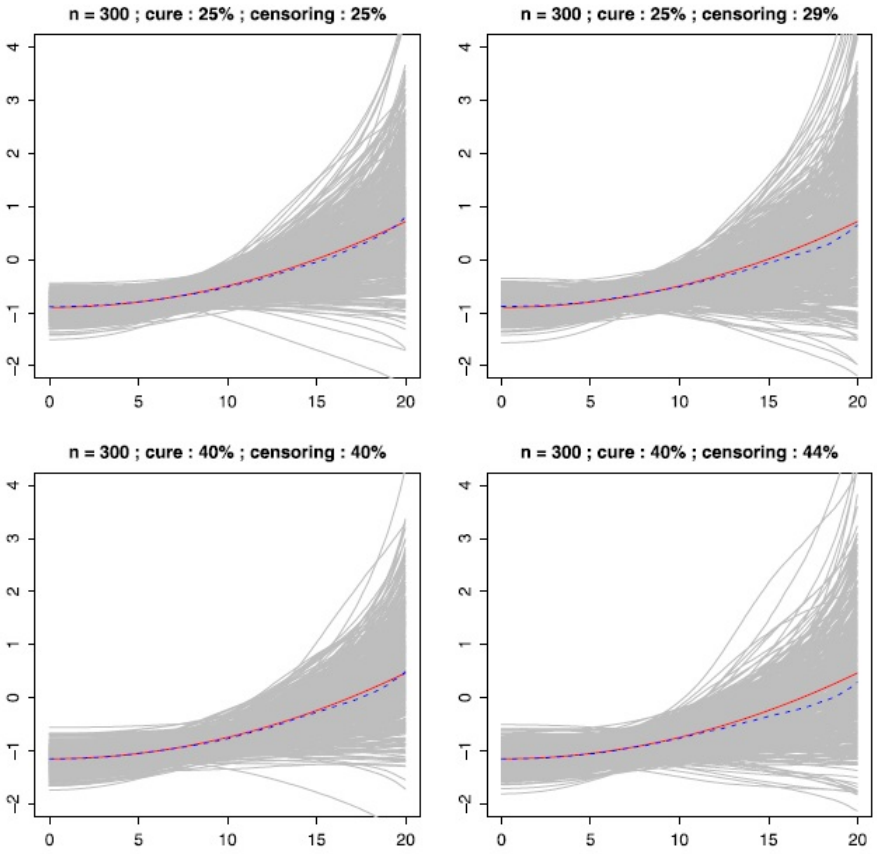


شکل ۱.۴: برآورد $S_0(t)$ در مثال شبیه‌سازی با پیگیری طولانی مدت برای ۵۰۰ مجموعه داده و $n = 300$. هر منحنی خاکستری برآورد تابع بقا در یک مجموعه داده است. هر ردیف شکل به درصدی از افراد شفایافته (ردیف بالا: ۲۵٪ و ردیف پایین: ۴۰٪) با نرخ کلی سانسور از راست (چپ: سناریو ۱ و راست: سناریو ۲) اختصاص دارد. منحنی خط ممتد نمودار واقعی $S_0(t)$ و منحنی خط‌چین میانه حاصل از ۵۰۰ تکرار است.

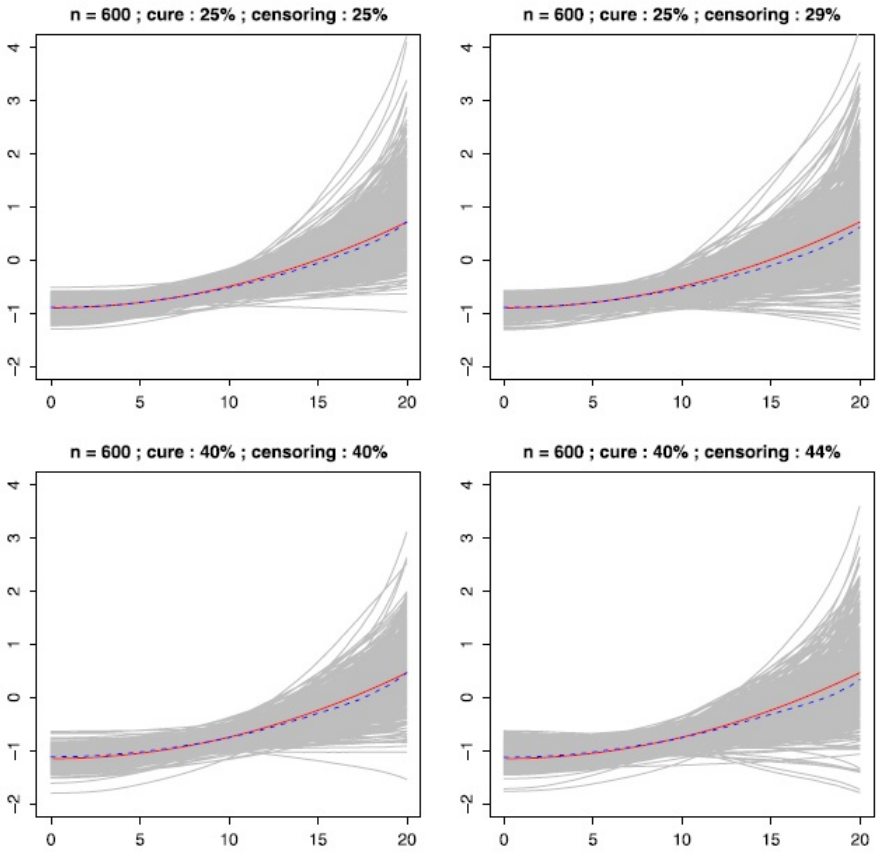


شکل ۲.۴: برآورد $S_0(t)$ در مثال شبیه‌سازی با پیگیری طولانی مدت برای ۵۰۰ مجموعه داده و $n = 600$. هر منحنی خاکستری برآورد تابع بقا در یک مجموعه داده است. هر ردیف شکل به درصدی از افراد شفایافته (ردیف بالا: ۲۵٪ و ردیف پایین: ۴۰٪) با نرخ کلی سانسور از راست (چپ: سناریو ۱ و راست: سناریو ۲) اختصاص دارد. منحنی خط ممتد نمودار واقعی $S_0(t)$ و منحنی خط‌چین میانه حاصل از ۵۰۰ تکرار است.

