# رگرسیون بقا به شیوه بیز ناپارامتری برای داده‌های سانسور شده با استفاده از فرایند دیریخله وابسته

زهره شاه‌حسینی ، سید مرتضی امینی
بخش آمار دانشکده ریاضی آمار و علوم کامپیوتر، دانشگاه تهران

**چکیده**

در این مقاله تحلیل بیز ناپارامتری براساس پیشین فرآیند دیریخله وابسته را برای تحلیل بقای داده‌های سانسور شده با در نظر گرفتن عوامل تأثیرگذار در طول عمر معرفی کرده و مورد بررسی و استفاده قرار می‌دهیم. همچنین الگوریتم‌های محاسباتی و نمونه‌گیری زنجیره مارکف را معرفی و تشریح کرده و برای استفاده در این مسأله توسعه می‌دهیم. در نهایت به تحلیل بقای داده‌های واقعی مربوط به طول عمر بیماران سرطانی با استفاده از مدل‌های کاکس، کپلان مایر و نیز روش تحلیل واریانس دیریخله سلسله مراتبی وابسته می‌پردازیم. نشان داده می‌شود که روش‌های کلاسیک تحلیل بقا در بعضی زمان‌ها مقدار ثابتی را اختیار می‌کنند. با این حال حضور داده‌های سانسور شده در براورد تابع بقا تحت مدل تحلیل واریانس دیریخله سلسله مراتبی وابسته چنین مشکلی را ایجاد نمی‌کند.

**واژه‌های کلیدی:** تحلیل بقا، داده‌های سانسور شده، زنجیر مارکف مونت کارلو، فرآیند دیریخله آمیخته، فرآیند دیریخله وابسته، مدل سلسله مراتبی، نمونه‌بردار گیبس.

## ۱ مقدمه

پیشرفت فناوری و بالا رفتن سطح زندگی مردم با استفاده از دستاوردهای صنعتی، رقابت‌های شدیدی بین تولیدکنندگان برای ارائه محصولات و خدمات مرغوب‌تر پدید آورده است. وجود محصولات مشابه و منابع مختلف اطلاعاتی، به مشتریان این امکان را می‌دهد که خصوصیات محصولات مورد نظر را از نظر قیمت، قابلیت اطمینان و جنبه‌های دیگر مقایسه کرده و بهترین گزینه را انتخاب کنند. با توجه به این مسأله، شرکت‌های معتبر تولیدی در این رقابت به روش‌های آماری مختلف، برای بالا بردن قابلیت اطمینان محصولات خود نیاز دارند.

از سویی دیگر، با افزایش آمار مرگ و میر، شیوع بیماری‌های مختلف و پیشرفت‌های پزشکی این سوالات مطرح می‌شود که

۱. طول عمر یک انسان سالم چقدر است؟

۲. چه عواملی بر طول عمر یک انسان اثرگذار هستند؟

۳. در صورتی که فردی دچار بیماری خاصی شود و یا تحت

شرایطی درمان شود طول عمر آن چه خواهد بود؟

امروزه تحلیل بقا در اکثر مطالعات علمی که شامل بررسی مدت زمان وقوع یک پیشامد باشد، مورد استفاده قرار می‌گیرد. در کاربرد توزیع داده‌های طول عمر نامعلوم است. بدین سبب به‌کار بردن روش‌های ناپارامتری در مقایسه با روش‌های پارامتری معمول‌تر است. در مطالعه‌های بقا اگر هدف، توصیف زمان بقا بدون در نظر گرفتن متغیر کمکی باشد، از روش‌های ناپارامتری غیر وابسته همچون جداول عمر و روش کاپلان-مایر استفاده می‌شود. با این حال تعیین عوامل و متغیرهای اثرگذار در زمان رخداد پدیده‌ها از اهمیت زیادی برخوردار است. برای مدل‌سازی چنین عواملی در تحلیل داده‌های بقا و پیش‌بینی از مدل‌های رگرسیون بقا مانند مدل کاکس به عنوان یک روش نیمه ناپارامتری استفاده می‌شود که نخستین بار توسط کالب فلیش [۱۳] با در نظر گرفتن پیشین فرآیند گاما برای تابع توزیع تجمعی هازارد در مدل کاکس به کار گرفته شد. همچنین کریستنسن و جانسون [۳] تحلیل بیز برای مدل‌های زمان شکست شتابیده با در نظر گرفتن پیشین فرآیند دیریخله را مطرح کردند. والکر و مالیک [۱۶] با به کارگیری روش‌های MCMC و با بهره‌گیری از پیشین فرآیند درخت پولیا این شیوه‌های تحلیل بقا را توسعه دادند. همچنین در سال‌های اخیر پیشین فرآیند درخت پولیای آمیخته برای تابع بقای پایه در مدل‌های کاکس، زمان شکست شتابیده و بخت‌های متناسب استفاده شده است. هانسون و جانسون [۱۱] و هانسون، جانسون و لاد [۱۲]، کیو و مالیک [۱۴] و ژلفاند و کوتاس [۱۰] از پیشین فرآیند دیریخله آمیخته برای تابع بقای پایه در مدل‌های زمان شکست شتابیده استفاده کردند. اما این مدل‌ها نیز دارای فرضیات محدود کننده‌ای هستند.

برای حذف چنین محدودیت‌هایی استفاده از شیوه‌های ناپارامتری راه حل مناسبی است. با این حال شیوه‌های ناپارامتری فراوانی گرای موجود انعطاف لازم برای مدل‌بندی اثرات متغیرهای کمکی و عوامل تأثیرگذار را ندارند. بنابراین شیوه‌های بیز ناپارامتری برای این منظور به کار گرفته می‌شود. شیوه‌های بیز ناپارامتری وابسته با در نظر گرفتن مدل‌های انعطاف پذیر رگرسیونی در ابر پارامترهای توزیع‌های پیشین خود می‌توانند جایگزین مناسبی برای شیوه‌های پارامتری و نیمه پارامتری وابسته مانند مدل کاکس باشند. در این مقاله به طور خاص به فرآیند دیریخله وابسته و کاربرد آن برای تحلیل رگرسیون بقا برای داده‌های سانسور شده می‌پردازیم. فرآیند دیریخله یکی از متداول‌ترین شیوه‌های بیز ناپارامتری است که دارای خصوصیات بسیار خوب نظری و محاسباتی است.

ساختار کلی این مقاله بدین صورت است که ابتدا مفهوم‌ها و تعریف‌هایی از تحلیل بقا، نمادهای مورد نیاز آن و نیز چند روش خاص تحلیل بقا معرفی خواهد شد. پس از آن به بیان فرایند دیریخله و نیز رهیافت بیز ناپارامتری پرداخته خواهد شد. همچنین در این مقاله تعمیم‌های این رهیافت از جمله فرایند دیریخله وابسته و فرایند دیریخله سلسله مراتبی و نیز نحوه محاسبه توزیع پسین در این مدل با استفاده از روش الگوریتم گیبس بیان می‌شوند. در ادامه روش‌های بیز ناپارامتری بر اساس پیشین فرایند دیریخله سلسله مراتبی وابسته، در حالتی که داده‌ها سانسور شده از راست باشند و نیز الگوریتم‌های لازم بیان خواهد شد. در پایان مطالب بالا برای بررسی داده‌های مربوط به بیماران سرطانی مورد استفاده قرار می‌گیرد.

## ۲ تحلیل بقا

تحلیل بقا مجموعه‌ای از روش‌های آماری تحلیل داده‌های طول عمر است. طول عمر یک متغیر تصادفی مثبت است

که می‌تواند زمان شکست یک مؤلفه فیزیکی و یا زمان مرگ یک واحد زنده باشد. فرض کنید یک سیستم، یا قطعه الکتریکی، یا هر وسیله یا شیء دیگر، دارای طول عمر $T$ است و مقدارهای خود را در بازه‌ی $(۰, \infty)$ اختیار می‌کند. اگر $f(t)$ و $F(t)$ به ترتیب نمایانگر تابع چگالی احتمال و تابع توزیع تجمعی احتمال این متغیر تصادفی باشند، آنگاه $S(t) = ۱ - F(t) = P(T > t)$ را تابع بقای سیستم تا زمان $t$ می‌گوییم. معمولا در کاربردهای غیر پزشکی به تابع بقا، تابع قابلیت اعتماد گفته می‌شود و آن را با $R(t)$ نمایش می‌دهند.

یکی دیگر از معیارهای مهم در مطالعه قابلیت اطمینان و طول عمر تابع نرخ خطر است. تابع نرخ خطر احتمال شکست یک موجود در یک بازه زمانی کوچک $(\delta)$ به شرط آن که تا زمان $t$ از کار نیفتاده باشد را بیان می‌کند و آن را با $h(t)$ نمایش می دهند.

