تحلیل بقاء

نرگس سهرابی

1390

فهرست مطالب

فهرست جداول

فهرست اشکال

1 فصل اول: مفاهیم اولیه

1-1 مقدمه

1-2 معرفی داده های بقاء[1]

آنالیز بقاء[2] یا تحلیل ماندگاری یا تجزیه و تحلیل بقاء یکی از مباحث علم آمار است که در رشته های مختلفی از جمله؛ علوم کامپیوتر، اپیدومیولوژی[3] و کشاورزی کاربرد دارد (اسفندیاری،1389). تحلیل بقاء مجموعه ای از تکنیک های آماری متنوّع جهت تحلیل متغیّرهای تصادفی است که دارای مقادیر نامنفی می باشند. مقدار این متغیّرتصادفی، زمان شکست یک مؤلّفه فیزیکی و یا زمان مرگ یک واحد بیولوژیک می باشد.

در تحلیل داده های بقاء مسئله اصلی یافتن مدل مناسبی برای همبستگی زمان بقاء با متغیرهای عوامل مختلف می باشد. اگرداده ها دارای مقادیر ناتمام نباشد، می توان این ارتباط را به وسیله رگرسیون چندگانه بیان نمود. البته به علّت وجود چولگی بایستی در انجام این روش از تبدیل لگاریتم و یا تکنیک معکوس تابع و بسط آن استفاده نماید. ولی اگر داده های ناتمام داشته باشیم، استفاده از آنالیز رگرسیون به دلیل نداشتن مقدار دقیق متغیّر وابسته غیرممکن خواهد بود. در ساختن مدل بقاء[4]، می توان تابع مخاطره[5] را برای هرفرد (به عنوان مثال، بیمار) به عنوان یک تابع از متغیّرها با زمان ثابت در نظرگرفت؛ از آنجاییکه ممکن است در طول مطالعه همه متغیّرها یا بعضی از آنها با زمان تغییر کند، مدل می تواند با استفاده از متغیّرهای وابسته به زمان ، اطّلاعات مفیدی را درباره چگونگی تأثیر تاریخچه بیماران روی بقاء در اختیار بگذارد. اما در حقیقت تصویر کاملی از وضعیّت بیمار در آینده را نمی دهد و نمی توان آنالیز صحیح جهت پیشگویی در وضعیّت های خاص بیمار را ارائه نماید. برای رفع مشکل، معمولا ازمدل های چند حالته استفاده می کنند. اخیرا در مسائل اپیدمیولوژی نیز از این مدل چند حالته برای بررسی های بقاء در جامعه بخصوص در مورد افراد سرطانی اقداماتی صورت داده اند که متأسّفانه در مدل های ارائه شده بعضی نکات اساسی در نظر گرفته نشده است.

تحلیل بقاء از نظرعلم آمار عبارت است از: استفاده از فنون مختلف آماری در تحلیل متغیّرهای تصادفی نامنفی. نوعا مقدار این متغیّرتصادفی زمان شکست یک مؤلّفه فیزیکی (مکانیکی یا الکتریکی) یا زمان مرگ یک واحد زنده (سلّول، بیمار، حیوان وغیره) است. ممکن است این متغیّر، زمان یادگیری یک مهارت باشد، یا حتی امکان دارد به زمان هیچ ارتباطی نداشته باشد. برای مثال متغیّر می تواند مبلغ پرداختی

---

[1] Survival data
[2] Survival analysis
[3] Epidemiology
[4] Survival model
[5] Hazard function

یک شرکت بیمه در وضعیّت خاصّی باشد در برخی موارد، یک بیمار بهبود یافته و مبلغ کل پرداختی بیمه او معلوم است. در موارد دیگر، بیماری هنوز ادامه دارد و تنها مبلغ پرداختی تا آن زمان معلوم است. پیشینه و منشأ تحلیل بقاء کارهایی است که در گذشته در مورد جداول طول عمر انجام شده است. شکل جدید تحلیل بقاء از نیم قرن گذشته با کاربردهای مهندسی مورد توجه و مطالعه قرارگرفته است (لیوایت و اولشن، 1974).

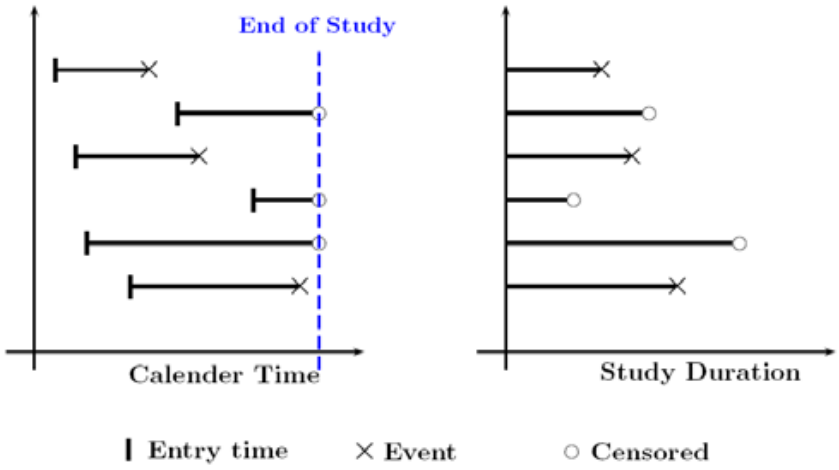
هنگامي كه با تحليل داده هاي بقاء سروكار داريم، دو هدف عمده مورد توجه است. يكي مدل بندي براي پيدا كردن تركيب مناسبي از متغيرهاي كمكي كه طول بقاي اعضاي مورد مطالعه را تحت تأثير قرار می دهند و هدف بعدي پيدا كردن برآوردهاي مطمئن براي تابع مخاطره در زيرگروه هاي مورد بررسي مي باشد. به طور معمول دو راه كار اساسي در تحليل داده هاي آماري، روش هاي نيمه پارامتري و پارامتري هستند. هنگامی که در تحلیل بقاء با این دیدگاه به بررسی مسئله می پردازیم، با توجه به هر کدام از اهداف فوق یکي از راه كارهاي ذكر شده از اهميت بيشتري برخوردار خواهد بود (كلين و موئشبرگر، 1997). امروزه به دلیل استفاده روز افزون از تحلیل بقاء در مطالعات پزشكي نياز به مدل هاي كارا و با انعطاف بيشتر براي داده هاي بقاء بيش از پيش احساس مي شود .

مطالعه اي را در نظر بگيريد كه هدف از آن بررسي اثر داروي متيل پردنيزولون در كاهش خطر رد پيوند كليه باشد. فرض كنيد مطالعه به اين صورت انجام شود كه از ابتداي سال 1385 تا انتهاي سال 1386 تمام بيماراني كه در بيمارستان سينا تحت پيوند كليه قرار مي گيرند، به طور تصادفي به دو گروه تقسيم شوند: گروهي كه متيل پردنيزولون دريافت مي كنند و گروهي كه دارونما دريافت مي كنند. تمام بيماران تا پايان سال 1386 پيگيري مي شوند و تمام موارد رد پيوند ثبت مي شوند. اگر قرار باشد شما داده هاي حاصل از اين مطالعه را آناليز كنيد، از كدام آزمون آماري استفاده مي كنيد؟ (بديهي است كه پاسخ درست يكي از تست هاي آناليز بقاء است! اما چرا؟)

اجازه دهيد مطالعه اي ديگر را مثال بزنيم كه با هدف مشابهي انجام شده، اما به جاي آن كه بيماران تا پايان سال 1386 پيگيري شوند تا پايان عمرشان پيگيري شده اند. در چنين حالتي آناليز كردن داده ها آسان است. شما در مورد هر بيمار دقيقاً مي دانيد كه آيا تا پايان عمرش دچار رد پيوند شده يا نشده و مي توانيد درصد وقوع رد پيوند در هر گروه را گزارش و مقايسه كنيد. حتي مي توانيد به دقت تعيين كنيد كه فاصله زماني بين پيوند و وقوع رد پيوند، در هر يك از دو گروه مطالعه، به طور متوسط چه قدر بوده است. اما مشكلي كه درمورد اين مطالعه وجود دارد آن است كه عملاً اجراي آن بسيار دشوار است. (مثلاً به اين دليل كه همه بيماران بايد تا پايان عمر پيگيري شوند، ممكن است مطالعه 30 سال يا بيشتر ادامه پيدا كند!)