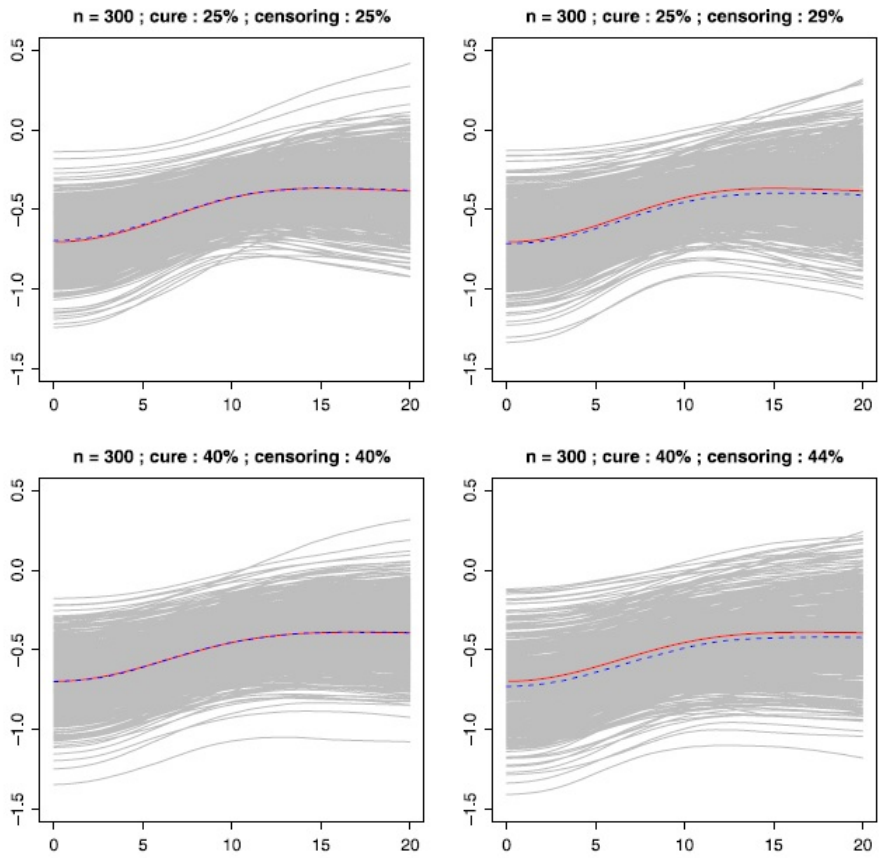
شکل‌های ۳.۴ و ۴.۴، برآورد نسبت لگاریتم مخاطره کل جامعه را برای به ترتیب $n = 300$ و $n = 600$ نمایش می‌دهند. نتایج ذکر شده برای برآوردهای $S_0(t)$ در نمودارهای ۱.۴ و ۲.۴، برای نسبت لگاریتم مخاطره نیز برقرار هستند و قابل درک است که نسبت افراد شفایافته تاثیری بر دقت این برآوردها نداشته‌اند. البته نسبت لگاریتم مخاطره برای افراد غیرمصون (۴.۴ برای $n = 300$) تحت هر سناریو به خوبی برآورد شده است.



شکل ۳.۴: برآورد نسبت لگاریتم مخاطره جامعه $\log(HR_p(t))$ در مثال شبیه‌سازی با پیگیری طولانی مدت برای ۵۰۰ مجموعه داده و $n = ۳۰۰$. هر یک منحنی خاکستری برآورد در یک مجموعه داده است. ردیف بالا به ۲۵٪ و ردیف پایین به ۴۰٪ از افراد شفایافته اشاره دارند. سمت چپ سناریو ۱ و سمت راست سناریو ۲ سانسور از راست را نشان می‌دهند.



شکل ۴.۴: برآورد نسبت لگاریتم مخاطره جامعه $\log(HR_p(t))$ در مثال شبیه‌سازی با پیگیری طولانی مدت برای ۵۰۰ مجموعه داده و $n = ۶۰۰$. هر یک منحنی خاکستری برآورد در یک مجموعه داده است. ردیف بالا به ۲۵٪ و ردیف پایین به ۴۰٪ از افراد شفایافته اشاره دارند. سمت چپ سناریو ۱ و سمت راست سناریو ۲ سانسور از راست را نشان می‌دهند.



شکل ۵.۴: برآورد نسبت لگاریتم مخاطره افراد غیرمصون در مثال شبیه‌سازی با پیگیری طولانی مدت برای ۵۰۰ مجموعه داده و $n = ۳۰۰$. هر یک منحنی خاکستری برآورد در یک مجموعه داده است. ردیف بالا به ۲۵٪ و ردیف پایین به ۴۰٪ از افراد شفایافته اشاره دارند. سمت چپ سناریو ۱ و سمت راست سناریو ۲ سانسور از راست را نشان می‌دهند.

۲.۱.۴ پیگیری کوتاه مدت

هدف از این مثال، استفاده محدود از مدل نرخ شفایافتگی پیشرفت زمان هنگامی‌که پیگیری مطالعه به‌اندازه کافی طولانی نیست، می‌باشد. مجموعه داده‌ها با استفاده از روش شرح داده‌شده در بخش ۱.۱.۴ تولید شدند. از آن‌جا که پیگیری مطالعه به‌اندازه کافی بلند نیست، برای شناسایی‌پذیر بودن مدل، متغیر تبیینی $W_۱$ را تنها برای بخش تولید یک تومور قابل تشخیص $(Z = W_۱)$ و متغیر تبیینی $W_۲$ را تنها برای بخش نسبت شفایافتگی $(X = W_۲)$ در نظر گرفتیم. دو مقدار ۲۵٪ و ۴۰٪ نیز برای نسبت افراد شفایافته در نظر گرفته شدند. برای اطمینان از اینکه بزرگترین زمان سانسور کوچکتر از بزرگترین زمان شکست مشاهده‌شده (ویژگی پیگیری کوتاه مدت) است، یک توزیع وایبل با میانگین ۱۷٫۹ و انحراف معیار ۶٫۵ که در ۱۳٫۷ (سناریوی ۳) و ۱۰٫۶ (سناریوی ۴) بریده شده اند، در نظر گرفتیم. مشابه ۱.۱.۴

و با در نظر گرفتن ۱۲ گره با فاصله‌های برابر روی $[۰, t_{Rcens}]$ که در آن t_{Rcens} برابر با ۱۳٫۷ یا ۱۰٫۶ است، شبیه‌سازی برای نمونه‌های به حجم $n = ۳۰۰$ و $n = ۶۰۰$ با ۵۰۰ تکرار انجام داده شد. مشابه مثال قبلی زنجیری به طول ۲۳۰۰۰ با دوره داغیدن ۳۰۰۰ تولید شد که آماره Z معیار جی‌وک تمام پارامترهای مدل برای بررسی همگرایی الگوریتم $MCMC$ بررسی و همگرایی زنجیر تایید شد.