روش‌های محاسبه تابع بقا بر اساس دیدگاه‌های پارامتری و ناپارامتری و نیز فراوانی‌گرا یا بیزی دسته‌بندی می‌شوند. به عنوان مثال روش درستنمایی ماکزیمم دیدگاه پارامتری فراوانی‌گرا است که در آن، براورد تابع چگالی براساس براورد پارامترها انجام می‌شود. از دیگر نمونه‌های این دیدگاه می‌توان به روش نیمه پارامتری کاکس اشاره کرد که در آن، متغیرهای کمکی نیز در مدل‌بندی نقش دارند و بر پایه فرض متناسب بودن تابع‌های نرخ خطر مشاهده‌ها و استقلال زمان‌های رخداد پیشامدها تشکیل شده است. دیدگاه بیز پارامتری شیوه‌ای استنباطی پارامتری براساس اطلاعات پیشین در خصوص پارامترهای مجهول است. در این شیوه پارامترهای مجهول توزیع پارامتری معلوم به عنوان متغیرهای تصادفی در نظر گرفته شده و با توجه به آگاهی‌های پیشین برای آن‌ها توزیع‌های پیشین در نظر گرفته می‌شود. در دیدگاه فراوانی گرای ناپارامتری همچون روش کاپلان مایر[1]، بدون در نظر گرفتن توزیع معینی برای نمونه تصادفی با روش‌های آزاد توزیع به براورد تابع مورد استفاده در تحلیل بقا مانند تابع توزیع، تابع چگالی و تابع نرخ خطر پرداخته می‌شود. دیدگاه روش‌های بیز ناپارامتری که موضوع اصلی این مقاله نوع خاصی از این شیوه‌ها است، به وسیله دو ویژگی فضای پارامتر نامتناهی-بعد و نیز ترکیب اندازه‌های احتمال روی فضاهای احتمال نامتناهی-بعد از دیگر شیوه‌ها متمایز می‌شود که در بخش سوم به تفصیل شرح داده خواهد شد.

## ۳ استنباط بیز ناپارامتری بر اساس فرآیند دیریخله

در شیوه استنباط بیزی پارامتری، برای نمونه‌های متناهی یک توزیع معلوم با پارامترهای نامعلوم در نظر گرفته می‌شود و با در نظر گرفتن پارامترهای مدل به عنوان متغیرهای تصادفی و با تعیین توزیع پیشین برای آن‌ها، به استنباط بیزی در خصوص پارامترها می‌پردازیم. در عمل با موارد بسیاری روبرو هستیم که تعیین توزیع معلوم برای داده‌ها دور از واقعیت است. در این شرایط می‌توان از استنباط بیزی ناپارامتری استفاده کرد.

در شیوه‌های بیز ناپارامتری تابع احتمال را به عنوان یک فرآیند تصادفی روی $\sigma-$جبر فضای نمونه در نظر می‌گیریم. به عبارت دیگر تابع احتمال $P$ که تمامی تابع‌های مورد استفاده در تحلیل بقا را می‌توان بر اساس آن تعریف کرد را روی $\sigma-$جبر $\mathcal{A}$ که $P$ روی آن تعریف شده است به عنوان یک فرآیند تصادفی در نظر می‌گیریم به طوری که به ازای هر پیشامد $A \in \mathcal{A}$، $P(A)$ یک متغیر تصادفی در نظر گرفته می‌شود. فرآیندهای پیشین متفاوتی برای تحلیل

[1] Kaplan-meier

بیزی ناپارامتری معرفی شده‌اند، که معروف‌ترین آن‌ها فرآیند دیریخله[۲] است، و توسط فرگوسن [۱۹] معرفی شده است. امروزه روش‌های بیز ناپارامتری بسیار متنوع و هر یک دارای خصوصیت‌های فراوان و پرکاربرد هستند.

برای تعریف فرآیند دیریخله ابتدا لازم است با توزیع دیریخله [۳] آشنا شویم.

گوییم $(p_۱, p_۲, \cdots, p_k)$ به ازای $p_i \geq ۰ \quad , \quad \sum_{i=۱}^{k} p_i = ۱$ دارای توزیع دیریخله با پارامترهای $(\alpha_۱, \ldots, \alpha_k)$ است و با نماد

$$(p_۱, p_۲, \cdots, p_k) \sim \ \mathrm{Dir}(\alpha_۰, \alpha_۱, \ldots, \alpha_k)$$

نشان می‌دهیم، اگر $p = (p_۱, p_۲, \cdots, p_k)$ دارای تابع چگالی احتمال زیر باشد

$$f(p_۱, p_۲, \cdots, p_k) = \frac{\Gamma(\sum_{i=۱}^{k} \alpha_i)}{\prod_{i=۱}^{k} \Gamma(\alpha_i)} \prod_{i=۱}^{k} P_i^{\alpha_i - ۱}, \quad (۱)$$

که در آن $\Gamma(.)$ تابع گاما را نشان می‌دهد.

حال به تعریف فرآیند دیریخله می‌پردازیم.

**تعریف:** فرض کنید $\Theta$ فضای نمونه و $\alpha(.)$ یک اندازه جمع پذیر روی $\Theta$ باشد. می‌گوییم فرآیند تصادفی $P(.)$ روی فضای نمونه $\Theta$ یک فرآیند دیریخله با پارامتر $\alpha(.)$ است، اگر به ازای هر افراز متناهی اندازه پذیر $A_۱, \ldots, A_r$، $r \geq ۱$، از فضای نمونه $\Theta$، بردار تصادفی $(P(A_۱), \ldots, P(A_r))$ دارای توزیع دیریخله با پارامترهای $(\alpha(A_۱), \ldots, \alpha(A_r))$ است و آن را با نماد $P \sim DP(\alpha)$ نشان می‌دهیم.

بنابراین اگر افراز $(-\infty, x] \cup (x, \infty)$ از فضای نمونه $\Theta = \mathbb{R}$ را در نظر بگیریم و تابع توزیع تجمعی متناظر با $P(.)$ را به صورت $G(x) = P((-\infty, x])$ تعریف کنیم، داریم

$$G(x) \sim \mathrm{Beta}\left(\alpha((-\infty, x]), \alpha((x, \infty))\right) , \quad (۲)$$

که یک توزیع پیشین مناسب برای توزیع تجمعی است. در این صورت

$$E(G(x)) = \frac{\alpha(-\infty, x]}{\alpha(\mathbb{R})}$$

و

$$\mathrm{Var}(G(x)) = \frac{\alpha(-\infty, x]\alpha(x, \infty)}{\alpha(\mathbb{R}) + ۱}$$

گاهی اوقات اندازه $\alpha(A)$ را به صورت $\alpha(A) = \alpha_۰ Q(A)$ تعریف می‌کنند، که در آن $Q$ یک تابع احتمال روی $\Theta$ و $\alpha_۰ > ۰$ یک پارامتر تمرکز است. در این صورت اگر $G_۰$ تابع توزیع تجمعی متناظر با $Q$ باشد داریم:

$$E(G(x)) = G_۰(x), \quad \mathrm{Var}(G(x)) = \frac{G_۰(x)(۱ - G_۰(x))}{\alpha_۰ + ۱}, \quad (۳)$$

توزیع $G_۰$ را می‌توان به عنوان حدسی از $G$ و نیز پارامتر دقت $\alpha_۰$ را به عنوان درجه تمرکز توزیع $G$ و یا شدت قطعیت $G$ حول $G_۰$ در نظر گرفت. برای مقادیر بزرگ $\alpha_۰$، یک نمونه از $G$ با احتمال زیاد مقادیر نزدیک به $G_۰$ را می‌گیرد و برای مقادیر کوچک $\alpha_۰$، نمونه $G$ با احتمال زیاد جرم احتمال زیادی را تنها در تعداد معین اتم‌ها می‌گذارد. همچنین نشان داده شده است که فرآیند دیریخله یک پیشین مزدوج است. بنابراین توزیع پسین نیز فرآیند دیریخله است. برای مطالب بیشتر در این خصوص می‌توانید به اسکوبار و وست [۴] و وست، مولر و اسکوبار [۱۷] مراجعه کنید.

در صورتیکه در تعریف فوق $G$ نامعلوم باشد، می‌توان براساس دیدگاه بیز ناپارامتری برای $G$ توزیع پیشینی در نظر گرفت. اگر این توزیع پیشین فرآیند دیریخله باشد،

[۲] Dirichlet Process
[۳] Dirichlet distribution

مدل حاصل، مدل فرآیند دیریخله آمیخته نامیده می‌شود. بنابراین مدل فرآیند دیریخله آمیخته به صورت زیر تعریف می‌شود:

$$Y_i|\theta_i \sim H(.|\theta_i)\ , \quad \theta_i|G \sim G(.)\ , \quad G|\alpha \sim \mathrm{DP}(.|\alpha). \tag{۴}$$

به عبارت دیگر، اگر $Y_1, Y_2, \ldots, Y_n$ هم توزیع با توزیع نامعلوم $F$ باشند، آن‌گاه

$$F(y_i) = \int H(y_i \mid \theta) dG(\theta)\ , \quad G \sim \mathrm{DP}(\alpha)$$

که در آن $H(. \mid \theta)$ یک توزیع پارامتری است، که هسته آمیخته نامیده می‌شود و توسط یک پارامتر با بعد محدود $\theta$ اندیس‌گذاری شده است.