مطالعه اول از لحاظ اجرايي منطقي تر است، اما در پايان اين مطالعه همه داده هايي كه در مطالعه دوم در اختيار پژوهشگر بود، حاصل نشده اند. مثلاً چون بيماران فقط تا پايان سال 1386 پيگيري شده اند، اگر

بيماري در فروردين 1385 پيوند دريافت كند و تا پايان مطالعه آن را پس نزند، مي توان گفت به مدت دو سال رد پيوند در وي رخ نداده، اما بيماري كه در اسفند 1386 وارد مطالعه مي شود، فقط يك ماه پيگيري مي شود، و ما نمي توانيم بگوييم كه او هم تا دو سال دچار رد پيوند نخواهد شد؛ چه بسا او در فروردين 1387دچار رد پيوند شود!



شكل 1-1. پيگيری بيماران در محور زمان

شكل 1-1، وضعيت پيگيري چند بيمار در محور زمان را نشان مي دهد. در سمت چپ، محور افقي نشان دهنده تاريخ (بر اساس تقويم) است. همان طور كه مي بينيد، زمان ورود بيماران با مطالعه متفاوت است. علامت ضربدر نشان مي دهد كه بيمار دچار رد پيوند شده است. علامت دايره نشان مي دهد كه بيمار تا پايان مطالعه پيگيري شده ولي دچار رد پيوند نشده است. در سمت راست، مي بينيد كه اطلاعات همان بيماران دوباره نشان داده شده است، با اين تفاوت كه محور افقي نشان دهنده زمان « از ورود به مطالعه » است. در واقع اطلاعاتي كه پژوهشگر در پايان مطالعه دارد از همين جنس است.

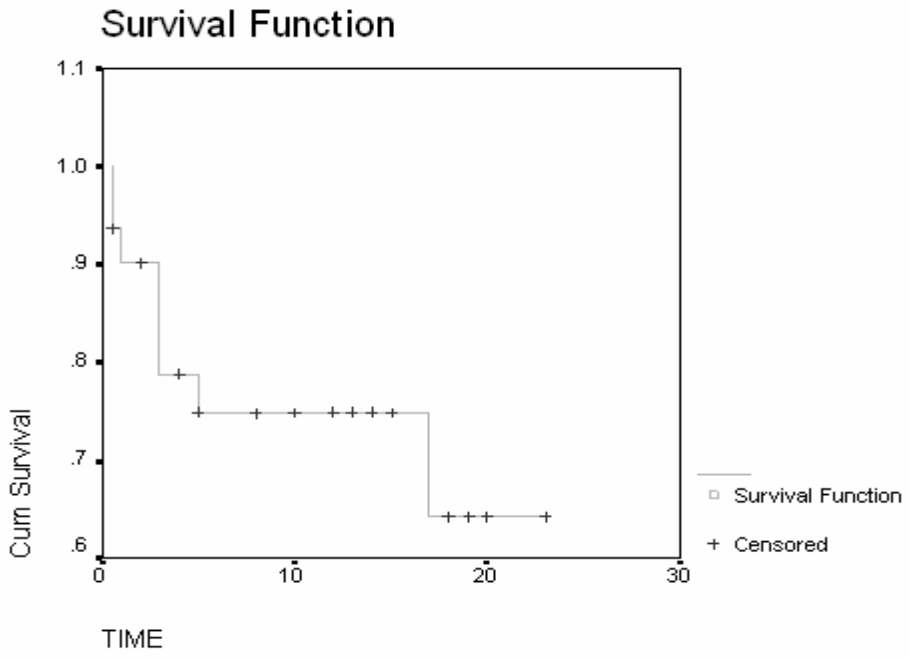
وقتي بيماري تا پايان مطالعه پيگيري شود و پيامد مورد نظر ما (مثل رد پيوند) در وي رخ ندهد، در مطالعات آناليز بقا، به اين داده اصطلاحا « سانسور شده »[1] گفته مي شود. هدف مطالعه آناليز بقاء آن است كه با وجود داده هاي سانسور شده، بتواند تخميني از شاخص هاي مرتبط با زمان را به دست بياورد؛ مثلاً در مطالعه اول، هدف آن است كه بتوانيم تخميني از متوسط عمر پيوند در بيماران، يا ريسك پس زدن پيوند بعد از دو سال به دست بياوريم.

جالب است بدانيد كه اولين تلاش های اين چنين، توسط ستاره شناس معروف، « ادموند هالی »، صورت گرفت. وی در دوره ای اززندگي اش به ثبت سن مرگ همشهري هاي خود پرداخت و جدولي از اين اطلاعات فراهم كرد كه نشان مي داد چند درصد از آدم ها در هر دوره ی سنی می ميرند. امروزه به چنين

[1] Censored

جدولی « جدول طول عمر »[1] گفته مي شود. از آن جا كه در اولين تلاش هاي اين چنيني، پيامدي كه بررسي مي شد مردن يا زنده ماندن افراد بود، به روش تحليل اين داده ها آناليز بقاء نام دادند.

براي نشان دادن بقاي افرادي كه در مطالعه وارد شده بودند، معمولا از نمودارهای پلکانی[2] استفاده مي شود، اگرچه مي توان ازانواع ديگر نمودارها نيز استفاده كرد. آن چه در شكل 2-1 مي بينيد، يك نمودار پلكاني متداول است. محور افقي نشان دهنده زمان از ورود به مطالعه است. محور عمودي نشان دهنده بقاء است، اما بر خلاف نمودارهاي ساده، اين محور به صورت « تجمعی »[3] ترسيم شده است. با نگاه كردن به اين نمودار مي توان فهميد كه، مثلاً پس از گذشت 10 ماه از انجام پيوند، حدود 75٪ بيماران زنده مانده اند و 25٪ ديگر مرده اند.



شكل 2-1. نمودار پلكانی بقای بيماران

از مهم ترين نتايجي كه از چنين مطالعاتي حاصل مي شود، تعيين بقای بيماران است. احتمالا جملاتی نظير اينكه « بقای 5 ساله بيماران 40٪ است» را شنيده ايد. در نمودار صفحه قبل مي بينيد كه بقاي 2 ساله (24ماهه) بيماران در مطالعه حدود 65٪ است. با كمك روش هاي آناليز بقاء، نظير محاسبه حد حاصل ضرب كاپلان- مير[4]، می توان از مقدار بقای واقعی بيماران، بر اساس آن چه در يك مطالعه مشاهده شده، تخميني به دست آورد. به طور خلاصه، اين روش امكان محاسبه بازه اطمينان 95٪ براي شاخص بقاء را فراهم مي كند.

[1] Lifetime Table
[2] Step plot
[3] Cumulative
[4] Kaplan-Meier Product-Limit

براي مقايسه بقاي بيماران در دو گروه نيز، روش هاي آماري اختصاصي وجود دارد كه آزمون گهان[1] يا برسلو[2] و آزمون رتبه لگاريتمی[3]، از آن جمله اند. اين روش ها پژوهشگر را قادر مي سازد كه وجود تفاوت معني دار بين دو گروه را با ذكر مقدار P-value گزارش كند (داوسون و تراپ، 2001).

1-3 نمونه هایی از داده های بقاء

معمولا نوع داده هایی که در اینجا، مورد تجزیه و تحلیل قرار می گیرند داده های خاصی هستند که به داده های بقاء موسوم اند. داده های مربوط به زمان وقوع حوادثی چون مرگ پس از تشخیص بیماری، عود بیماری پس از بهبود، طلاق پس از ازدواج، روی آوری مجدد به اعتیاد پس از ترک آن و ... ، به روش های آماری ویژه ای برای پیش بینی، برآورد و تجزیه و تحلیل داده ها نیاز دارد. برای نمونه، مثال های زیر را در نظر بگیرید.