نتایج حاصل برای پارامترهای رگرسیونی $\tilde{\boldsymbol{\beta}}$، $\boldsymbol{\beta}_۰$ و $\tilde{\gamma}$ برای حجم نمونه $n = ۳۰۰$ در جدول ۳.۴ گزارش شده‌اند.

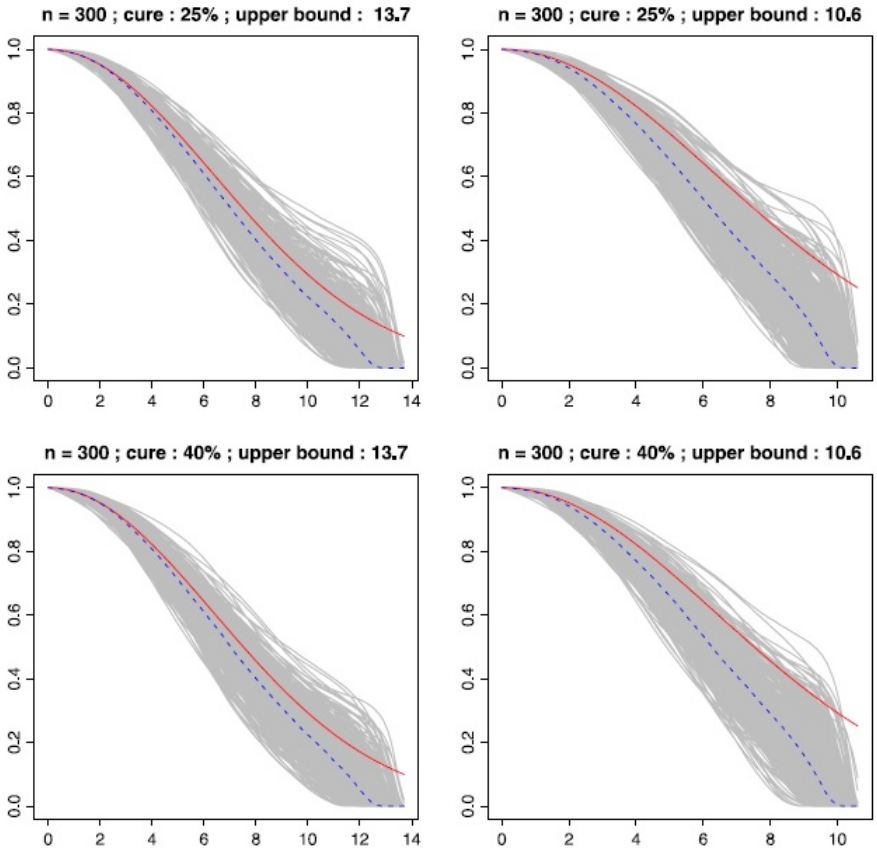
جدول ۳.۴: نتایج شبیه‌سازی با پیگیری کوتاه مدت برای $n = ۳۰۰$ و ۵۰۰ مجموعه داده

شفایافته	سناریو	پارامترها	اریبی	$CV_{\%۹۰}$	$CV_{\%۹۵}$	ESE	$RMSE$
		$\beta_۰ = ۰٫۷۰$	$-۰٫۰۹۳$	۸۰٫۶	۸۷٫۸	۰٫۱۲۹	۰٫۰۲۵
	۳	$\beta_۱ = -۰٫۷۰$	۰٫۰۰۷	۸۸٫۴	۹۵٫۶	۰٫۱۵۶	۰٫۰۲۴
٪۲۵		$\gamma_۱ = ۰٫۴۰$	$-۰٫۰۲۹$	۸۶٫۶	۹۲٫۲	۰٫۱۱۱	۰٫۱۱۴
		$\beta_۰ = ۰٫۷۰$	$-۰٫۲۶۶$	۴۶٫۶	۶۰٫۶	۰٫۱۵۷	۰٫۰۹۶
	۴	$\beta_۱ = -۰٫۷۰$	۰٫۰۰۷	۹۱٫۰	۹۳٫۸	۰٫۱۶۹	۰٫۰۲۹
		$\gamma_۱ = ۰٫۴۰$	$-۰٫۰۴۶$	۸۵٫۶	۹۱٫۰	۰٫۱۲۴	۰٫۱۳۰
		$\beta_۰ = ۰٫۳۰$	$-۰٫۱۱۲$	۷۷٫۴	۸۵٫۰	۰٫۱۴۲	۰٫۰۳۳
	۳	$\beta_۱ = -۰٫۸۰$	۰٫۰۲۴	۹۱٫۰	۹۶٫۰	۰٫۱۷۸	۰٫۰۳۲
٪۴۰		$\gamma_۱ = ۰٫۴۰$	$-۰٫۰۳۹$	۸۶٫۸	۹۳٫۰	۰٫۱۲۷	۰٫۱۳۳
		$\beta_۰ = ۰٫۳۰$	$-۰٫۲۹۲$	۴۵٫۴	۵۶٫۶	۰٫۱۵۹	۰٫۱۰۴
	۴	$\beta_۱ = -۰٫۸۰$	۰٫۰۲۴	۹۲٫۰	۹۶٫۲	۰٫۱۹۵	۰٫۰۳۸
		$\gamma_۱ = ۰٫۴۰$	$-۰٫۰۴۹$	۸۶٫۲	۹۲٫۲	۰٫۱۳۷	۰٫۱۴۵

با توجه به اعداد جدول، نتایج زیر قابل طرح هستند:

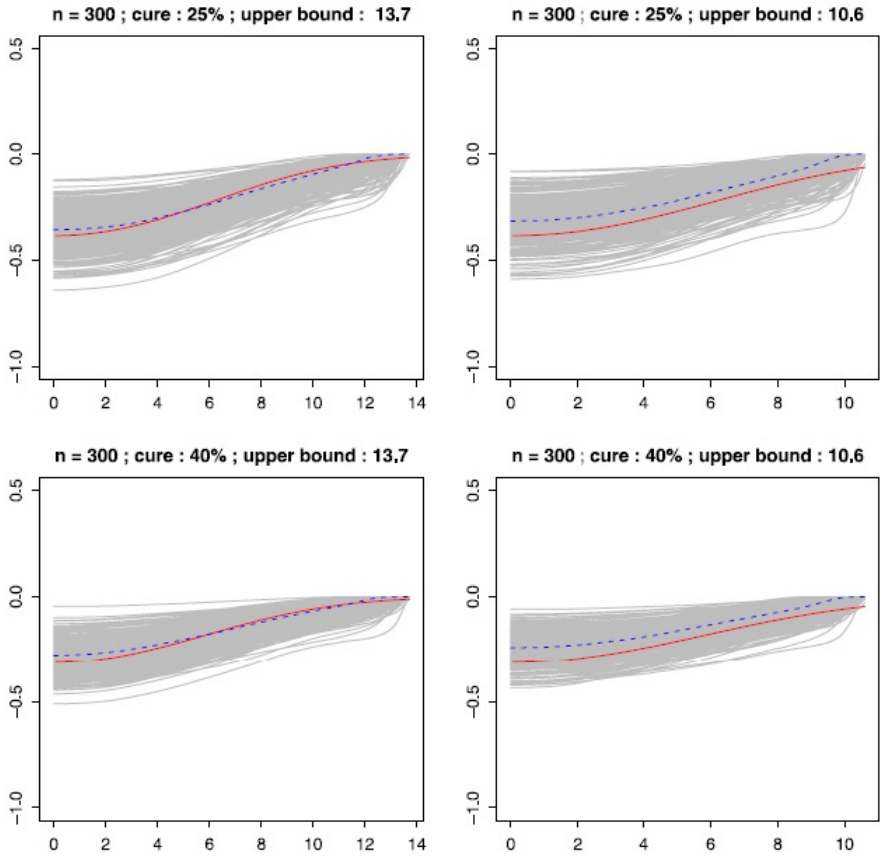
۱) احتمال‌های پوشش فواصل اعتبار ۹۰ و ۹۵ درصد به مقادیر اسمی خود نزدیک هستند.

۲) برآوردهای ضرایب رگرسیونی برای حجم نمونه در نظر گرفته‌شده اریبی ناچیزی دارند.



شکل ۶.۴: برآورد $S_0(t)$ در مثال شبیه‌سازی با پیگیری کوتاه مدت برای ۵۰۰ مجموعه داده و $n = 300$. هر منحنی خاکستری برآورد تابع بقا در یک مجموعه داده است. هر ردیف شکل به درصدی از افراد شفایافته (ردیف بالا: ۲۵٪ و ردیف پایین: ۴۰٪) با نرخ کلی سانسور از راست (چپ: سناریو ۳ و راست: سناریو ۴) اختصاص دارد. منحنی خط ممتد نمودار واقعی $S_0(t)$ و منحنی خط‌چین میانه حاصل از ۵۰۰ تکرار است.

شکل ۶.۴ نشان می‌دهد که توزیع پایه $S_0(t)$ به‌خاطر محدودیت در دم صفر برآورد نمی‌شود.



شکل ۷.۴: برآورد نسبت لگاریتم مخاطره افراد غیرمصون در مثال شبیه‌سازی با پیگیری کوتاه مدت برای ۵۰۰ مجموعه داده و $n = ۳۰۰$. هر منحنی خاکستری برآورد تابع بقا در یک مجموعه داده است. ردیف بالا به ۲۵٪ و ردیف پایین به ۴۰٪ از افراد شفایافته اشاره دارند. سمت چپ سناریو ۳ و سمت راست سناریو ۴ سانسور از راست را نشان می‌دهند.

نسبت لگاریتم مخاطره برای افراد غیرمصون (شکل ۷.۴ برای $n = ۳۰۰$) تحت هر سناریو به خوبی برآورد شده است. همان‌طور که در شکل ۷.۴ مشاهده می‌شود، نسبت مخاطره در افراد شفانیافته زمانی که کران بالای پیگیری کوچک است، بیش‌برازش می‌شود. این نتیجه زمانی حاصل می‌شود که توزیع بقای پایه در رابطه (۳.۲) در ماکسیمم زمان سانسور ممکن کوچک‌تر از یک باشد. در واقع در کوتاه‌ترین پیگیری (سناریوی ۴) $F_۰(۱۰/۶) = ۰/۷۵$ و در مطلوب‌ترین پیگیری (سناریوی ۳) $F_۰(۱۳/۷) = ۰/۹$.

نتایج مشابهی را می‌توان برای نمونه به حجم ۶۰۰ نیز بیان کرد. به‌طور کلی نتایج به‌دست آمده از این مثال شبیه‌سازی، لم ۲.۵.۲ را تایید می‌کنند.

۲.۴ مثال واقعی

این مطالعه بر روی ۲۸۴ بیمار مبتلا به ملانومای پوستی توسط گروه $ECOG$[۵] انجام شده است. در این مطالعه که شامل یک آزمایش بالینی تصادفی دومرحله‌ای است، تاثیر داروی اینترفرون آلفا-۲بی ارزیابی شده است (کرک‌وود و همکاران، ۱۹۹۶). مطالعه آن‌ها نشان داد که اینترفرون آلفا-۲بی تاثیر مثبتی بر زمان بقا بدون بازگشت بیماری دارد. از میان بیماران حاضر در مطالعه ۱۴۴ نفر (۵۱٪) درمان IFN دریافت کرده‌اند، که ۱۷۱ نفر (۶۰٪) از آن‌ها مرد هستند و عود سرطان برای ۱۹۶ نفر (۶۹٪) آن‌ها اتفاق افتاده است. سایر افراد در این مطالعه از راست سانسور شده‌اند. میانگین سن بیماران ۴۷ سال و انحراف معیار آن‌ها ۱۳ سال است. شکل ۸.۴ منحنی برآوردشده کاپلان مایر را برای تابع بقای بیماران نشان می‌دهد. با توجه به این شکل معلوم است که پیگیری مطالعه به‌اندازه کافی طولانی است. بنابراین فرض می‌کنیم جنس، سن و درمان تصادفی به‌طور هم‌زمان بر احتمال شفایافتگی و زمان لازم برای تولید یک تومور قابل تشخیص، تاثیر دارند.

داده‌های بقای این آزمایش بالینی قبلا توسط چن و همکاران (۱۹۹۹) و کونر و همکاران (۲۰۰۷) مورد مطالعه قرار گرفته اند. با این وجود، در تحلیل آن‌ها فرض بر این بود که این متغیرهای کمکی فقط بر احتمال شفایافتگی تاثیر دارند. اما می‌توان گفت که متغیرهای کمکی قادر هستند تا زمان لازم برای تولید یک تومور قابل تشخیص را نیز تحت تاثیر قرار دهند.