برای یافتن یک برآورد بیز ناپارامتری برای $G(x)$، تعریف آمیخته فرایند دیریخله ضروری است. آمیخته فرآیند دیریخله تعمیم فرآیند دیریخله است، به صورتی که توزیع نهایی پیشین آمیخته‌ای از فرآیند دیریخله با اندازه‌های پایه گوناگون است. با در نظر گرفتن توزیع آمیزنده[۴] $H$ می‌نویسیم

$$G\Big|\eta \sim \mathrm{DP}(\alpha_\eta)\ , \quad \eta \sim H(\eta) \tag{۵}$$

و در نتیجه

$$G \sim \int \mathrm{DP}(\alpha_\eta) dH(\eta). \tag{۶}$$

برای مدل (۴)، توزیع پسین $G$ به صورت زیر به دست می‌آید:

$$G|Y_1, Y_2, \ldots, Y_n) \sim \int \mathrm{DP}(\alpha + \textstyle\sum_{i=1}^{n} \delta_{\theta_i}) \quad (۷)$$
$$dP(\theta_1, \ldots, \theta_n \Big| Y_1, Y_2, \ldots, Y_n),$$

که در آن $\delta$ تابع نشانگر و $dP(\theta_1, \ldots, \theta_n \Big| Y_1, Y_2, \ldots, Y_n)$ توزیع پسین $\theta_1, \theta_2, \ldots, \theta_n$ به شرط $Y_1, \ldots, Y_n$ است. بنابراین یک براورد بیز ناپارامتری برای $G(x)$ براساس تابع زیان درجه دوم خطا برابر است با

$$E\left((G(x) \mid Y_1, \ldots, Y_n)\right) = \quad (۸)$$
$$\int \frac{\alpha((-\infty, x]) + \sum_{i=1}^{n} \delta_{\theta_i}((-\infty, x)}{\alpha(\mathbb{R}) + n}$$
$$dP(\theta_1, \ldots, \theta_n \Big| Y_1, Y_2, \ldots, Y_n).$$

با استفاده از نمونه به دست آمده، نمونه بردار گیبس و رابطه (۷) می‌توان یک براورد بیز ناپارامتری از $G(x)$ با تقریب رابطه (۸) به صورت زیر به دست آورد:

$$\frac{1}{M} \sum_{j=1}^{M} \int \frac{\alpha((-\infty, x]) + \sum_{i=1}^{n} \delta_{\theta_{ij}}((-\infty, x)}{\alpha(\mathbb{R}) + n},$$

که در آن $\theta_j = (\theta_{1j}, \ldots, \theta_{nj}), j = 1, \ldots, M$، نمونه‌های تولید شده در نمونه‌برداری گیبس پس از طی دوره داغیدن هستند. جهت اطلاعات بیشتر در رابطه با روش‌های شبیه‌سازی مونت کارلو مانند نمونه بردار گیبس به داس و هافر [۷] مراجعه شود.

[۴] Mixing distribution

از طرفی بر اساس مدل بلکول و مک کویین [۲] می‌توان

نوشت

$$G(\theta_{n+1}|\theta_1,\theta_2,\ldots,\theta_n) \propto \alpha((-\infty,\theta_{n+1}]) + \sum_{j=1}^{n}\delta_{\theta_j}((-\infty,\theta_{n+1}]) = \alpha((-\infty,\theta_{n+1}]) + \sum_{j=1}^{n} n_j\delta_{\theta_j^*}((-\infty,\theta_{n+1})), \quad (۹)$$

که در آن $\theta_j^*$ها مقدارهای متمایز $\theta_1,\ldots,\theta_n$ و $n_j$ تعداد تکرارها $\theta_j^*$ در میان $\theta_1,\ldots,\theta_n$ است. فرض کنید $\theta = (\theta_1,\theta_2,\ldots,\theta_n)$ و $Y = (Y_1,Y_2,\ldots,Y_n)$ باشد. بنابراین توزیع پسین $P(\theta|Y)$ در (۷) به صورت زیر به‌دست می‌آید:

$$P(\theta|Y) \propto H(Y|\theta)\cdot G(\theta), \quad (۱۰)$$

که در آن $G(\theta) = \prod_{i=1}^{n} G(\theta_i|\theta_1,\theta_2,\ldots,\theta_{i-1})$ براساس رابطه (۱۰) به دست می‌آید. بنابراین داریم

$$P(\theta|Y) \propto \prod_{i=1}^{n} H(Y_i|\theta_i)\cdot\alpha((-\infty,\theta_i])+\sum_{j=1}^{i-1}\delta_{\theta_j}((-\infty,\theta_i]). \quad (۱۱)$$

برای $\alpha(A) = \alpha_\circ Q(A)$ و با در نظر گرفتن $G_\circ$ به عنوان تابع توزیع تجمعی متناظر با $Q$ و همچنین مقدارهای غیر تکراری $\theta_j^*$ با فراوانی $n_j$، $j = 1,\ldots,k_i$ در دنباله $\theta_1,\ldots,\theta_n$، می‌توان نوشت

$$P(\theta|Y) = \prod_{i=1}^{n} H(Y_i|\theta_i)\cdot\mathcal{P}_{\theta_i}, \quad (۱۲)$$

که در آن

$$\mathcal{P}_{\theta_i} = \frac{\alpha_\circ G_\circ(\theta_i) + \sum_{j=1}^{k_i} n_{ij}\delta_{\theta_j^*}((-\infty,(\theta_i])}{\alpha_\circ + i - 1}.$$

در نمونه‌گیری گیبس از توزیع $G(\theta)$، برای تولید نمونه $\theta_1,\ldots,\theta_n$ از توزیع شرطی هر $\theta_i$ به شرط

$$\theta_{-i} = \{\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_n\}$$

به صورت زیر استفاده می‌شود.

$$G(\theta_i|\theta_{-i}) \propto \alpha_\circ G_\circ(\theta_i) + \sum_{j=1,j\neq i}^{n}\delta_{\theta_j}((-\infty,\theta_i]). \quad (۱۳)$$

بنابراین می‌توان توزیع پسین شرطی را به صورت زیر نوشت

$$P(\theta_i|\theta_{-i},Y) \propto H(Y|\theta)\cdot G(\theta_i|\theta_{-i} = \alpha_\circ H(Y_i|\theta_i)G_\circ(\theta_i) + \sum_{j=1,j\neq i}^{n} H(Y_i|\theta_j)\delta_{\theta_j}((-\infty,\theta_i]) \quad (۱۴)$$

از طرفی

$$P(\theta_i|Y_i) = \frac{H(Y_i|\theta_i)G_\circ(\theta_i)}{\int H(Y_i|\theta_i)dG_\circ(\theta_i)}, \quad (۱۵)$$

بنابراین

$$\alpha_\circ H(Y_i|\theta_i)G_\circ(\theta_i) = \alpha_\circ\int H(Y_i|\theta_i)dG_\circ(\theta_i)P(\theta_i|Y_i), \quad (۱۶)$$

از طرف دیگر از آن‌جا که از رابطه (۱۵) داریم

$$P(\theta_i|Y_i) \propto H(Y_i|\theta_i)G_\circ(\theta_i).$$

می‌توان نوشت

$$\alpha_\circ H(Y_i|\theta_i)G_\circ(\theta_i) \propto q_\circ t H(Y_i,\theta_i)G_\circ(\theta_i),$$

که در آن $q_0 \propto \int H(Y_i,\theta_i)dG_0(\theta_i)$ است.
بنابراین

$$P(\theta_i|\theta_{-i},Y) = q_0 H(Y_i|\theta_i)G_0(\theta_i) + \sum_{j=1,j\neq i}^{n} q_j \delta_{\theta_j}((-\infty,\theta_i]) \quad (۱۷)$$

که در آن $q_i \propto H(Y_i|\theta_j)$ و $q_0 + \sum_{j\neq i} q_j = 1$ است.
بنابراین نمونه بردار گیبس در سه گام نمونه‌ای از $\theta$ را از طریق رابطه (۱۷) به صورت زیر تولید می‌کند:
**گام۰:** یک مقدار اولیه برای $\theta = (\theta_1,\theta_2,\ldots,\theta_n)$ انتخاب می‌کنیم.
**گام۱:** مقدار $\theta_1$ را از توزیع $P(\theta_1|\theta_{-1},Y)$ تولید می‌کنیم، سپس از جایگذاری مقدار $\theta_1$ در $\theta$ مقدار $\theta_2$ را از توزیع $P(\theta_2|\theta_{-2},Y)$ تولید می‌کنیم و این کار را تا تولید $\theta_n$ ادامه می‌دهیم.
**گام۲:** گام ۱ را تا رسیدن به همگرایی الگوریتم تکرار می‌کنیم. بنابراین تعداد نمونه‌هایی که تا این جا تولید کرده‌ایم به عنوان دوره سوخت کنار می‌گذاریم و تعداد نمونه‌های پس از همگرایی را به عنوان نمونه تولید شده از پسین در نظر می‌گیریم.