مثال1: داده های مربوط به بیماران قلبی عروقی: در طی دهه های گذشته استفاده از آنالیز بقاء در مباحث پزشکی نظیر سرطان و بیماری های قلبی عروقی رواج بسیاری یافته است. یکی از علل عمده مرگ و ناتوانی در جهان، بیماری های قلبی عروقی می باشند. مطالعات انجام شده یک مطالعه آینده نگر تاریخی است و بر روی بیمارانی که از نیمه دوم سال 1383 لغایت نیمه اول 1385 در مرکز درمانی فاطمه الزهرا (س) ساری (مرکز قلب مازندران) تحت آنژیوپلاستی قرارگرفته بودند، انجام شد. مطالعات انجام شده در این زمینه بیشتر معطوف به شناسایی عوامل خطر برای این گروه از بیماران بوده است. درحالی که این مطالعه با هدف پیش بینی زمان بروز مجدد عوارض بالینی آنژیوپلاستی به همراه میزان تأثیر متغیرهای مرتبط با این عامل بر اساس آنالیز داده های بقاء انجام شده است (یوسف نژاد و همکاران، 1390).

مثال2: داده های مربوط به بیماران مبتلا به سرطان کوروکتال: سرطان کوروکتال یک بیماری کشنده و نسبتا شایع (5000 مورد جدید در سال در ایران) است. میزان بروز استاندارد شده سنی سرطان کوروکتال در ایران در مردان در سال 1382 ،1383 و 1384 ، به ترتیب 9/27 ،9/64 و9/90 در هر100000 مرد ایرانی و 9/12، 9/47 و 9/13 در هر 100000 زن ایرانی بوده است. بر طبق همین مطالعات، سالیانه 5000 نفر در ایران به سرطان کوروکتال مبتلا می شوند. انتظار می رود شیوع سرطان کوروکتال به علت بروز بالا و رو به افزایش آن و همچنین افزایش نسبت بقاء، یک معظل عمده مدیریت سرطان در ایران باشد. با توجه به اینکه طول عمر برخی از بیماران سرطانی به 5 سال نمی رسد و در بعضی از انواع آن ها، طول عمر، کوتاهتر است و از طرف دیگر نیازهای بیماران سرطانی در سال های اول، دوم، سوم، چهارم و پنجم پس از تشخیص، با هم متفاوت است، معمولا شیوع1، 3-2 و 5-4 ساله به منظور ارزیابی درمان اولیه (1ساله)، پیگیری کلینیکی (3-2 ساله) و زمان بهبودی یا توانبخشی (5-4 ساله) استفاده می شود. برای

[1] Gehan test
[2] Breslow
[3] Log-rank test

محاسبه موارد شیوع در سال های متفاوت پس از تشخیص سرطان، پیسانی و همکارانش از داده های بقاء و میزان های بروز اختصاصی سنی استفاده کرده و شیوع 1، 3-2 و 5-4 ساله سرطان ها را در 25 منطقه جهان برآورد کرده اند. نظر به اینکه در ایران برای اغلب سرطان ها، طول عمر 5-1 ساله موجود نیست، با وجود میزان بروز، میزان شیوع به درستی مشخص نیست. با توجه به اهمیت خاص شیوع در مدیریت سرطان، میزان شیوع سرطان ها که معرف موارد جدید به اضافه افرادی است که با سرطان از سالهای قبل زندگی می کنند، نشان می دهد که چه نیازهایی وجود داشته و سیستم ارائه خدمات بایدچه امکاناتی را برای پاسخگویی به این نیازها در اختیار داشته باشد. به عبارت بهتر، شیوع سرطان، بیانگر تعداد افرادی در جامعه است که در یک زمان مشخص نیاز به دریافت خدمات تعریف شده ای دارند. بنابراین در این مطالعه شیوع 1، 3-2 و 5-4 ساله سرطان کولون براساس داده های بقاء تعیین می شود تا بتوان بیماران نیازمند به درمان اولیه، پیگیری کلینیکی و زمان بهبودی و همچنین توانبخشی و حمایت های تسکینی را مشخص نمود (پیسانی و همکاران، 2002).

مثال3: داده های مربوط به بیماران مبتلا به سرطان معده: در دهه هاي گذشته مطالعات بسياري در زمينه تحليل بقاء در سرطان معده پايه ريزي شده اند. داشتن اطلاعات به روز در زمينه عوامل مؤثر بر بقاء سرطان به عنوان يكي از اركان مهم جهت پزشكان، انكولوژيست ها، اپيدميولوژيست ها و تمامي حرفي كه درگير مسائل باليني پژوهش مي باشند، ضروري است. داده هاي اين پژوهش با استفاده از يك مطالعه كوهورت تاريخي بر روي تمامي بيماران مبتلا به سرطان معده ثبت شده در مركز ثبت تومور استان فارس در حد فاصل سال های 1380 تا 1384به دست آمد. بيماران در طول دوره مطالعه به طور دوره اي مورد پيگيري تلفني واقع شده و وضعيت بقاء آنها تا پايان سال به عنوان زمان شكست مشخص گرديد.

در اين مطالعه با استفاده از فرم هاي ثبت سرطان، اطلاعات مربوط به متغيرهايي از قبيل؛ سن تشخيص بيماري، جنسيت، گروه قومي (فارس، لر، كرد، ترك، ساير)، وضعيت تأهل، شغل مردان (كارمند و بازنشسته، كشاورز و دامدار، كارگر ساده و فنی، ساير)، شغل زنان (شاغل، خانه دار)، سابقة فاميلي ابتلا به يك نوع سرطان در بستگان درجة اول و درجة 2و3 ، سابقة ابتلا به بيماريهاي گوارشي (بيماري هاي التهابي معده بر طبق اطلاعات فرمهاي ثبت سرطان بيماران)، شاخص توده بدنی (BMI) مصرف دخانيات (مصرف كننده فعال، مصرف كننده غير فعا ل، هرگز مصرف نكرده)، نوع اولين درمان (عمل جراحي، ساير)، درجة تمايز يافتگي تومور، متاستاز به ساير ارگا ن ها و فاصلة زماني بين اولين علامت بيماري تا زمان تشخيص (كمتريا بيشتر از يك ماه)، جمع آوري شد (رجائی فرد و همکاران، 1388).

1-4 تابع بقاء[1]

در تحلیل بقاء برای مشخص کردن توزیع زمان های بقاء، معمولا از دو تابع استفاده می شود. یکی از

[1] Survival function

این توابع، تابع بقاء می باشد. هدف اولیه در تجزیه و تحلیل داده های بقاء شناخت تابع بقاء است که با علامت S نشان داده می شود.

فرض کنید متغیر تصادفی T $(T > 0)$ دارای تابع چگالی احتمال $f(t)$ و تابع توزیع $F(t)$ [1]باشد. تابع بقاء $S(t)$ به صورت زیر تعریف می شود:

$$S(t) = \Pr(T > t) = \int_t^{\infty} f(u)\, du = 1 - F(t)$$

که در آن t مدت زمان است وT ، دلالت بر زمان مرگ یا شکست دارد و Pr ، بیانگر این احتمال است که عنصر مورد نظر حداقل تا زمان T ماندگاری داشته باشد.

تابع چگالی بقاء را می توان به صورت زیر تعریف کرد:

$$s(t) = \acute{S}(t) = \frac{d}{dt}S(t) = \frac{d}{dt}\int_t^{\infty} f(u)\, du = \frac{d}{dt}(1 - F(t)) = -f(t)$$

هم چنین، تابع بقاء در مسائل بقای بیولوژیکی، تابع بازمانده[2] یا تابع بازماندگی، و در مسائل بقای مکانیکی، تابع قابلیت اطمینان[3] نامیده می شود. در مورد دوم تابع قابلیت اطمینان را با $R(t)$ نشان می دهند.