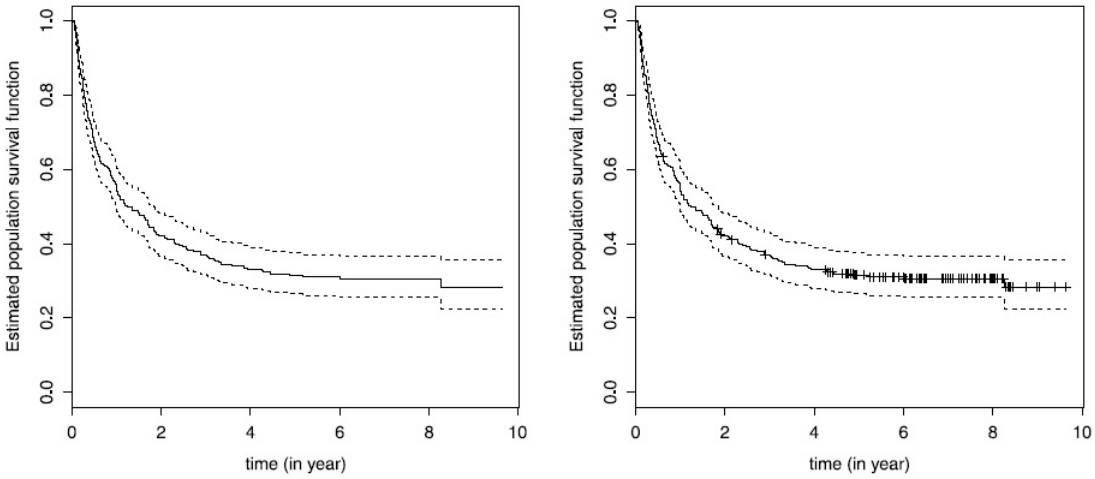
با در نظر گرفتن این داده‌ها و الگوریتم $MCMC$ پیشنهادی برای مدل بیزی ذکر شده در فصل ۳، زنجیری به‌طول ۲۳۰۰۰ از توزیع پسین تولید شد. برای بررسی همگرایی الگوریتم $MCMC$، ردیابی تمام پارامترهای مدل مورد بررسی قرار گرفت. آماره Z جی‌وک برای تمام پارامترهای مدل بین ۰٫۹۶ و ۱٫۹۶ یافت شد که نشانگر همگرایی الگوریتم $MCMC$ است. جدول ۴.۴ برآوردهای پارامترها را نشان می‌دهد.

می‌توان نتیجه گرفت که تنها درمان، تاثیر قابل توجهی بر احتمال شفایافتگی دارد. به‌عبارت دیگر، این مدل نشان می‌دهد که اینترفرون آلفا-۲بی به‌طور قابل توجهی سلول‌های سرطان‌زا را کاهش می‌دهد اما بر زمان انکوباسیون سلول تاثیر نمی‌گذارد. این نتیجه‌گیری در شکل‌های ۹.۴ و ۱۰.۴ نیز نشان داده شده است. منحنی سمت راست شکل ۹.۴ تابع بقا جمعیت و منحنی سمت چپ تابع بقا افراد شفانیافته را نشان می‌دهد. تنها اختلاف بین گروه‌های درمان در تابع بقا جمعیت، مربوط به متوسط سن افراد است. شکل ۱۰.۴ نیز برآورد نسبت لگاریتم مخاطره جمعیت (منحنی چپ) و نسبت لگاریتم مخاطره افراد شفانیافته (منحنی راست) را برای میانگین سن مردان نشان می‌دهد. نتایج مشابهی را می‌توان برای زنان نیز به‌دست آورد.

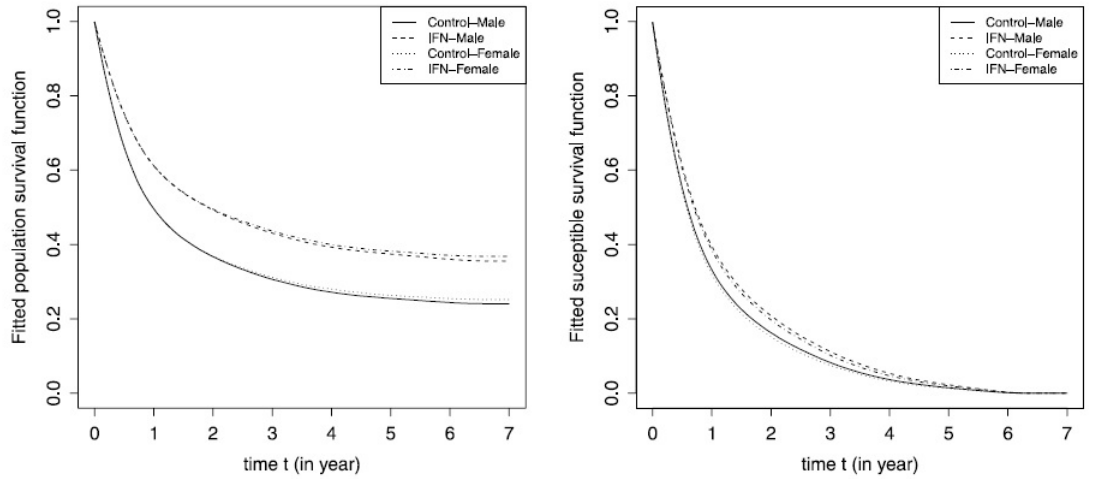
[۵] Eastern Cooperative Oncology Group

جدول ۴.۴: برآورد میانه‌های پسین، فاصله اعتبار ۹۵٪ و انحراف استاندارد پسین پارامترها در مثال واقعی.

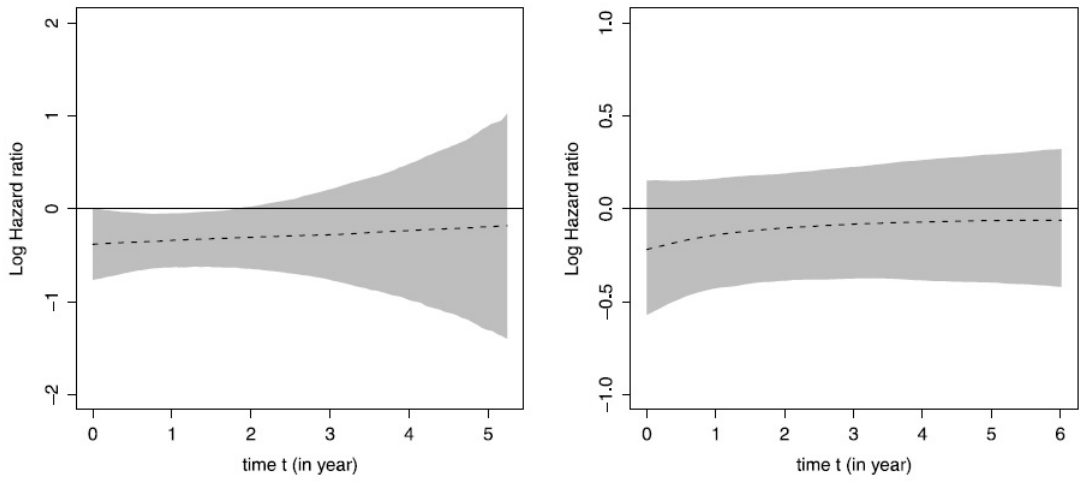
	پارامترها	برآورد	$HPD_{\%95}$	sd_{post}
$\theta(\mathbf{x})$	برش	۰٫۳۵۱	[۰٫۰۸۸, ۰٫۵۶۴]	۰٫۱۲۳
	سن	۰٫۱۰۰	[−۰٫۰۴۶, ۰٫۲۴۸]	۰٫۰۷۴
	IFN	−۰٫۳۲۱	[−۰٫۶۳۵, −۰٫۰۳۲]	۰٫۱۵۵
	جنس	−۰٫۰۳۱	[−۰٫۳۳۷, ۰٫۲۷۶]	۰٫۱۶۱
$F(t\|\mathbf{z})$	سن	−۰٫۱۳۶	[−۰٫۳۰۳, ۰٫۰۳۳]	۰٫۰۸۵
	IFN	−۰٫۰۶۰	[−۰٫۴۱۷, ۰٫۳۴۱]	۰٫۱۹۰
	جنس	۰٫۰۵۳	[−۰٫۳۳۳, ۰٫۴۳۰]	۰٫۱۹۵