# ۴ مدل بیز ناپارامتری دیریخله سلسله مراتبی

روش بیز سلسله مراتبی یک روش اساسی در آمار بیزی است. در مدل‌های بیز سلسله مراتبی برای پارامترهای توزیع پیشین (ابر پارامترها) توزیع‌های پیشینی در نظر گرفته می‌شود. به کارگیری روش‌های سلسله مراتبی در بیز ناپارامتری منجر به مدل‌های بیز ناپارامتری سلسله مراتبی می‌شود. در این بخش ابتدا به معرفی این مدل پرداخته می شود و پس از آن روش محاسبه توزیع پسین در حالتی که مدل فرایند دیریخله سلسله مراتبی باشد مختصرا بیان خواهد شد. به منظور معرفی فرایند دیریخله سلسله مراتبی داریم:

فرض کنید $Y_1,Y_2,\ldots,Y_n$ نمونه‌ای تصادفی باشند به طوری که برای $i = 1,\ldots,n$، $Y_i \sim H(.|\theta_i,\zeta_i)$ و $Y_i$ ها به شرط $\theta_i$ و $\zeta_i$ مستقل هستند. هم چنین قرار دهید

$$Y = (Y_1,Y_2,\ldots,Y_n)\ , \quad \theta = (\theta_1,\theta_2,\ldots,\theta_n)\ ,$$

$$\zeta = (\zeta_1,\zeta_2,\ldots,\zeta_n)\ .$$

توزیع‌های پیشین زیر را در نظر بگیرید.

$$\theta_i \overset{i.i.d}{\sim} G(.|\lambda)\ , \quad \zeta_i \overset{i.i.d}{\sim} H_1(.)\ , \quad \lambda \sim H_2(.).$$

علاوه بر آن فرض کنید $Y$ و $\zeta$ به شرط $\theta$ از $\lambda$ مستقل هستند.

همچنین فرض کنید که $\theta$ و $G$ به شرط $Y$، $\zeta$ و $\boldsymbol{\lambda}$ مستقل از هر پارامتر دیگر در مدل سلسله مراتبی باشند. عبارت سلسله مراتبی در این جا، درگسترده ترین معنی آن، به معنی یک مدل دارای سطوح مختلف است که به وسیله دنباله‌ای از توزیع‌های شرطی آشیانه‌ای مدل‌بندی شده باشد. یک مدل پارامتری بیز سلسله مراتبی که می‌خواهیم به توصیف آن بپردازیم، مدلی است که عدم قطعیت در خصوص شکل تابعی توزیع پیشین $G$ را توسط یک مدل پیشین فرآیند دیریخله مدل‌بندی می‌کند. بنابراین می توان چنین ادعا کرد که محدودیت معلوم بودن شکل توزیع پیشین که در استنباط بیز پارامتری مفروض است، در این مدل وجود ندارد.

## ۱.۴ محاسبه توزیع پسین در مدل فرایند دیریخله سلسله مراتبی

به منظور نمونه‌گیری از پارامتر $\theta$ با توجه به بخش سوم و رابطه (۱۷) می‌دانیم که

$$P(\theta_i|\theta_{-i},Y,\zeta,\lambda) \propto q_0 H(Y_i|\theta_i,\zeta_i) G_0(\theta_i|\lambda) + \sum_{j=1, j\neq i}^{n} q_j \delta_{\theta_j}((-\infty,\theta_i]), \quad (۱۸)$$

که در آن داریم

$$q_0 \propto \alpha_0 \int H(Y_i|\theta_i,\zeta_i)\, \mathrm{d}G_0(\theta_i|\lambda),$$

همچنین $q_j \propto H(Y_i|\theta_j,\zeta_i)$ و $q_0 + \sum_{j\neq i} q_j = 1$ است.

همچنین همان‌طور که در بخش سوم گفتیم رابطه (۱۸) را می‌توان با خوشه‌بندی $\theta_i$ ها به صورت $\theta_j^*, j = 1,\ldots,k$ در $k$ خوشه با تعداد تکرار $n_j$ برای $\theta_j^*$ متمایز در مجموعه $\theta_1,\ldots,\theta_n$ به صورت زیر بازنویسی کرد:

$$P(\theta_i|\theta_{-i},Y,\zeta,\lambda) \propto q_0 H(Y_i|\theta_i,\zeta_i) G_0(\theta_i|\lambda) + \sum_{j=1,\theta_j^*\neq\theta_i}^{k} n_j q_j^* \delta_{\theta_j^*}((-\infty,\theta_i]), \quad (۱۹)$$

که در آن $q_j^* \propto H(Y_i|\theta_j^*,\zeta_i)$ و $q_0 + \sum_j n_j q_j^* = 1$ است.

علاوه بر آن برای جلوگیری از پایدار شدن زنجیر مارکف روی تعداد کمی از خوشه‌ها می‌توان از نمونه‌گیری خوشه‌ها به شیوه بازآمیختن با استفاده از توزیع پسین زیر نمونه‌گیری کرد.

$$P(\theta_j^*|Y,S,k,\zeta,\lambda) \propto \prod_{i\in J_j} \int H(Y_i|\theta_j^*,\zeta_i)\, \mathrm{d}G_0(\theta_i|\lambda). \quad (۲۰)$$

در مدل سلسله مراتبی ابرپارامترهای $\zeta$ و $\lambda$ نیز دارای توزیع‌های پسینی هستند که می‌توان از آن‌ها در الگوریتم نمونه بردار گیبس برای نمونه‌گیری از این ابرپارامترها و نهایتاً نمونه‌گیری از $\theta_i$ ها استفاده کرد. همان طور که در ابتدای بخش گفتیم معمولاً فرض می‌شود که $\zeta$ به شرط $Y$ و $\theta$ از $\lambda$ و $\alpha_0$ مستقل است. بنابراین

$$\Pi(\zeta|Y,\theta,\lambda,\alpha_0) = \Pi(\zeta|Y,\theta) \propto \prod_{i=1}^{n} H(Y_i|\theta_i,\zeta_i) G_0(\theta_i|\lambda) H_1(\zeta). \quad (۲۱)$$

همچنین همان‌طور که در ابتدای بخش بیان کردیم توزیع شرطی ابر پارامتر $\lambda$ به شرط $\alpha_0$ و $\zeta$ و $\theta$ و $Y$ تنها به $\theta$ بستگی دارد. بنابراین

$$\Pi(\lambda|Y,\theta,\zeta,\alpha_0) = \Pi(\lambda|Y,\theta) \propto G_0(\theta|\lambda) H_2(\lambda). \quad (۲۲)$$

سپس با توجه به عبارات فوق، با استفاده از الگوریتم نمونه‌گیری گیبس اقدام به نمونه‌گیری از $\theta$ می کنیم. اما با توجه مطالب بیان شده در بخش‌های فوق، انتخاب یک مقدار ثابت برای $\alpha_0$ در توزیع پیشین فرآیند دیریخله به طرز قابل توجهی روی نمونه‌گیری از $\theta_i$ ها تأثیر می‌گذارد. به منظور کاهش میزان تاثیر این پارامتر می‌توان یک مدل سلسله مراتبی برای این پارامتر با در نظر گرفتن یک توزیع پیشین برای $\alpha_0$ و یادگیری در مورد آن در نمونه برداری گیبس ساخت که به شرح زیر می‌باشد:

فرض کنید $p(\alpha_0)$ چگالی پیشین پیوسته برای $\alpha_0$ باشد، که می‌تواند به حجم نمونه، $n$، وابسته باشد.

آنتونیاک [۱] نشان داد که توزیع تعداد خوشه‌ها، $k$، به شرط معلوم بودن $n$ و $\alpha_0$ به صورت زیر است:

$$p(k|\alpha_0,n) = C_n(k)\alpha_0^k f\Gamma(\alpha_0)\Gamma(\alpha_0+n), \quad k = 1,\ldots,n, \quad (۲۳)$$

که در آن ثابت $C_n(k)$ به $\alpha_\circ$ بستگی ندارد. البته در صورت نیاز ثابت‌های $C_n(k)$ به آسانی و با استفاده از رابطه بازگشتی برای اعداد استرلینگ[۵] محاسبه می‌شوند.

همچنین

$$p\left(k|n\right)=E\left(p\left(k|\alpha_\circ,n\right)\right)=\int_\circ^\infty p\left(k|\alpha_\circ,n\right)p(\alpha_\circ)\mathrm{d}\alpha_\circ\,.$$

با فرض این که توزیع $\alpha_\circ$ به شرط $k$ و $n$ از $\theta$ و $Y$ مستقل باشد، توزیع پسین $\alpha_\circ$ به شرط معلوم بودن $k$ و $n$ برابر است با

$$\begin{aligned} p\left(\alpha_\circ|k,n,\theta,Y\right) &= p\left(\alpha_\circ|k,n\right) \\ &= \tfrac{p(k|\alpha_\circ,n)p(\alpha_\circ)}{p(k|n)}\,. \end{aligned} \qquad (۲۴)$$

معمولاً توزیع پیشین برای $\alpha_\circ$ توزیع گاما یا آمیخته‌ای از چند توزیع گاما در نظر گرفته می‌شود. در این حالت نمونه‌گیری از توزیع پسین $\alpha_\circ$ به سادگی انجام می‌پذیرد. در این جا ما توزیع پیشین $\alpha_\circ$ را گاما در نظر می‌گیریم. می‌توانید تعمیم این پیشین به آمیخته‌ای از چند پیشین گاما را در وست [۱۸] ببینید.