معمولا فرض می شود، $S(0) = 1$ باشد. اگرچه می تواند کمتر از یک نیز باشد و این در صورتی خواهد بود که احتمال شکست یا مرگ فوری وجود داشته باشد. تابع بقاء باید کاهشی باشد. به این معنی که :

$$S(u) \leq S(t) \quad ; \; u \leq t$$

این ویژگی به طور مستقیم از $F(t) = 1 - S(t)$ پیروی می کند. که انتگرال یک تابع غیر منفی است و نشان دهنده این مفهوم است که بقای طولانی در صورتی امکان پذیر است که تمام سنین جوانی به موفقیت سپری شده باشد. باید به این نکته توجه شود که تابع بقاء معمولا با افزایش سن بدون محدودیت به صفر میل می کند. یعنی: $S(t) \to 0 \; ; t \to \infty$.

1-5 نمودار بقاء

نمودار تابع بقاء، در مقابل زمان بقاء، **منحنی بقاء**[4] نامیده می شود. این منحنی که از روی داده های مشاهده شده برآورد می شود، در تحلیل داده های زیستی از اهمیت زیادی برخوردار است. منحنی بقاء (log – log) یک منحنی برآوردی است که در نتیجه دو بار لگاریتم گیری ازتابع بقاء برآوردی، $\hat{S}$ ،حاصل می

[1] $F(t)$ را تابع توزیع طول عمر نیز می گوییم که در بخش بعد معرفی می کنیم.
[2] Survivor function
[3] Reliability function
[4] Survival curve

شود و در واقع انتقالی از $\hat{S}$ می باشد و آن را به طور خلاصه به صورت (Ln(-Ln(s)) – ، نشان می دهند (حسینی، 1390).

1-6 توزیع طول عمر[1]

تابع توزیع طول عمر، که طبق قرارداد با علامت F نشان داده می شود، به عنوان مکملی از تابع بقاء تعریف شده است:

$$F(t) = \Pr(T \leq t) = 1 - S(t)$$

و مشتق F ، که تابع چگالی[2] توزیع طول عمر می باشد و با f نشان داده می شود، به صورت زیر است:

$$f(t) = \acute{F}(t) = \frac{d}{dt}\, F(t).$$

تابع f ، گاهی اوقات، چگالی رویداد نامیده می شود. این تابع چگالی نرخ مرگ یا شکست رویدادها در واحد زمان است. طول عمر آینده[3] در یک زمان معین t_0 ، زمان باقیمانده تا زمان مرگ، بقاء معینی تا سن t_0 است. بنابراین، $T - t_0$ در حال حاضر نماد است. طول عمر آینده مورد انتظار[4]، ارزش مورد انتظار طول عمر در آینده است. احتمال مرگ در سن $t + t_0$ و یا قبل از آن به صورت زیر است:

$$P(T \leq t + t_0 \mid T > t_0) = \frac{P(t_0 < T \leq t + t_0)}{P(T > t_0)} = \frac{F(\, t + t_0) - F(t)}{S(t_0)}.$$

بنابراین چگالی احتمال طول عمر آینده به صورت زیر خواهد بود:

$$\frac{d}{dt}\frac{F(t + t_0) - F(t_0)}{S(t_0)} = \frac{f(t + t_0)}{S(t_0)}$$

و طول عمر آینده مورد انتظار به فرم زیر می باشد:

$$\frac{1}{S(t_0)}\int_0^\infty t f(t + t_0)dt = \frac{1}{S(t_0)}\int_{t_0}^\infty S(t)dt$$

که در آن، عبارت دوم از ادغام بخش ها به دست آمده است که در زمان $t_0 = 0$ ، یعنی زمان در بدو تولد، به طول عمر مورد انتظار کاهش پیدا می کند.

---

[1] Lifetime distribution
[2] Density function
[3] Future lifetime
[4] Expected future lifetime

در مسائل قابلیت اعتماد[1]، طول عمر مورد انتظار، زمان متوسط تا شکست[2] و طول عمر آینده مورد انتظار، طول عمر باقیمانده متوسط[3] نامیده می شود. احتمال بقای یک فرد تا سن t یا بعد از آن، S(t) است. طبق تعریف تعداد مورد انتظار بازماندگان تا سن t ، خارج از جمعیت اولیه n نوزاد، برابر $n \times S(t)$ است. با این فرض که تابع بقاء همه افراد یکسان در نظر گرفته می شود. بنابراین نسبت انتظار بازماندگان S(t) است. اگر بقای افراد مختلف مستقل باشد، تعداد بازماندگان در سن t دارای توزیع دو جمله ای با پارامترهای n و S(t) خواهد بود و واریانس نسبت بازماندگان به صورت $S(t) \times {(1 - S(t))}/{n}$ می باشد.

سنی که در آن یک نسبت مشخصی از بازماندگان باقی می مانند، را می توان با حل معادله $S(t) = q$ برای t به دست آورد. که در آن q مخفف کلمه question ، به معنی چارک ، در مسئله می باشد. نوعا در طول عمر میانه[4]، $q = \frac{1}{2}$ مورد علاقه می باشد یا چارک های دیگری مثل: $q = 0.90$ یا $q = 0.99$.

هم چنین می توان استنتاج های پیچیده ای را از توزیع بقاء ایجاد کرد. در مسائل قابلیت اعتماد مکانیکی، می توان هزینه (یا به طور کلی، ابزار) را در نظر گرفت و در نتیجه مسائل مربوط به تعمیر و یا جیگزینی را حل نمود. این امر منجر به مطالعه نظریه نوسازی و تئوری قابلیت اطمینان از پیری و طول عمر می شود.

1-7 تابع خطر[5]

اینک به معرفی تابع خطر (مخاطره) یا تابع نرخ شکست می پردازیم:

تابع خطر با علامت λ نشان داده می شود و به عنوان نرخ رویداد[6] در زمان t ، مشروط به بقاء تا زمان t و یا بعد از آن تعریف می شود.

تابع بقای $S(t) = 1 - F(t)$ را در نظر بگیرید. $F(t)$ تابع چگالی احتمال تجمعی می باشد. تابع چگالی احتمال $f(t)$:

$$f(t) = \lim_{\Delta t} \frac{pr\,[\,t < T < t + \Delta t]}{\Delta t} = -\frac{d\,S(t)}{dt},$$

بیانگر احتمالی است که شکست در فاصله زمانی t و Δt اتفاق بیافتد.

---

[1] Reliability
[2] Mean time to failure
[3] Mean residual lifetime
[4] Median lifetime
[5] Hazard function
[6] Event rate

تابع مخاطره به شکل زیر تعریف می شود:

$$\lambda(\mathrm{t}) = \frac{\mathrm{f(t)}}{1-\mathrm{F(t)}}$$

که ،

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{pr\,[t \le T < t + \Delta t \mid T \ge t]}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{d\ \log S(t)}{dt},$$

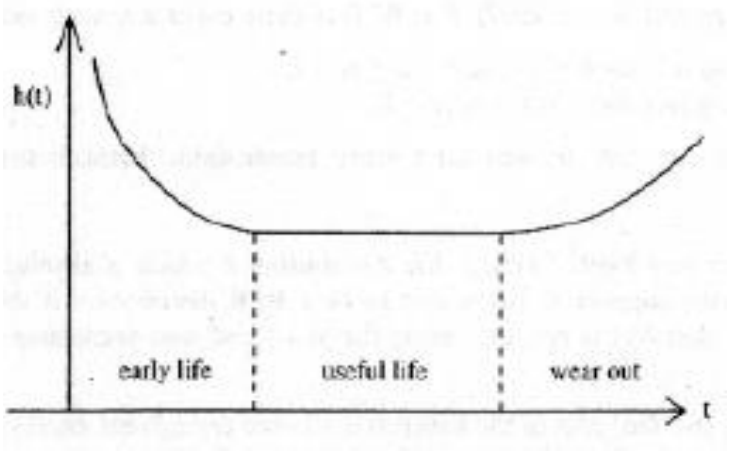
بیانگر احتمال شرطی است که عنصر مورد نظر در فاصله زمانی $[t, t + \Delta t)$ با فرض اینکه تا زمان t ماندگاری داشته باشد، دچار شکست شود. تمام روابط بالا به هم وابسته هستند.