شکل ۸.۴: برآورد کاپلان–مایر منحنی بقا در مثال واقعی. نمودار سمت راست: برای همه جامعه با علامت صلیب برای افراد سانسورشده و نمودار سمت چپ: برای افراد سانسورشده.



شکل ۹.۴: منحنی‌های برآوردشده بقا به تفکیک گروه‌های بیمار. چپ: همه جامعه و راست:بیماران شفانیافته.



شکل ۱۰.۴: نمودارهای برآوردشده نسبت لگاریتم مخاطره با ناحیه اعتبار ۹۵٪ برای، چپ: همه مردان و راست: مردان شفانیافته.

۳.۴ نتیجه‌گیری و پیشنهادات

یک نسخه انعطاف‌پذیر از مدل نرخ شفایافتگی پیشرفت زمان هنگامی‌که متغیرها به‌طور هم‌زمان بر احتمال شفایافتگی و زمان لازم برای سلول به منظور تولید یک تومور قابل تشخیص تاثیر می‌گذارند، پیشنهاد شد. اگر چه مدل پیشنهادی ساختار مخاطرات متناسب در بیمار را ندارد، اما یک برآورد ساده از نسبت‌های خطر در طول زمان را فراهم می‌کند. هنگامی‌که پیگیری مطالعه به‌اندازه کافی طولانی نیست، باز هم می‌توان از مدل زمان پیشرفت با برخی محدودیت‌ها استفاده کرد. در این زمینه، ثابت شده است که اگر از متغیرهای تبیینی مشترکی به‌طور هم‌زمان برای مدل احتمال شفایافتگی و زمان لازم برای تشخیص رشد تومور سلول سرطانی استفاده نشود، اثرات متغیرها شناسایی‌پذیر هستند.

استفاده از یک پیوند لجیت برای مدل‌سازی احتمال شفایافتگی مورد بررسی قرار گرفت تا سعی کند مشکل شناسایی‌پذیری را حل کند.

با توجه به محتوی پایان‌نامه می‌توان پیشنهادات زیر را برای بررسی و انجام ارائه داد:

۱) مدل‌بندی ناپارامتری تابع بقای پایه در دیدگاه بیزی با سایر روش‌ها مانند روش هسته یا چندجمله‌ای موضعی.

۲) در نظر گرفتن اثرات ناپارامتری برای متغیرهای رگرسیونی به‌جای اثر خطی آن‌ها در هر دو قسمت نسبت شفایافتگی و تابع بقای افراد غیرمصون.

۳) وارد کردن وابستگی داده‌ها به مدل پیشنهادی، مانند وابستگی طولی یا فضایی داده‌های بقا.

پیوست آ

برآوردگر ناپارامتری کاپلان-مایر

برای برآورد تابع مخاطره پایه می‌توان از روش کاپلان-مایر استفاده کرد. در این روش برآورد تابع مخاطره پایه در زمان $t_{(i)}$ از i امین پیشامد به‌صورت

$$\hat{h}_{\circ(t_{(i)})} = \frac{d_{(i)}}{\sum_{j \in R_{(t_{(i)})}} \exp(\hat{\beta}^T x_j)}$$

تعریف می‌شود، که در آن $d_{(i)}$ تعداد مرگ‌ها در زمان $t_{(i)}$ است و $R_{(t_{(i)})}$ مجموعه‌ای از افرادی است که می‌توانند در آن زمان بمیرند.

برآورد تابع مخاطره پایه تخمینی مجزا برای یک تابع پیوسته است. بنابراین با توجه به

$$S(t) = \exp\{-\int_{\circ}^{t} h(\tau)\mathrm{d}\tau\}$$

می‌توان از $\hat{h}_{\circ(t_{(i)})}$ برای برآورد $\int_{t_{(i-1)}}^{t_{(i)}} h_{(\circ)}(\tau)\mathrm{d}\tau$ استفاده کرد. در این صورت برآورد تابع بقای پایه به‌صورت زیر به‌دست می‌آید:

$$\hat{S}_{\circ(t_{(i)})} = \exp[-\sum_{j \leqslant i} \hat{h}_{\circ(t_{(j)})}].$$

پیوست ب

معیار تشخیص همگرایی جی‌وک برای الگوریتم‌های $MCMC$

جی‌وک در سال ۱۹۹۲ یک آماره آزمون تشخیص همگرایی که برابری میانگین‌های بخش اول و آخر زنجیرهای مارکوف را اندازه‌گیری می‌کند، ارائه داد. این دو بخش به‌طور معمول، ۱۰ درصد اول و ۵۰ درصد آخر زنجیر در نظر گرفته می‌شوند. آزمون تشخیص همگرایی برای این‌که چه تعداد از تکرارهای آغازی باید نادیده گرفته شوند، مورد استفاده قرار می‌گیرد. به عبارتی این آزمون به‌عنوان یک برآوردگر برای دوره داغیدن شناخته می‌شود.