با در نظر گرفتن $\alpha_\circ\sim\Gamma(a,b)$ داریم:

$$p\left(\alpha_\circ|k,n\right)\propto\frac{\alpha_\circ^{k+\alpha+۲}\Gamma(\alpha_\circ)e^{-\alpha_\circ b}}{\Gamma(\alpha_\circ+n)}\,, \qquad (۲۵)$$

که توزیع پسین مشخص و سهل‌الوصولی نیست و نمونه‌گیری از آن به راحتی امکان‌پذیر نیست.

یک راه‌کار برای نمونه گیری از پسین $\alpha_\circ$ در این حالت، که به آن افزایش داده[۶] گفته می‌شود، به صورت زیر است.

با توجه به رابطه

$$\frac{\Gamma(\alpha_\circ)}{\Gamma(\alpha_\circ+n)}=\frac{(\alpha_\circ+n)}{\Gamma(n)}\int_\circ^۱\eta^{\alpha_\circ}(۱-\eta)^{n-۱}\mathrm{d}\eta\,,$$

رابطه (۲۳) را می‌توان به صورت زیر نوشت:

$$p(k|\alpha_\circ,n)=C_n(k)\alpha_\circ^{k-۱}\frac{(\alpha_\circ+n)}{\Gamma(n)}\int_\circ^۱\eta^{\alpha_\circ}(۱-\eta)^{n-۱}\mathrm{d}\eta,$$

که در آن $k=۱,...,n$. بنابراین

(۲۶)

$$p\left(\alpha_\circ|k,n\right)\propto\left(\alpha_\circ+n\right)\alpha_\circ^{k+a-۲}e^{-\alpha_\circ b}\int_\circ^۱\eta^{\alpha_\circ}(۱-\eta)^{n-۱}\mathrm{d}\eta.$$

حال متغیر پنهان $\eta$ را به صورت زیر در نظر بگیرید (اسکوبار و وست [۴])

$$\left(\eta|\alpha_\circ,n\right)\propto\mathrm{Beta}\left(\alpha_\circ+۱,n\right). \qquad (۲۷)$$

با دقت در رابطه‌های (۲۵) و (۲۶) می‌توان دریافت که $p\left(\alpha_\circ|k,n\right)$ توزیع حاشیه‌ای $(\alpha_\circ,\eta)$ برای $\alpha_\circ$ است که از توزیع توأم $\alpha_\circ$ و $\eta$ : به صورت زیر به دست آمده است

$$\begin{aligned} p\left(\alpha_\circ,\eta|k,n\right) &\propto \alpha_\circ^{a+k-۲}(\alpha_\circ+n) \\ &\quad \exp\left\{-\alpha_\circ\left(b-\log(\eta)\right)\right\} \\ &\propto \alpha_\circ^{a+k-۱}\exp\left\{-\alpha_\circ\left(b-\log(\eta)\right)\right\} \\ &\quad +n\alpha_\circ^{a+k-۲}\exp\left\{-\alpha_\circ\left(b-\log(\eta)\right)\right\}. \end{aligned} \qquad (۲۸)$$

بنابراین توزیع پسین $\alpha_\circ$ به شرط متغیر پنهان $\eta$، $p\left(\alpha_\circ|k,n,\eta\right)$، به صورت آمیخته‌ای از دو توزیع گاما به

[۵] Stirling numbers
[۶] Data augmentation

صورت زیر است:

$$p(\alpha_0|k,\eta) \propto \pi_\eta \Gamma(a+k, b-\log(\eta)) + (1-\pi_\eta)\Gamma(a+k-1, b-\log(\eta)), \quad (۲۹)$$

که در آن $\pi_\eta = \frac{a+k-1}{a+k-1+n(b-\log(\eta))}$ است.

بنابراین برای نمونه‌گیری از $\alpha_0$ به شرط معلوم بودن $k$ کافیست ابتدا $\eta$ را از توزیع بتا (۲۶) به شرط $\alpha_0$ مرحله قبل نمونه‌گیری گیبس تولید کنیم و سپس به شرط $\eta$ و $k$ معلوم $\alpha_0$ را از پسین آمیخته (۲۹) نمونه‌گیری نماییم.

# ۵ مدل‌های بیز ناپارامتری دیریخله سلسله مراتبی وابسته برای داده‌های سانسور شده از راست

همانطور که در بخش اول بیان شد، مدل کاکس رابطه‌ای برای مدل‌بندی تابع نرخ خطر و دیگر توابع بقا بر حسب متغیرهای پیشگو تأثیرگذار بر روی بقا می باشد اما این مدل با فرضیات محدود کننده متناسب بودن مخاطرات و استقلال زمان‌های رخداد پیشامدها در نظر گرفته شده است. مدل‌های نیمه پارامتری دیگر نظیر مدل زمان شکست شتابیده[۷] و یا مدل بخت‌های متناسب[۸] نیز با در نظر گرفتن مدل‌های پارامتری برای توابع بقا دارای چنین فرض‌های محدود کننده‌ای هستند. برای داده‌های بقای پیچیده، پیش فرض‌های این گونه مدل‌ها مشکلات و محدودیت‌هایی در مدل سازی به وجود می‌آورند. استفاده از شیوه‌های بیز ناپارامتری در تحلیل مدل‌های رگرسیون بقا با حذف چنین محدودیت‌هایی، انعطاف‌پذیری بسیار زیادی به تحلیل داده‌های بقا بخشیده است. در این بخش ابتدا فرایند دیریخله وابسته معرفی شده و پس از آن مدل تحلیل واریانس بر پایه این فرآیند دیریخله وابسته بیان می شود.

## ۱.۵ فرآیند دیریخله وابسته

یکی از جذاب‌ترین شیوه‌های بیز ناپارامتری برای تحلیل مدل‌های رگرسیون بقا مدل فرآیند دیریخله وابسته است که در این بخش به آن می‌پردازیم. مدل فرآیند دیریخله وابسته که توسط مک ایچرن [۱۵] معرفی شد، به صورت زیر تعریف می‌شود.

فرض کنید یک مجموعه از توابع توزیع‌های $\{F_x, x \in X\}$ توسط بردار $p$ بعدی $x = (x_1, \ldots, x_p)$ اندیس گذاری شده‌اند. برای مثال در یک آزمایش بالینی $F_{(x_1,x_2)}$ تابع توزیع زمان‌های بهبودی بیماران با دو دوز دارو $x_1$ و $x_2$ می‌باشد. اگر در این مجموعه برای هر $x$، توزیع پیشین $F_x$ فرآیند دیریخله با پارامتر دقت $\alpha_0$ و اندازه پایه $F_{0x}$ باشد، $F_x \sim DP(\alpha_0 F_{0x})$، آنگاه چنین فرآیندی را فرآیند دیریخله وابسته می‌نامیم.

به بیان دقیق‌تر و کلی‌تر، فرض کنید $P_x(A)$ یک فرآیند تصادفی اندیس گذاری شده توسط $x$ و $A$ باشد، به طوری که $x$ روی $X$ و $A$ روی $\Theta$ تغییر می‌کند. همچنین فرض کنید $\alpha_x(.)$ به ازای هر $x \in X$ یک اندازه جمع پذیر روی $\Theta$ باشد. آنگاه می‌گوییم $\{P_x; x \in X\}$ یک فرآیند دیریخله وابسته است، اگر به ازای هر $x \in X$ و هر افراز $A_1, \ldots, A_r$ دلخواه از $\Theta$

$$(P_x(A_1), \ldots, P_x(A_r)) \sim \mathrm{Dir}(\alpha_x(A_1), \ldots, \alpha_x(A_r)).$$

[۷] Accelerated Failure Time
[۸] Proportional Odds

## ۲.۵ تحلیل واریانس فرآیند دیریخله وابسته

د آوریو و همکاران[۶] مدل های ناپارامتری وابسته ای را برای یک مجموعه از تابع ها یا توزیع های احتمال تصادفی در نوعی از مدل تحلیل واریانس مطرح کردند که به شرح زیر است:

فرض کنید $\{F_x, x \in X\}$ آرایه‌ای از توزیع‌های تصادفی باشد که توسط متغیر کمکی $x$ اندیس گذاری شده است. وابستگی میان توزیع‌های تصادفی توسط وابستگی تحقق‌های آنان $(\theta_{xh})$ ایجاد شده است که مدل تحلیل واریانس برای $\theta_{xh}$ بصورت زیر می باشد

$$\theta_{xh} = m_h + A_{vh} + B_h z\ , \quad h = ۱, ۲, \dots\ , \qquad (۳۰)$$

که در آن متغیر کمکی $x$ را بتوان به صورت $x = (v, z)$ افراز کرد به طوری که $v$ یک متغیر رسته‌ای شامل سطوح $۱, ۲, \dots, V$ و $z$ یک متغیر پیوسته با تکیه‌گاه $\mathbb{Z}$ باشد. برای مثال، متغیر کمکی $v$ می‌تواند سطوح یک تیمار در یک آزمایش بالینی و متغیر کمکی $z$ می‌تواند میزان یک متغیر پیوسته مربوط به هر بیمار مثل فشار خون باشد، که در این حالت $F_x$ می‌تواند توزیع تصادفی زمان بهبودی هر بیمار در سطح تیمار معین $v$ و فشار خون معین $z$ باشد. همچنین $m_h \sim p^\circ_m$، $A_{vh} \sim p^\circ_{A_v}$ و $B_h \sim p^\circ_B$ و $m_h$ها، $A_{vh}$ها و $B_h$ها به ازای مقادیر مختلف $h$ و $v$ از هم مستقل هستند. در این حالت مدل احتمالاتی توأم فرآیند $\{F_x, x \in X\}$ را تحلیل واریانس فرآیند دیریخله وابسته می‌نامیم و می‌نویسیم $\{F_x, x \in X\} \sim \text{ANOVADDP}(\alpha_\circ p^\circ)$، که در آن $p^\circ$ توزیع توأم $(m_h, A_{vh}, B_h)$ است و اندازه پایه روی اثرات تحلیل واریانس در رابطه (۳۰) نامیده می‌شود. از آن جا که برای هر $x = (v, z)$ توزیع تصادفی $F_x$ دارای پیشین فرآیند دیریخله با اندازه $\alpha_\circ F_{\circ x}$ است، بنابراین اندازه پایه $F_x$ در حالت کلی پیچش[۹] توابع $\alpha_\circ p^\circ_m$، $\alpha_\circ p^\circ_{A_v}$ و $\alpha_\circ p^\circ_{B_w}$ می‌باشد.

میزان وابستگی توزیع‌های تصادفی توسط ساختار کواریانس جرم‌های نقطه‌ای $\theta_{xh}$ تعریف می‌شود. از طرفی در مدل تحلیل واریانس استاندارد (پارامتری) سطح وابستگی توسط واریانس‌های $p^\circ_m$، $p^\circ_{A_v}$ و $p^\circ_{B_w}$ مشخص شده است. برای مثال به ازای نقاط ثابت $z_۱$ و $z_۲$، دو نقطه $x = (v, z_۱)$ و $x' = (v, z_۲)$ و نمونه‌های تصادفی $y_x \sim F_x$ و $y_{x'} \sim F_{x'}$ را در نظر بگیرید. همچنین قرار دهید $p^\circ_m = N(\mu_m, \sigma^۲_m)$، $p^\circ_{A_v} = N(\mu_{A_v}, \sigma^۲_A)$ و $p^\circ_B = N(\mu_B, \sigma^۲_B)$. در مدل تحلیل واریانس استاندارد، $\text{Cov}(y_x, y_{x'}) = \sigma^۲_m + \sigma^۲_A$ است. در مدل تحلیل واریانس فرآیند دیریخله وابسته می‌توان نشان داد

$$\text{cov}(y_x, y_{x'}) = \frac{\sigma^۲_m + \sigma^۲_A}{\alpha_\circ + ۱}.$$

در واقع کواریانس پاسخ‌های مشاهده شده $y_x$ همانند کواریانس تحت مدل تحلیل واریانس استاندارد است که توسط ضریب $\frac{۱}{\alpha_\circ + ۱}$ کاهش یافته است.

یک مزیت مدل تحلیل واریانس فرآیند دیریخله وابسته قابلیت تفسیر ساده آن بر اساس اصطلاحات مربوط به مفاهیم تحلیل واریانس استاندارد می‌باشد. در واقع $m_h$ را می‌توان به عنوان میانگین کلی[۱۰] و $A_{vh}$ و $B_h$ به عنوان اثرات اصلی[۱۱] برای سطح $v$ و متغیر پیوسته $z$ تفسیر کرد.

[۹] Convolution
[۱۰] Overall mean
[۱۱] Main effects

## ۳.۵ مدل تحلیل واریانس دیریخله سلسله مراتبی برای داده‌های سانسور شده وابسته

مدل تحلیل واریانس دیریخله سلسله مراتبی وابسته برای داده‌های سانسور شده به صورت زیر بیان می‌شود.

فرض کنید $T_1,\ldots,T_n$ مشاهدات حاصله باشند و نیز بعضی از مشاهدات از راست سانسور شده باشند، یعنی به ازای برخی از مقادیر $i$، اگر $T_i \le C_i$ باشد، $T_i$ مشاهده شده باشد و اگر $T_i > C_i$ باشد، $T_i$ سانسور شده باشد. اگر تابع چگالی متناظر با $H(T_i|\theta_i,\zeta)$ برابر $h(T_i|\theta_i,\zeta)$ باشد و مقادیر مشاهده شده $T_1^*,\ldots,T_n^*$ را به صورت $T_i^* = \min(T_i, C_i)$ تعریف کنیم، آنگاه با توجه به مطالب بیان شده در بخش سه برای $i = 1,\ldots,n$ مدل فرایند دیریخله سلسله مراتبی وابسته بصورت زیر ارائه می شود:

$$T_i|X_i = x,\theta_{ix},\zeta,\lambda \sim H(T_i|\alpha_i\eta_x,\theta_{0i},\zeta_i),$$

که در آن $x = (v,z)$، $\alpha_i = (m_i, A_{1i},\ldots,A_{Vi},B_i)$ و $\eta_x = (1, e_v, z)'$.

همچنین فرض کنید $\delta_i = I(Y_i \le C_i)$ و $h$ تابع چگالی متناظر با $H$ باشد، آنگاه برای $i = 1,\ldots,n$ :

$$Y_i^*|X_i = x,\theta_{ix},\zeta_i \sim [h(Y_i^*|\alpha_i\eta_x,\theta_{0i},\zeta_i)]^{\delta_i} \times [1 - H(Y_i^*|\alpha_i\eta_x,\theta_{0i},\zeta_i)]^{1-\delta_i}. \quad (۳۱)$$

از طرف دیگر فرض کنید

$$(\alpha_i,\theta_{0i}) \sim G$$

و

$$G(.)|\alpha_0, G_0\ ,\ \lambda \sim DP\left(.|\alpha_0 G_0(.|\lambda)\right)$$

آنگاه داریم

$$f_x(y_i^*) = \int [h(Y_i^*|\alpha_i\eta_x,\theta_{0i},\zeta_i)]^{\delta_i} [1 - H(Y_i^*|\alpha_i\eta_x,\theta_{0i},\zeta_i)]^{1-\delta_i} dG\left((\alpha_i,\theta_{0i})\right), \quad (۳۲)$$
$$G(.)|\alpha_0, G_0, \lambda \sim DP\left(.|\alpha_0 G_0(.|\lambda)\right),$$

که در آن $f_x$ نشان‌دهنده تابع چگالی احتمال متناظر با $F_x$ است.

همچنین فرض کنید

$$\zeta \sim H_1(\zeta), \quad (۳۳)$$
$$\lambda \sim H_2(\lambda),$$
$$\alpha_0 \sim \Gamma(a,b).$$

در این صورت نمونه‌گیری گیبس برای نمونه‌گیری از $(\alpha_i,\theta_{0i})$ها به صورت زیر انجام می‌شود:

**گام ۰:** پارامترهای $\alpha_0$، $\zeta$ و $\lambda$ را به ترتیب از توزیع‌های پیشین $\Gamma(a,b)$، $H_1$ و $H_2$ تولید می‌کنیم؛

**گام ۱:** برای هر $i = 1,2,\ldots,n$، $(\alpha_i,\theta_{0i})$ از توزیع پسین زیر نمونه‌گیری می‌کنیم

$$\begin{aligned}&((\alpha_i,\theta_{0i})|(\alpha_{-i},\theta_{0-i}),Y^*,\zeta,\lambda) \sim \\ &q_0\,[h(Y_i^*|\alpha_i\eta_x,\theta_{0i},\zeta_i)]^{\delta_i} \\ &\times [1 - H(Y_i^*|\alpha_i\eta_x,\theta_{0i},\zeta_i)]^{1-\delta_i}\, dG_0\left((\alpha_i,\theta_{0i})|\lambda\right) \\ &+ \textstyle\sum_{j=1,j\neq i}^{n} q_j \delta_{(\alpha_j,\theta_{0j})}(\alpha_i,\theta_{0i}),\end{aligned}$$

که در آن

$$q_{\circ} \propto \alpha_{\circ} \int [h(Y_i^*|\alpha_i\eta_x, \theta_{\circ i}, \zeta_i)]^{\delta_i} \quad (۳۴)$$
$$[۱ - H(Y_i^*|\alpha_i\eta_x, \theta_{\circ i}, \zeta_i)]^{۱-\delta_i} G_{\circ}(d(\alpha_i, \theta_{\circ i})|\lambda),$$

$$q_j \propto [h(Y_i^*|\alpha_i\eta_x, \theta_{\circ i}, \zeta_i)]^{\delta_i} [۱ - H(Y_i^*|\alpha_i\eta_x, \theta_{\circ i}, \zeta_i)]^{۱-\delta_i}$$

و $q_{\circ} + \sum_{j\neq i} q_j = ۱$ است.