در علم امراض مسری، تابع $\lambda(t)$ را نرخ مرگ و میر نامند. نیروی مرگ و میر مترادف تابع خطر است که در جمعیت شناسی و علوم آماری کاربرد ویژه ای دارد و آن را با علامت μ نشان می دهند. نرخ شکست[1] اصطلاح مترادف دیگری است. نرخ شکست را می توان به سه دسته زیر تقسیم کرد:

- نرخ شکست یکنواخت که تابع نرخ مخاطره یا صعودی و یا نزولی است.
- نرخ شکست U شکل که تابع نرخ مخاطره در این حالت به شکل وان حمامی است. یعنی ابتدا کاهشی و سپس در یک بازه زمانی ثابت و پس از آن روند افزایشی دارد.
- نرخ شکست U شکل تعمیم یافته که تابع نرخ مخاطره در اینجا به فرم یک چند جمله ای است و شکل آن حالت پیچ و خم دارد (علی حسینی).

تابع خطر باید مثبت باشد. یعنی $\lambda(t) \ge 0$. انتگرال آن بر بازه $[0, \infty]$ باید بینهایت باشد. اما در هر بازه ای غیر از این بازه، حتما لازم نیست که بینهایت شود و ممکن است افزایشی یا کاهشی، غیر یکنواخت و یا ناپیوسته باشد. به عنوان مثال تابع خطر وان حمامی را در نظر بگیرید که در شکل 3-1 نشان داده شده است.

[1] Hazard rate



شکل 3-1 . نرخ مخاطره U شکل

تابع مخاطره به صورت زیر قابل تفسیر است:

$$\lambda(t)\,dt \cong pr(t \le T < t + dt | T \ge t)$$

= pr { رخ دهد (t , t + dt) پیشامد در بازه | { بقا بیشتر از t

از $\lambda(t)$ انتگرال می گیریم، داریم:

$$\int_0^t \lambda(u)\,du = \int_0^t \frac{f(u)}{1-F(u)}\,du = -\log[1-F(u)]_0^t$$

$$= -\log[1-F(t)] = -\log S(t)$$

درنتیجه می توان رابطه مهم زیر را نتیجه گرفت:

$$S(t) = exp \int_0^t \lambda(u)du$$

توجه نمایید که $F(\infty) = 1$ (یعنی$\acute{S}(\infty) = 0$) خواهد بود، اگر و فقط اگر $\int_0^\infty \lambda(u)du = \infty$ باشد. مفاهیم بالا را می توان برای حالتی که T دارای تابع چگالی نیست نیز تعمیم داد. به عبارت دیگر، هنگامی که تابع توزیع F گسسته باشد (میلر، 2001).

تابع خطر می تواند در شرایط استفاده از تابع خطر تجمعی[1] نیز قرار بگیرد که با Λ نشان داده می شود و به صورت زیر تعریف می کنیم:

[1] Cumulative hazard function

$$\Lambda(t) = -\log S(t)$$

بنابراین با جا به جا کردن روابط و تبدیل کردن به حالت نمایی داریم:

$$S(t) = exp\big(-\Lambda(t)\big)$$

با مشتق گرفتن (با قاعده زنجیری) داریم:

$$\frac{d}{dt}\Lambda(t) = -\frac{\acute{S}(t)}{S(t)} = \lambda(t)$$

نام «خطر تجمعی» مشتق شده است. قطع نظر از اینکه

$$\Lambda(t) = \int_0^t \lambda(u)\, du$$

که انباشت خطر در طول زمان را نشان می دهد. بنا بر تعریف $\Lambda(t)$ می بینیم که، $\Lambda(t)$ با میل کردن t به سمت بینهایت (فرض کنید $S(t)$ به صفر میل کند) افزایش می یابد. این بدان معنی است که $\lambda(t)$ نباید خیلی سریع کاهش یابد. چون طبق تعریف، خطر تجمعی باید دراین حالت باید خاصیت واگرایی داشته باشد. به عنوان مثال $exp(-t)$ تابع خطر هیچ توزیع بقایی نیست. زیرا انتگرال آن به عدد یک همگرا است.

1-8 سانسور(برش)[1] و انواع آن

ویژگي اصلي داده هاي بقاء در مقایسه با دیگر داده هاي آماري، وجود داده هاي سانسور مي باشد. آنچه تحلیل بقاء را از سایر مباحث آماری جدا می کند، عمل برش است. به طور کلی، یک مشاهده بریده شده، فقط شامل قسمتی از اطلاعات مربوط به متغیرهای تصادفی مورد نظر است.

در اینجا سه نوع برش را معرفی می کنیم. فرض کنید T_1 ، T_2 ، ... ، T_n ، متغیرهای مستقل و هم توزیع(iid) با تابع توزیع F باشند. برش ها را به شرح زیر بررسی می کنیم:

1-8-1 برش نوع اول

فرض کنید t_c عددی ثابت باشد، که آن را زمان برش می نامیم. به جای مشاهده T_1 تا T_n (متغیرهای مورد توجه)، فقط متغیرهای Y_1 تا Y_n را به شرح زیر مشاهده می کنیم:

$$Y_i = \begin{cases} T_i & T_i \le t_c \\ t_c & T_i > t_c \end{cases}$$

[1] censoring

توجه کنید که تابع توزیع Y در y =t_c ، دارای جرم مثبت $P(T > t_c) > 0$، است.

1-8-2 برش نوع دوم

فرض کنید r < n عددی ثابت و $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$ آماره های مرتب T_1 تا T_n باشند. اگر r امین شکست رخ دهد، مشاهده در این حالت پایان می پذیرد. بنابراین، فقط می توانیم $T_{(1)}$ تا $T_{(n)}$ را مشاهده کنیم. مشاهدات مرتب شده کامل به شرح زیراند:

$$
\begin{aligned}
Y_{(1)} &= T_{(1)} \\
Y_{(2)} &= T_{(2)} \\
&\vdots \qquad \vdots \\
Y_{(r)} &= T_{(r)} \\
Y_{(r+1)} &= T_{(r)} \\
&\vdots \qquad \vdots \\
Y_{(n)} &= T_{(r)}
\end{aligned}
$$

هر دو نوع، برش اول و دوم را در مهندسی به کار می برند. برای مثال، تعدادی لامپ یا ترانزیستور موجود است. تمام آنها را در زمان t=0 مورد آزمایش قرار می دهیم و زمان شکست را یادداشت می کنیم. ممکن است طول عمر بعضی از ترانزیستورها طولانی باشد و نخواهیم به مدت طولانی منتظر بمانیم تا آزمایش به پایان برسد. بنابراین، امکان دارد، آزمایش را در زمان t_c ، که از قبل تعیین شده است، پایان دهیم. در این صورت برش نوع اول انجام شده است. درحالت دیگر از قبل ندانیم چه زمانی برای برش مناسب است. لذا، تصمیم بگیریم، که در انتظار بمانیم تا نسبت معینی از ترانزیستورها، مانند: $\frac{r}{n}$ بسوزند. در اینجا، برش نوع دوم رخ داده است.

1-8-3 برش نوع سوم (برش تصادفی)

این نوع برش که به برش تصادفی نیز معروف است، به صورت زیر تعریف می شود:

فرض کنید، C_1 تا C_n ، متغیرهای مستقل و هم توزیع (iid) با تابع توزیع G باشند. C_i زمان برش مربوط به T_i است. در این حالت فقط می توانیم (Y_1, δ_1) ، ... ، (Y_n, δ_n) را مشاهده کنیم.