اگر n تعداد تکرارها و $n_\circ$ برابر با اولین تکراری باشد که بخواهیم با انجام این آزمون بررسی کنیم که زنجیر همگرا می‌شود یا نه، باید آماره آزمون جی‌وک که یک آماره Z[1] استاندارد است، محاسبه شود. فرض کنید $\theta(X)$ کمیت مورد دلخواه باشد و $\theta^t = \theta(X^{(t+n_\circ)})$. اگر n_A تکرار اول، تعداد تکرارهای بخش A و n_B تکرار آخر، تعداد تکرارهای بخش B، که در آن n^* مقداری است که $n - n^* + 1 = n_B$، باشند آن‌گاه بخش‌های A و B و میانگین‌های آن‌ها به صورت زیر تعریف می‌شوند:

$$
\begin{aligned}
A &= \{t : 1 \leqslant t \leqslant n_A\},\\
B &= \{t : n^* \leqslant t \leqslant n\},\\
1 < n_A < n^* < n, \quad \frac{n_A + n_B}{n} &< 1,\\
\bar{\theta}_A &= \frac{1}{n_A}\Sigma_{t\in A}\theta^t,\\
\bar{\theta}_B &= \frac{1}{n - n^* + 1}\Sigma_{t\in B}\theta^t.
\end{aligned}
$$

آزمون جی‌وک بیان‌گر این مطلب است که اگر زنجیر در $n_\circ$ همگرا شود، آن‌گاه نمونه‌های تولیدشده از توزیع مانای زنجیر گرفته شده‌اند و در نتیجه میانگین‌های محاسبه‌شده از اولین و آخرین بخش زنجیر، معادل هستند. بنابراین آماره جی وک که به‌طور مجانبی دارای توزیع نرمال استاندارد است، عبارت

[1] Z-score

است از

$$Z_n = \frac{\bar{\theta_A} - \bar{\theta_B}}{\sqrt{\frac{\hat{S}_\theta^A}{n_A} + \frac{\hat{S}_\theta^B}{n-n^*+1}}} \to N(0,1) \quad n \to \infty,$$

که در آن $\frac{\hat{S_\theta}^A}{n_A}$ و $\frac{\hat{S_\theta}^B}{n-n^*+1}$ واریانس‌های مجانبی $\bar{\theta}_A$ و $\bar{\theta}_B$ هستند. حال با توجه به مطالب ذکرشده می‌توان فرضیه صفر، فرضیه برابری میانگین‌ها، را روی فضای حالت θ آزمون کرد. در این آزمون، اگر مقدار $|Z_n|$ بزرگ باشد، آن‌گاه فرضیه صفر رد می‌شود و این یعنی زنجیر هنوز همگرا نشده است.

پیوست پ

کد تولید داده‌ها از مدل نرخ شفایافتگی زمان پیشرفت

```
library(mvtnorm)
library(survival)
library(mixdist)
library(matlib )
library(FAmle)

weibull.par=weibullpar(8,4.18)

K=12
n=K
W11=rnorm(n,0,1);W12=rbinom(n,1,0.5)
W21=rnorm(n,0,1);W22=rbinom(n,1,0.5)
W1=rnorm(n,0,1);W2=rbinom(n,1,0.5)
x=matrix(c(W1,W2),ncol=2)
z=matrix(c(W1,W2),ncol=2)

x1=matrix(c(W11,W12),ncol=2)
z1=matrix(c(W11,W12),ncol=2)
x2=matrix(c(W21,W22),ncol=2)
z2=matrix(c(W21,W22),ncol=2)

beta0=0.75 ; beta1=0.8 ; beta2=-0.5
Beta=c(beta1,beta2)
Beta_real=c(beta0,beta1,beta2)
```

```
gamma1=0.4 ; gamma2=-0.4
Gamma=c(gamma1,gamma2)
theta=exp(beta0+x%*%Beta)

T=c()
N=rpois(n,theta)
for(i in 1:n){
if(N[i]==0){T[i]=999}
else
{y=rweibull(N[i],weibull.par$shape,weibull.par$scale)
T[i]=min(y)}
}
T
r_cens=runif(n,20,25)
r_cens
nou=as.numeric(r_cens>=23)

D=matrix(c(T,nou,x[,1],x[,2],z[,1],z[,2]),ncol=6)
I <- matrix(0, K, K);diag(I) <- 1
epsilon=10^-6
P=(D%*%t(D))+(epsilon*I)
solve(P)

t_Rcens=25
B=3
n=300
S=500
```

مراجع

[1] Bremhorst, V., Lambert, P. (2016). Flexible estimation in cure survival models using P-splines, *Computational Statistics and Data Analysis,* **93**, 270-284.

[2] Berkson, J., Gage, R.P. (1952). Survival curve for cancer patients following treatment, *Journal of American Statistical Association,* **47**, 501–515.

[3] Cetinyurek, A., Lambert, P. (2011). Smooth estimation of survival functions and hazard ratios from interval-censored data using Bayesian penalized B-splines, *Statistics in Medicine,* **30**, 75–90.

[4] Chen, M.H., Ibrahim, J.G., Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction, *Journal of American Statistical Association,* **94**, 909–919.

[5] Cooner, F., Banerjee, S., Carlin, B.P., Sinha, D. (2007). Flexible cure rate modelling under latent activation schemes, *Journal of American Statistical Association,* **102**, 560–572.

[6] Cox, D.R.(1972). *Regression Models and Life-Tables, Journal of the Royal Statistical Society: series B*, **34**, 187-220.

[7] Cuzick, J. (1992). Semiparametric additive regression, *Journal of the Royal Statistical Society: series B*, **54**, 831-843.

[8] Eilers, P.H.C. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood, *Statistical Modeling*, **7**, 239–254 .

[9] Eilers, P.H.C., Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with discussion), *Statistical Society.* **11**, 89–121.

[10] Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales, *Journals of American Statistical Association*, **81**, 310- 320.

[11] Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics,* **38**, 1041–1046.

[12] Farewell, V.T. (1986). Mixture models in survival analysis: are they worth the risk?, *Canadian Journal of Statistics,* **14**, 257–262.

[13] Friedman, J. H. (1990). Multivariate adaptive regression splines. *The Annals of Statistics.*, **19**, 11- 41.

[14] Gelman, A., Roberts, G.O., Gilks, W.R. (1996). *Efficient Metropolis jumping rules.* In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 5. Oxford University Press, UK.

[15] Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments, In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 4. Clarendon Press, Oxford, UK.

[16] Haario, H., Saksman, E., Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, **7**, 223–242.

[17] Hekman, N. E. (1986). Spline smoothing in partly linear models. *Journal of the Royal Statistical Society: series B*, **48**, 244-248 .

[18] Ibrahim, J.G., Chen, M.-H., Sinha, D. (2001). Bayesian semiparametric models for survival data with a cure fraction, *Biometrics,* **57**, 383–388.

[19] Jullion, A., Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models, *Computational Statistics and Data Analysis*, **51**, 2542–2558.