**گام ۲:** پارامتر $\zeta$ را از توزیع پسین زیر نمونه‌گیری می‌کنیم:

$$\pi(\zeta|Y^*, \alpha, \theta_{\circ}, \lambda) \propto \prod_{i=۱}^{n} [h(Y_i^*|\alpha_i\eta_x, \theta_{\circ i}, \zeta_i)]^{\delta_i} [۱ - H(Y_i^*|\alpha_i\eta_x, \theta_{\circ i}, \zeta_i)]^{۱-\delta_i} \times dG_{\circ}((\alpha_i, \theta_{\circ i})|\lambda)\, h_{۱}(\zeta_i),$$

که در آن $h_{۱}$ تابع چگالی احتمال متناظر با $H_{۱}$ است.

**گام ۳:** پارامتر $\lambda$ را از توزیع پسین زیر نمونه‌گیری می‌کنیم:

$$\Pi(\lambda|Y, \alpha, \theta_{\circ}, \zeta, \alpha_{\circ}) \propto G_{\circ}(\alpha, \theta_{\circ}|\lambda) H_{۲}(\lambda).$$

**گام ۴:** پارامتر پنهان $\eta$ را از توزیع $\text{Beta}(\alpha_{\circ}+۱, n)$ و سپس پارامتر $\alpha_{\circ}$ جاری را از توزیع آمیخته ۲۷ نمونه‌گیری می‌کنیم؛

**گام ۵:** به گام ۱ باز می‌گردیم و تا زمان رسیدن به همگرایی زنجیر الگوریتم را تکرار می‌کنیم. تعداد نمونه‌های تولید شده تا قبل از همگرایی را به عنوان دوره سوخت کنار می‌گذاریم. پس از طی دوره سوخت بر اساس نمونه $j = ۱, \ldots, M$، $(\alpha_j, \theta_{\circ j}, \zeta_j, \lambda_j)$ تابع توزیع تجمعی

$$F_x(y_i) = \int H(y_i|\alpha_i\eta_x, \theta_{\circ i}, \zeta) dG(\alpha_i, \theta_{\circ i}).$$

به صورت زیر برآورد می‌شود:

$$\hat{F}_x(t) = \frac{۱}{M} \sum_{j=۱}^{M} H\left(t|\alpha_i\eta_x, \theta_{\circ j}, \zeta_j\right).$$

# ۶ تحلیل داده های سرطان

در این بخش به بررسی یک مجموعه از داده‌های مربوط به طول عمر بیماران مبتلا به سرطان معده در استان اردبیل پرداخته خواهد شد. اگرچه عوامل بسیاری بر بقای بیماران اثر می‌گذارند، با این حال در اینجا ما اثر عامل دو سطحی شیمی‌درمانی مدنظر قرار گرفته است که تابع بقای بیماران با استفاده از تحلیل واریانس آمیخته سلسله مراتبی برای داده‌های سانسور شده از راست براورد خواهد شد.

## ۱.۶ تحلیل بقای داده های سرطان در حضور متغیر کمکی شیمی درمانی

جدول ۱ خلاصه ای از داده‌های مربوط به بیماران سرطانی را به تفکیک شیمی درمانی و نیز سانسور نشان می‌دهد.

جدول ۱: خلاصه داده‌های سرطان

| متغیرکمکی | سطوح | رخداد | سانسور | جمع |
|---|---|---|---|---|
| شیمی درمانی | بله | ۶۹ | ۲۲ | ۹۱ |
| | خیر | ۴۲ | ۱۴ | ۵۶ |

ملاحظه می‌شود که داده‌ها شامل $n = ۱۴۷$ مشاهده طول عمر سانسور شده از راست $T_{۱}^*, \ldots, T_n^*$ هستند. مشاهدات سانسور شده به صورت $T_i^* = \min(T_i, C_i)$ تعریف می‌شوند به طوری که $C_{۱}, \ldots, C_n$ زمان‌های سانسور معین هستند. حال فرض کنید برای $i = ۱, \ldots, n$،

$$T_i|X_i = x, \theta_{ix} \sim H(T_i|\beta_i\ \eta_x, \gamma_i),$$

که در آن، $\eta_x = (۱, e_x)'$، $x$ متغیر رسته‌ای دو سطحی شیمی‌درمانی (۱ = تحت شیمی‌درمانی قرار گرفته‌اند، ۲ = تحت شیمی‌درمانی قرار نگرفته‌اند)، $\beta_i = (\beta_{۰i}, \beta_{۱i})$ و $e_x$ به صورت زیر است

$$e_x = \begin{cases} ۱ & , \quad v = ۱ \\ -۱ & , \quad v = ۲ \end{cases}$$

هسته $H$ را توزیع لگ نرمال در نظر گرفته و با توجه به سانسور از راست مشاهدات، داریم

$$f_{T_i^*}(t_i^*) = \int [h\left(T_i^*|\beta_i\eta_x, gamma_i\right)]^{\delta_i} [۱ - H\left(T_i^*|\beta_i\eta_x, \gamma_i\right)]^{۱-\delta_i} dG(\beta_i, \gamma_i|\lambda).$$

ساختار توزیع‌های پیشین در این تحلیل به صورت زیر است:

$$\begin{aligned} T_i^*|\beta_i\eta_x, \gamma_i &\sim [h(T_i^*|\beta_i\eta_x, \gamma_i)]^{\delta_i}[۱ - H(T_i^*|\beta_i\eta_x, \gamma_i)]^{۱-\delta_i} \\ (\beta_i, \gamma_i) &\sim G \quad , i = ۱, \ldots, n, \\ G(.)|\alpha_۰, m, B, S &\sim \mathrm{DP}(.|\alpha_۰ G_۰(.|m, B, S)), \\ G_۰(\beta, \gamma) &= \Gamma(\gamma^{-۱}|\tfrac{s}{۲}, \tfrac{sS}{۲})N_۲(\beta|m, B), \\ S &\sim \Gamma(۰/۵q, \tfrac{۰/۵q}{R}), \\ m &\sim N_۲(a, A), \\ B^{-۱} &\sim Wish_۳((cC)^{-۱}, c), \\ \alpha_۰ &\sim \Gamma(a_۰, b_۰), \end{aligned}$$

که در آن $q = s = ۵$، $C = A = ۰/۰۰۴l_۲$، $R = ۱$، $a_۰ = b_۰ = ۱$ و $a = (۱, ۱)$ است.

الگوریتم نمونه‌گیری $(\beta_i, \gamma_i)$ها مشابه بخش قبل است و تنها در موارد زیر اختلاف دارد:

الف. مقادیر $(\beta_i, \gamma_i)$ را برای $i = ۱, ۲, \ldots, n$، از توزیع پسین زیر نمونه‌گیری می‌کنیم؛

$$\begin{aligned} ((\beta_i, \gamma_i)|(\beta_{-i}, \gamma_{-i}), T^*) &\sim q_۰ \left[h\left(T_i^*|\beta_i\eta_x, \gamma_i\right)\right]^{\delta_i} \\ &\times \left[۱ - H\left(T_i^*|\beta_i\eta_x, \gamma_i\right)\right]^{۱-\delta_i} dG_۰\left(\beta_i, \gamma_i|m, B, S\right) \\ &+ \textstyle\sum_{j=۱, j\neq i}^{n} q_j \delta_{(\beta_j, \gamma_j)}(\beta_i, \gamma_i), \end{aligned}$$

که در آن

$$q_۰ \propto \alpha_۰ \int \left[h\left(T_i^*|\beta_i\eta_x, \gamma_i\right)\right]^{\delta_i} \left[۱ - H\left(T_i^*|\beta_i\eta_x, \gamma_i\right)\right]^{۱-\delta_i} dG_۰\left(\beta_i, \gamma_i|m, B, S\right),$$

همچنین

$$q_j \propto \left[h\left(T_i^*|\beta_i\eta_x, \gamma_i\right)\right]^{\delta_i} \left[۱ - H\left(T_i^*|\beta_i\eta_x, \gamma_i\right)\right]^{۱-\delta_i}$$

و نیز

$$q_۰ + \sum_{j\neq i} q_j = ۱$$

ب. با انجام ۷۰۰ بار الگوریتم متروپولیس-هستینگز به ایستایی می‌رسد. شکل‌های ۱ و ۲ نمودار خودهمبستگی مقادیر $\beta_i^{(j)}$ها و $\gamma_i^{(j)}$ها را برای $j = ۱, \ldots, ۷۰۰$ نشان می‌دهد.

شکل ۳ برآورد تابع توزیع تجمعی تجربی برای مشاهدات طول عمر مربوط به دو سطح متغیر کمکی را نشان می‌دهد.