که در آن Y_i و δ_i به شرح زیر تعریف می شوند:

$$Y_i = \min(T_i , C_i) = T_i \wedge C_i$$

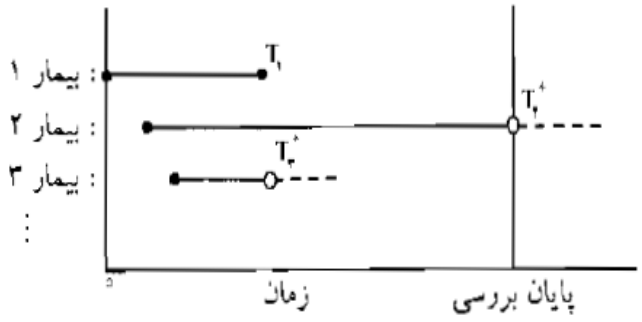
و

$$\delta_i = \mathrm{I}(T_i \leq C_i) = \begin{cases} 1 & T_i \leq C_i \\ 0 & T_i > t_c \end{cases} \qquad \begin{matrix} (\text{یعنی } T_i \text{ بریده نمی‌شود}) \\ (\text{یعنی } T_i \text{ بریده می‌شود}) \end{matrix}$$

توجه شود که Y_1 تا Y_n ، متغیرهای (iid) با توزیع H هستند. علاوه براین، δ_1 تا δ_n ، شامل اطلاعات مربوط به برش هستند. تذکر این که، در برش نوع اول و دوم می توانستیم ببینیم که چه مشاهداتی بریده می شوند. به این علت، تعریف δ_iها به طور آشکار ضروری نبود.

برش تصادفی در کارهای پزشکی و در مورد افراد یا آزمایش های بالینی استفاده می شود. در یک آزمایش بالینی، ممکن است بیماران برای معالجه در زمان های مختلف مراجعه کنند. سپس، هر بیمار به یکی از روش های ممکن معالجه شود. حال اگر بخواهیم طول عمر آنها را مطالعه کنیم، عمل برش به یکی از صورت های زیر انجام می شود:

1. **عدم مراجعه:** ممکن است بیمار تصمیم بگیرد به پزشک دیگری مراجعه کند. درننتیجه دیگر او را نخواهیم دید.
2. **قطع معالجه:** ممکن است به علت عوارض بد جانبی، درمان آن را متوقف کنیم. یا اینکه، ممکن است بیمار هنوز در تماس باشد ولی از ادامه معالجه خودداری کند.
3. **اتمام بررسی:** نمودار زیر یک بررسی ممکن را تشریح می کند:



در اینجا بیمار (1) در زمان t=0 مورد بررسی قرار گرفته و در زمان T_1 فوت شده است. در نتیجه، یک مشاهده بریده نشده به دست آمده است. بیمار(2) مورد بررسی قرار گرفته و در پایان بررسی هنوز زنده است. در نتیجه یک مشاهده بریده نشده T_2^+ به دست آمده است. بالاخره بیمار (3) مورد بررسی قرار گرفته و قبل از پایان بررسی، معالجه را قطع نموده است. در نتیجه، یک مشاهده بریده شده T_3^+ به دست آمده است.

درباره برش تصادفی، فرض اساسی زیر را در نظر می گیریم:

فرض: متغیرهای تصادفی C_i و T_i مستقل اند.

بدون این فرض نتایج کمتری در دسترس است. معقول به نظر می رسد که فرض کنیم: مطالعه در زمان های تصادفی شروع، به تصادف بررسی و قطع می شود. با این وجود، اگر دلیلی برای قطع بررسی در جریان مطالعه وجود داشته باشد، در این صورت ممکن است بین T_i و C_i ، رابطه ای وجود داشته باشد.

1-8-4 انواع دیگر برش

انواع دیگر برش در منابع دیگر مورد بررسی قرار گرفته اند. انواع قبلی برش به برش راست[1] و برش چپ[2] تقسیم می شوند. اگر متغیر مورد مطالعه بسیار بزرگ باشد و نخواهیم آن را به طور کامل مشاهده کنیم، آن را « برش راست » نامند. به طور مشابه « برش چپ » نیز قابل تعریف است. برای مثال، در برش تصادفی چپ، تنها می توان (Y_1 , ε_1) ، ... ، (Y_n , ε_n) را مشاهده نمود، که Y_i و ε_i به شرح زیراند:

$$Y_i = \max (T_i , C_i) = T_i \ \text{v} \ C_i$$

$$\varepsilon_i = \text{l} (C_i \leq T_i)$$

مثال: کودکان آفریقایی: در ای مثال هر دو برش راست و چپ انجام می شود. یک روانپزشک می خواهد سن گروه خاصی از کودکان آفریقایی را برای انجام یک عمل ویژه بداند. وقتی به روستای مورد نظر می رسد، تعدادی از بچه ها از قبل می دانند آن عمل را چگونه انجام دهند. در نتیجه، این کودکان مشاهدات بریده شده چپ را ارائه می کنند. بعضی از کودکان یادگیری را شروع کرده اند و سن یادگیری آنها را می توان ثبت کرد. هنگام بازگشت پژوهشگر، هنوز برخی از کودکان روستایی این عمل را یاد نگرفته اند. آنها مشاهدات بریده شده راست را نتیجه می دهند.

يكي از انواع پيچيده برش حالتي است كه زمان دقيق رخداد پيشامد مشخص نيست و فقط اين اطلاع موجود است كه پيشامد در يک بازه زماني اتفاق افتاده است كه به آن برش فاصله اي[3] مي گويند.

برش های راست و چپ؛ هردو حالت های خاصی از برش فاصله ای هستند که در آنها فقط مشاهداتی مورد توجه اند که در یک بازه قرار گیرند. اگر T_i برش تصادفی راست باشد، مشاهدات T_i در بازه$[C_i, \infty)$ قرار می گیرند. هم چنین، اگر T_i برش تصادفی چپ باشد، مشاهدات T_i در بازه $(-\infty, C_i]$ قرار می گیرند.

---

[1] Right censoring
[2] Left censoring
[3] Interval censoring

در برابر برش فاصله ای ، قطع مشاهدات[1] قرار دارد. دراینجا، اگر متغیر مورد علاقه در خارج بـازه معینی قرار گیرد، از آن چشم پوشی می شود. برای مثال، فرض کنید بخـواهیم توزیـع و میـانگین انـدازه یـک عنصر داخل سلول را تعیین کنیم. طبیعی است به علت محـدودیت وسـایل انـدازه گیـری، اگـر انـدازه عنصـر، کوچکتر از اندازه تعیین شده باشد، قابل تشخیص نیست (میلر، 2001).

2 فصل دوم

روش های برآوردیابی در تحلیل بقا

2-1 مقدمه

ارتقاء روش های تحلیل داده های بقاء به عنوان یکی از حوزه های علم آمار در سال های اخیـر بسـیار مورد توجه قرار گرفته است. این بدان خاطر است که در بسیاری از موقعیت هـای علمـی، محققـین تمایـل بـه بررسی زمان بقاء تا رخداد یک واقعه از قبیل؛ تاریخ انقضاء یک ماده، یا زمان مرگ یک بیمار و ... دارنـد.

[1] Truncation

به هر حال تعیین زمان دقیق رخداد یک واقعه قابل مشاهده نیست و تنها یک فاصله زمانی قابل مشاهده است (رجائی فرد و همکاران، 1388).

داده های مربوط به زمان بقاء به روش های آماری ویژه ای برای پیش بینی، برآورد و تجزیه و تحلیل نیاز دارد. در مدل های پارامتریک از روش برآورد درست نمایی ماکسیمم برای برآورد پارامترهای مجهول استفاده می شود. قائل شدن برخی مفروضات و انتخاب یک توزیع احتمال فرضی برای زمان های بقاء استنباط آماری را دقیق تر نموده و انحراف معیار برآوردها را نسبت به زمانی که چنین مفروضاتی وجود نداشته باشند را کوچکتر خواهد کرد (صانعی، 1380).