[20] Kirkwood, J.M., Strawderman, M.H., Ernstoff, M.S., Smith, T.J., Borden, E.C., Blum, R.H. (1996). Interferon alpha-2b adjuvant therapy of high-risk resected cutaneous melanoma: the eastern cooperative oncology group trial e1684, *Journal of Clinical Oncology,* **14**, 7–17.

[21] Kuk, A.Y.C, Chen, M.H. (1992). A mixture model combining logistic regression with proportional hazards regression, *Biometrika,* **79**, 531–541.

[22] Lambert, P. (2007). Archimedean copula estimation using Bayesian splines smoothing techniques, *Computational Statistics Data and Analysis,* **51**, 6307–6320.

[23] Lambert, P. (2013). Nonparametric additive location-scale models for interval censored data, *Statistical Computation,* **23**, 75–90.

[24] Lambert, P., Eilers, P. (2005). Bayesian proportional hazards model with time varying regression coefficients: a penalized Poisson regression approach, *Statistics in Medicine,* **24**, 3977–3989.

[25] Lang, S., Brezger, A. (2004). Bayesian P-splines, *Journal of Computiational and Graphical Statistics,* **13**, 183–212.

[26] Li, C.S., Taylor, J.M.G. (2002). A semi-parametric accelerated failure time cure model, *Statictics in Medicine,* **21**, 3235–3247.

[27] Liu, H., Shen, Y. (2009). A semiparametric parametric regression cure model for interval-censored data, *Journal of American Statististical Association,* **104**, 1168–1178.

[28] Lopes, C.C.M., Bolfarine, H. (2012). Random effects in the promotion times cure rate models, *Computational Statistics and Data Analysis,* **56**, 75–87.

[29] Lu, W. (2010). Efficient estimation for an accelerated failure time model with a cure fraction, *Statistica Sinica,* **20**, 661–674.

[30] Ma, Y., Yin, G. (2008). Cure rate model with mismeasured covariates under transformation, *Journal of American Statistical Association,* **103**, 743–756.

[31] Peng, Y. (2003). Estimating baseline distribution in proportional hazards cure models, *Computational Statistics and Data Analysis,* **42**, 187–201.

[32] Peng, Y., Dear, K.B.G. (2000). A nonparametric mixture model for cure rate estimation *Biometrics,* **56**, 237–243.

[33] Roberts, G.O., Rosenthal, J.S. (2001). Optimal scaling for various Metropolis Hastings algorithms, *Statistical Society,* **16**, 351–367.

[34] Schoenberg I.J. (1946). *Contribution to the problem of approximation of equidistant data by analytic functions, Parts A and B,* Quarterly Applied Mathematics 4 , 45–99, 112–141.

[35] Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, **50**, 413–436.

[36] Sy, J.P., Taylor, J.M.G. (2000). Estimation in a Cox proportional hazards cure model, *Biometrics,* **56**, 227–236.

[37] Taylor, J.M.G. (1995). Semi-parametric estimation in failure time mixture models, *Biometrics,* **51**, 899–907.

[38] Tsodikov, A.D. (1998). A proportional hazard model taking account of long-term survivors, *Biometrics,* **54**, 1508–1516.

[39] Tsodikov, A.D. (2002). Semi-parametric model of long- and short-term survival: an application to the analysis of breast cancer survival in Utah by age and stage, *Statics in Medicine,* **21**, 895–920.

[40] Wang, L., Du, P., Liang, H. (2012). Two-component mixture cure rate model with spline estimated nonparametric components, *Biometrics,* **68**, 726–735.

[41] Yakovlev, A.Y., Tsodikov, A.D. (1996). *Stochastic Models for Tumor of Latency and their Biostatistical Applications,* World Scientific Publication, Singapore.

[42] Yin, G., Ibrahim, J.G. (2005). Cure rate models: a unified approach, *Canadian Journal of Statistics,* **33**, 559–570.

[43] Zeng, D., Yin, G., Ibrahim, J.G. (2006). Semiparametric transformation models for survival data with a cure fraction, *Jounal of American Statistics Association,* **101**, 670–684.

[44] Zhang, J., Peng, Y., Li, H. (2013). A new semiparametric estimation method for accelerated hazards mixture cure model, *Computational Statistics and Data Analysis,* **59**, 95–102.

واژه‌نامه فارسی به انگلیسی

اسپلاین مکعبی ..... Cubic Spline
تابع بقا ..... Survival Function
تابع شکست ..... Failure Function
تابع مخاطره ..... Hazard Function
تحلیل بقا ..... Survival Analysis
شناسایی‌پذیری ..... Identification
مدل‌های نرخ شفایافتگی ..... Cure Rate Models
مدل‌های نرخ شفایافتگی غیرآمیخته ..... Non-Mixture Cure Models
مدل مخاطرات متناسب کاکس ..... Cox Proportional Hazard Model
متغیرهای کمکی ..... Covariates
نیمه‌پارامتری ..... Semi-Parametric

واژه‌نامه انگلیسی به فارسی

متغیرهای کمکی ..... Covariates
مدل مخاطرات متناسب کاکس ..... Cox Proportional Hazard Model
اسپلاین مکعبی ..... Cubic Spline
مدل‌های نرخ شفایافتگی ..... Cure Rate Models
تابع شکست ..... Failure Function
تابع مخاطره ..... Hazard Function
شناسایی‌پذیری ..... Identification
مدل‌های نرخ شفایافتگی غیرآمیخته ..... Non-Mixture Cure Models
نیمه پارامتری ..... Semi-Parametric
تحلیل بقا ..... Survival Analysis
تابع بقا ..... Survival Function

Abstract

In the analysis of survival data, it is usually assumed that any unit will experience the event of interest if it is observed for a sufficiently long time. However, it can be assumed that an unknown proportion of the population under study will never experience the interest event.The Promotion time model, which has a biological motivation, is one of the survival models taking this feature in to account. The survival function of uncured people is estimated by using a Cox proportional hazard model where the logarithm of the baseline hazard function is specified using Bayesian P-splines. We use a Bayesian framework for implementing Statistical inference in the proposed model. The identification issues of the model are discussed and a restricted use of the model when the follow-up of the study is not sufficiently long is suggested. The accuracy of our methodology is assumed through a simulation study and the model is illustrated on a dataset from a Melanoma clinical trial.

Keywords: Survival analysis, Promotion time model, P-spline, Cox proportional hazard model, Bayesian infrence.

Shahrood University of Technology

Faculty of Mathematical Sciences

Bayesian analysis of non-mixture cure rate models with a Cox semi-parametric survival function by using P-splines

Neda Shams

Supervisors

Dr. Hossein Baghishani and Dr. Negar Eghbal

Jul 2018