همچنین شکل ۴ برآورد تابع بقا با استفاده از سه روش کاپلان-مایر، مدل کاکس و مدل تحلیل واریانس دیریخله سلسله مراتبی را نشان می‌دهد. همانطور که مشاهده می شود، براوردهای مدل کاپلان مایر و کاکس در بعضی از زمان ها مقدار ثابتی اختیار می کنند. این در حالی است که حضور داده های سانسور شده در براورد تابع بقا تحت مدل

واریانس تاثیری نمی‌گذارند.

طبق براورد تابع بقا تحت مدل تحلیل واریانس فرایند دیریخله سلسله مراتبی می‌توان دریافت که با احتمال ۴۰ درصد شخصی که تحت شیمی درمانی قرار نگرفته است تا ۲۰ مال اول بعد از شروع بیماری زنده می‌ماند. این در حالی است که با احتمال ۴۰ درصد شخصی که تحت شیمی درمانی قرارگرفته است بیش از ۲۰ ماه بعد از شروع بیماری زنده می‌ماند. بنابراین شیمی درمانی بر روی بقای بیماران سرطانی تاثیرگذار است.

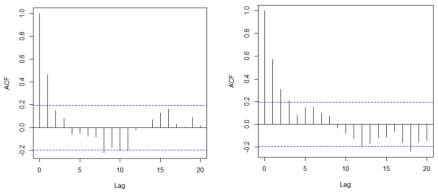
در شکل ۵ برآورد تابع چگالی احتمال، تابع توزیع تجمعی، تابع نرخ خطر و تابع نرخ خطر تجمعی براساس مدل تحلیل واریانس دیریخله سلسله مراتبی را نشان داده می‌شود.

## ۷ بحث و نتیجه‌گیری

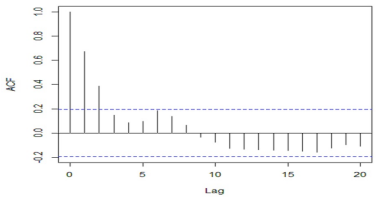
یکی از محدودیت‌های مدل‌های ناپارامتری نظیر روش برآورد کاپلان-مایر عدم حضور متغیر کمکی می‌باشد که تنها برای متغیر رسته‌ای می‌توان برای هر سطح تیمار منحنی‌های جداگانه‌ای رسم کرد. این در حالی است که مدل کاکس و تحلیل واریانس سلسله مراتبی برای تمامی مقادیر هر نوع متغیر کمکی به خوبی تابع بقا را برآورد می‌کنند.

همچنین در مدل‌هایی نظیر مدل کاکس فرضیه‌های محدودکننده‌ای برای ارتباط توابع بقا با متغیرهای کمکی وجود دارد که باعث کاهش کارایی این مدل‌ها می‌شود. مدل تحلیل واریانس دیریخله سلسله مراتبی دارای مفروضات اولیه بسیار کمتری است.

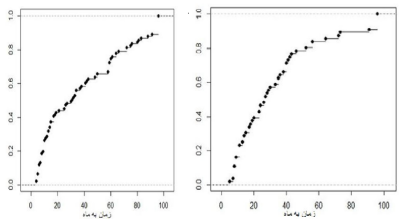
استفاده از تحلیل بیز سلسله مراتبی به طرز قابل توجهی وابستگی این روش بیزی را به توزیع‌های پیشین کاهش می‌دهد. همچنین استفاده از روش‌های عددی نمونه‌گیری گیبس و متروپولیس-هستینگز محاسبات پیچیده در این شیوه را به طرز قابل توجهی ساده می‌سازد. همچنین تحلیل نمودارهای هموار برآورد شده در خروجی بسیار ساده‌تر و مطلوب‌تر از سایر شیوه‌ها است.



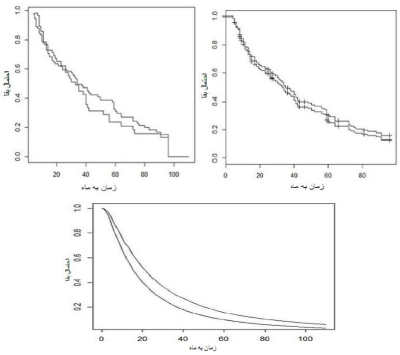

شکل ۱: نمودار خودهمبستگی مربوط به مقادیر $\beta_i^{(j)} = \left(\beta_{0i}^{(j)}, \beta_{1i}^{(j)}\right)$، $j = ۱, \ldots, ۷۰۰$.



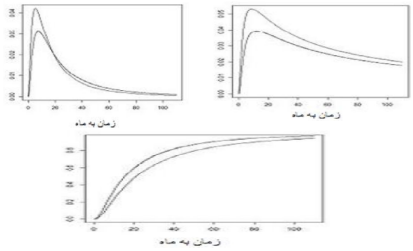

شکل ۲: نمودار خودهمبستگی مربوط به مقادیر $\gamma_i^{(j)}$، $j = ۱, \ldots, ۷۰۰$



شکل ۳: راست؛ برآورد تابع توزیع تجمعی تجربی مشاهدات طول عمر بیمارانی که تحت شیمی‌درمانی قرار نگرفته‌اند. چپ؛ برآورد تابع توزیع تجمعی تجربی مشاهدات طول عمر بیمارانی که تحت شیمی‌درمانی قرار گرفته‌اند.



شکل ۴: برآورد تابع بقا؛ سمت چپ مدل کاکس، سمت راست مدل کاپلان مایر، پایین مدل تحلیل واریانس دیریخله سلسله مراتبی



شکل ۵: برآورد توابع عمده مورد استفاده در تحلیل بقا: بالا راست؛ تابع چگالی احتمال، بالا چپ؛ تابع نرخ خطر، پایین ؛ تابع توزیع تجمعی.

# مراجع


[۱] عاطفه جاويدي، سميه راه پيما، مجيد جعفري خالدی (۱۳۹۲). معرفي پيشين فرايند ديريكله در چارچوب مدلهاي بيزي ناپارامتری. انديشه آماری، سال هجدهم، شماره ۲، ۷۲-۶۱.

[2] Antoniak, C. E. (1974). Mixture of Dirichlet Process with applications to Bayesian nonparametric problems. Annals of Statistics. 2, 1152-1174.

[3] Blackwell, D. and MacQueen, J. B. (1973). Ferguson Distribution via Polya Urn Schemes. Annals of Statistics. 1, 353-355.

[4] Christensen, R. and Johnson, W. (1988). Modeling Accelerated Failure Time with a Dirichlet Process. Biometrika. 75, 693-704.

[5] Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixture. Journal of the American Statistical Association. 90, 577-588.

[6] Escobar, M. D. and West, M. (1998). Computing Nonparametric Hierarchical Models. Springer Verlag. New York. USA.

[7] De Iorio, M., Johnson, W. O., Muller, P. and Rosner, G. L. (2009). Bayesian Nonparametric Nonproportional Hazard Survival Modeling. Biometrics. 65, 762- 771.

[8] Doss, H. and Huffer, F. W. (1999). Monte Carlo Method for Bayesian Analysis Survival Data using of Dirichlet Process Prior. Journal of Computational and Graphical Statistics. 282-307.

[9] Escobar, M. D. (1988). Estimating the means of several normal populations by estimating the distribution the means. Ph. D. thesis. Yale University. New Haven. USA.

[10] Escobar, M. D. and West, M. (1998). Computing Nonparametric Hierarchical Models. Springer Verlag. New York. USA.

[11] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. Annals of Statistics. 1, 209-230.

[12] Gelfand, A. E. and Kottas, A. (2002). Bayesian semiparametric regression for median residual life. Scandinavia Journal of Statistics. 30, 651-665.

[13] Hanson, T. and Johnson, W. O. (2002). Modeling regression error with a mixture of poly trees. Journal of the American Statistical Association. 97, 1020-1033.

[14] Hanson, T., Johnson, W. O. and Laud, P. (2008). A unified approach to semiparametric inference for survival models with step process covariates. Canadian Journal of Statistics. 37, 60-79.

[15] Kalbfleisch, J. (1978). Nonparametric Bayesian analysis of survival time data. Journal of the Royal Statistical Society. 40, 214-241.

[16] Kuo, L. and Mallick, B. (1997). Bayesian semiparametric inference for the accelerated failure- time model. Canadian Journal of Statistics. 25, 457-472.

[17] MacEachern, S. N. (1999). Dependent nonparametric processes. American Statistical Association. Virginia. USA.

[18] Walker, S. and Mallick, B. K. (1999). A Bayesian semiparametric accelerated failure time model. Biometrics. 55, 477-483.

[19] West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. ISDS Discussion Paper # 92-A 03. Duke University. USA.

[20] West, M., Muller, P. and Escobar, M. D. (1994). Hierarchical Priors and Mixture Models, with Applications in Regression and Density Estimation. John Wiley and Sons. London. UK.

[21] Zhou, M. (2004). Nonparametric Bayes estimator of survival function for doubly/ interval Censored data. Statistica Sinica. 14, 533-546.