یکی از موارد دشوار در تحلیل بقاء تنوع روش هاست. مدل های شتابدار زمان شکست از قبیل مدل وایبل، نمایی و لگ نرمال به عنوان روش های پارامتری[1] و روشهای کاپلان – میر و جداول طول عمر، به عنوان روش های ناپارامتری[2] از مهم ترین رویکردهای موجود برای تحلیل داده های بقاء می باشد. اگرچه استفاده از برآوردهای ناپارامتری بسیار گسترده شده اند، اما هنوز درمورد برآوردهای پارامتری که فرض می شود توزیع داده های بقاء معلوم است، نیازمند بحث هستند.

با توجه به اینکه در عمل توزیع واقعی داده ها معلوم نیست، بنابراین استفاده از برآوردهای ناپارامتری از جهت اینکه این روش ها نیازمند پیروی از توزیع خاصی نیستند، می توانند جایگزین مناسبی باشند (روشنی و همکاران، 1390).

2-2 توزیع های مورد استفاده در تحلیل بقاء (روش های پارامتری)

به دست آوردن توزیع پارامتری زمان بقا از اهمیت ویژه ای برخوردار است، به طوری که می توان احتمال بقا و احتمال خطر را در تمام زمان ها محاسبه نمود و در برآورد پارامترها از روش حداکثر درست نمایی استفاده کرد. در این زمینه از توزیع های ویژهای برای برازش به داده ها در تحلیل بقاء استفاده می شود. توزیع هایی که به طور معمول در تحلیل داده های بقاء کاربرد دارند عبارتند از: نمایی، وایبل، لگ – نرمال و توزیع گاما.

2-2-1 توزیع نمایی[3]

با توجه به سادگی محاسبات وخواص منحصر به فرد، توزیع نمایی یکی از مهمترین مدل های پارامتری پرکاربرد در تحلیل های آماری به شمار می رود. « ساده ترین توزیع در مطالعات بقاء، نمایی است که عموما برای توصیف الگوی طول عمر سیستم های الکترونیک به کار رفته است (صادقی و کاظمی، 1387)». ابتدا تعریفی ساده از این توزیع را مطرح می کنیم و سپس به کاربرد آن در تحلیل بقاء

[1] Parametric method
[2] Non – parametric method
[3] Exponential distribution

می پردازیم.

تعریف **1:** متغیر تصادفی[1] T دارای توزیع نمایی است هرگاه تابع چگالی آن به صورت زیر باشد:

$$f(t;\lambda) = \lambda e^{-\lambda t} \qquad ;t > 0\,,\lambda > 0$$

تابع توزیع آن به صورت زیر است:

$$F(t) = 1 - e^{-\lambda t} \qquad ;t \geq 0$$

زیرا؛ با تغییر متغیر u به جای t داریم:

$$F(u) = \int_0^t f(u)\,du = \int_0^t \lambda e^{-\lambda u}\,du = -e^{-\lambda u}\Big|_0^t = -e^{-\lambda t} + 1 = 1 - e^{-\lambda t}.$$

میانگین، واریانس و تابع مولد گشتاور توزیع نمایی به صورت زیر است:

میانگین	$E(T) = {}^{1}/_{\lambda}$
واریانس	$Var(T) = {}^{1}/_{\lambda^2}$
تابع مولد گشتاور	$M_y(u) = {}^{\lambda}/_{\lambda - u}$

جدول 1-2. میانگین، واریانس و تابع مولد گشتاور توزیع نمایی

یک ویژگی جالب متغیر تصادفی نمایی، T، این است که به ازای هر زوج داریم:

$$P[T > a + b] = P[T > a]\,P[T > b] \quad ;a > 0\,,b > 0$$

زیرا ،

$$P[T > a + b] = 1 - P[T \leq a + b] = 1 - \left[1 - e^{-\lambda(a+b)}\right] = e^{-\lambda a}\,e^{-\lambda b}$$

از طرفی،

$$P[T > a] = 1 - P[T \leq a] = 1 - [1 - e^{-\lambda a}] = e^{-\lambda a}$$

به طور مشابه، $P[T > b] = e^{-\lambda b}$. پس با جایگذاری این دو مقدار در نابرابری بالا داریم:

$$P[T > a + b] = P[T > a]\,P[T > b]$$

[1] Random variable

به عنوان مثال اگر متغیر تصادفی T، معرف مدت زمان کار یک مؤلفه ماشینی از موقع نسب تا موقع خرابی باشد و اگر $a = 50$ و $b = 35$ ساعت فرض شود، آنگاه احتمال اینکه مدت زمان کار مؤلفه بیش از $a + b = 85$ ساعت باشد مساوی احتمال این است که مدت زمان کار مؤلفه بیش از 50 ساعت باشد ضربدر احتمال اینکه مدت زمان کار مؤلفه بیش از 35 ساعت باشد.

از ویژگی های دیگر توزیع نمایی، خاصیت بی حافظگی متغیر تصادفی آن است. با اتکا به این ویژگی، از این توزیع می توان برای طول عمر چیزهای گوناگون مورد استفاده قرار داد. قضیه زیرا که موسوم به قضیه بی حافظگی است را داریم:

قضیه 1: اگر متغیر تصادفی T دارای توزیع نمایی با پارامتر λ باشد، آن گاه،

$$P\{T > a + b | T > a\} = P[T > b] \qquad ; a > 0\,, b > 0$$

اگر T نشان دهنده ی طول عمر قطعه معینی باشد، آن گاه قضیه بالا بیانگر آن است که احتمال شرطی[1] که این قطعه بیشتر از $a + b$ واحد زمان دوام بیاورد، با فرض اینکه بیشتر از a واحد زمان دوام کرده است، با احتمال اولیه آنکه بیشتر از b واحد زمان دوام بیاورد، برابر است. راه دیگر بیان این مطلب آن است که بگوییم یک قطعه «کهنه» درحال کار توزیع طول عمری همانند یک قطعه «نو» در حال کار دارد یا اینکه این قطعه کهنه در معرض خستگی و فرسودگی نیست (الکساندر و همکاران، 1913).

تابع مخاطره تابع نمایی با میانگین $1/\lambda$، برابر λ است. زیرا:

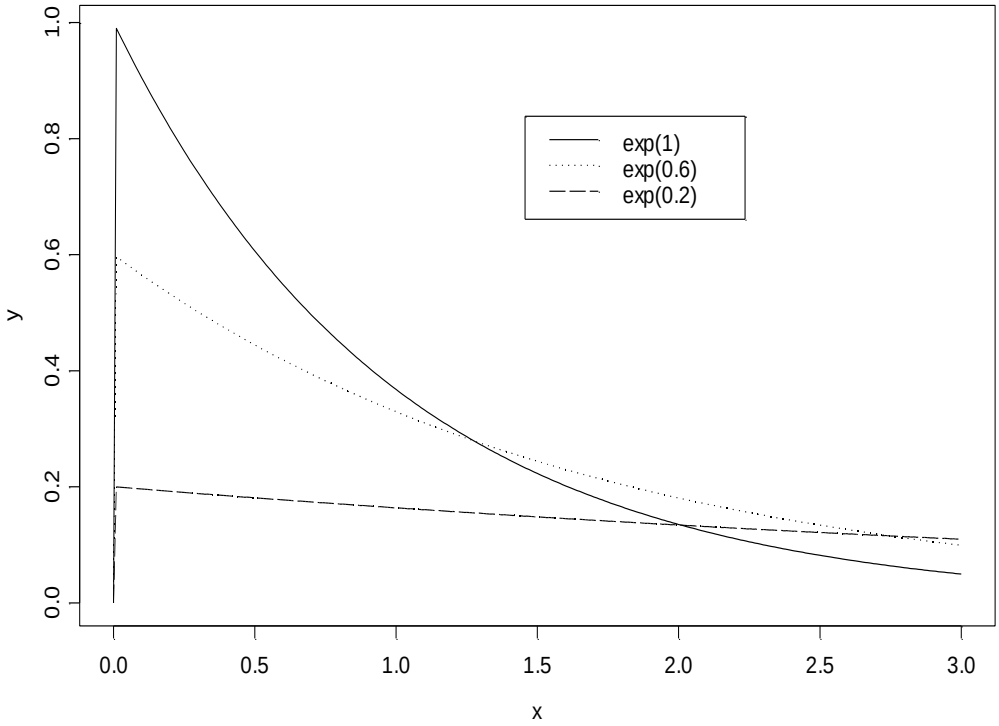
$$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{\lambda e^{-\lambda t}}{1 - 1 + e^{-\lambda t}} = \lambda$$

که مقدار کم λ نشان دهنده ریسک پایین و بقای طولانی تر است. هم چنین با محاسبه ای ساده تابع بقای این توزیع را به صورت زیر به دست می آوریم:

$$S(t|\,\lambda) = \mathrm{e}^{-\lambda \mathrm{t}}.$$

درشکل 1-2، نمودار چگالی توزیع نمایی به ازای پارامترهای مختلف نشان داده شده است:

[1] Conditional probability



شکل 1-2. نمودار چگالی توزیع نمایی به ازای پارامترهای مختلف

2-2-2 توزیع وایبل[1]

این توزیع در سال ۱۹۳۹ توسط والدی وایبل ابداع گردید و امروزه متداول ترین مدل مورد استفاده در مطالعات قابلیت اعتماد[2] است. این توزیع به طور وسیع در شاخه های مختلف مهندسی برای مدل بندی زمان های شکست کاربرد زیادی دارد.

دلیل اهمیت این توزیع رفتار تابع نرخ مخاطره آن است که با تغییر پارامترهای وایبل می تواند تابعی صعودی، نزولی و یا ثابت باشد و به خاطر رفتار تابع نرخ شکست آن این توزیع برای الگو سازی داده های مختلف دارای انعطاف پذیری زیادی است.

تعریف 2: متغیر تصادفی T دارای توزیع وایبل است هر گاه تابع چگالی آن به صورت زیر باشد:

$$f(t;\alpha,\beta) = \alpha\beta\, t^{\beta-1}\, e^{-\alpha t^{\beta}} \qquad ;\alpha,\beta > 0 \quad , \; t > 0$$

[1] Weibull distribution

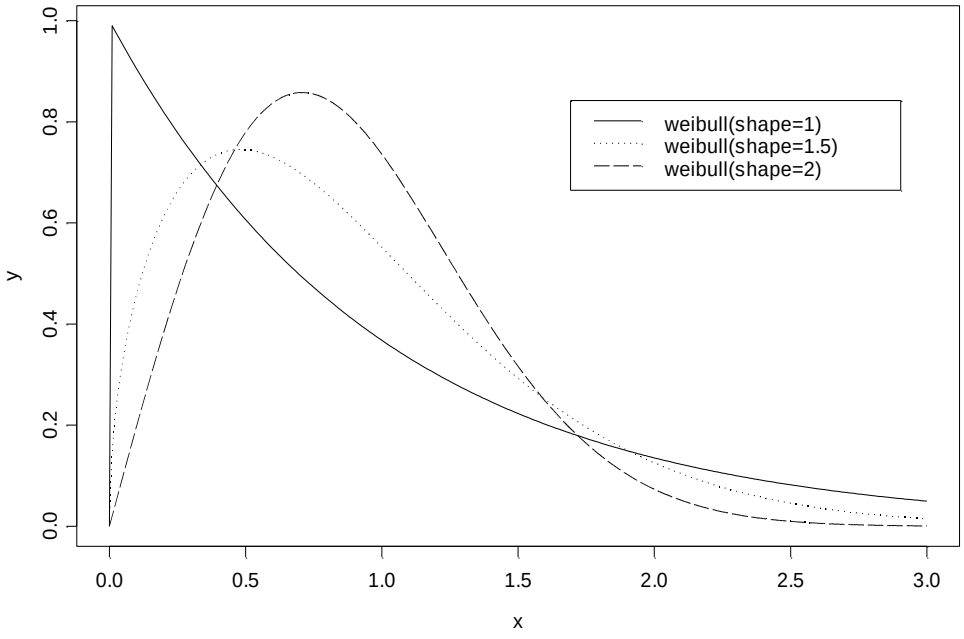
[2] مبحث قابلیت اطمینان یکی از شاخه هایی است که امروزه به طور وسیعی در شاخه های مختلف علوم از جمله پزشکی و مهندسی مورد استفاده قرار می گیرد.

این توزیع دارای دو پارامتر α و β می باشد. α ، پارامتر شکل و β پارامتر مکان است. برای $\beta=1$ چگالی وایبل به چگالی نمایی تبدیل می شود. این چگالی دارای میانگین و واریانس زیر است:

میانگین	$E(T) = (\frac{1}{\alpha})^{1/\beta} \Gamma(1+\beta^{-1}$
واریانس	$V(T) = (\frac{1}{\alpha})^{2/\beta} [\Gamma(1+2\beta^{-1}) - \Gamma^2(1+\beta^{-1})]$

جدول 2-2 . میانگین و واریانس توزیع وایبل

شکل 2-2، نمودار چگالی توزیع وایبل را به ازای پارامتر شکل مختلف و پارامتر مکان ثابت، نشان می دهد:



شکل 2-2. نمودار چگالی توزیع وایبل به ازای مقادیر مختلف پارامتر شکل

همانطورکه در نمودار بالا مشاهده می کنیم، منحنی توزیع وایبل با مقدار پارامتر شکل برابر 1، دقیقا با منحنی توزیع نمایی با پارامتر 1 برابر است. به همین دلیل می توان گفت که توزیع وایبل حالت خاصی از توزیع نمایی است.

توابع بقاء و خطر توزیع وایبل به صورت زیر است:

$$S(t|\,\alpha,\beta) = e^{(-\beta\, t^{\alpha})}.$$

$$h(t|\ \alpha, \beta) = \ \alpha\beta\, t^{\alpha-1}.$$

اثبات: اولا؛

$$F(t) = \int_0^{\infty} f(t)\, dt = \int_0^{\infty} \alpha\beta\ t^{\beta-1}\ e^{-\alpha t^{\beta}}\ dt =$$

بنابراین

$$S(t) = 1 - F(t) =$$

هم چنین

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{\alpha\beta\ t^{\beta-1}\ e^{-\alpha t^{\beta}}}{1 -} =$$

توزیع وایبل تعمیمی از توزیع نمایی است که در آن نرخ مخاطره ثابت نبوده و لذا کاربرد وسیعتری نسبت به توزیع نمایی دارد. اگر $\beta = 1$ ، آنگاه نرخ مخاطره در طول زمان ثابت است. اگر $\beta > 1$، آنگاه با افزایش زمان، نرخ مخاطره کاهش و وقتی که $\beta < 1$ نرخ مخاطره افزایش می یابد. یا به طریقی دیگر، برای این توزیع اگر $\alpha = 1$ باشد، خطر ثابت، اگر $\alpha > 1$ ، خطر افزایشی و اگر $\alpha < 1$ ، آنگاه خطر کاهشی خواهد بود. بنابراین توزیع وایبل برای مدل کردن توزیع بقای جامعه با ریسک افزایشی، کاهشی یا ثابت ممکن است مناسب تر باشد. توزیع وایبل در تمامی حالات که تابع خطر یکنوا یا ثابت است به کار می رود. از طرفی با تغییر پارامترهای این توزیع، چگالی وایبل می تواند برای زمان های چوله به راست و زمان های چوله به چپ قابل استفاده باشد. با توجه به خواص گفته شده دیده می شود که توزیع وایبل انعطاف پذیری زیادی دارد و به همین دلیل در تحلیل بقاء به صورت متداول کاربرد دارد.

در تحلیل پارامتری برای استنباط در مورد متغیرهای کمکی و برازش دقیق تر، فرض می شود پارامتر مکان با متغیرهای کمکی مرتبط است و پارامتر شکل به متغیرهای کمکی وابسته نیست و ثابت در نظر گرفته می شود. برای این منظور، پارامتر مکان را به صورت زیر با متغیرهای کمکی مرتبط می کنند:

$$\lambda\ = \exp\left(\beta_0 + \sum_{i=1}^{p} \beta_j * X_{\mathrm{j}}\right)$$

در عین حال، این فرض که پارامتر شکل توزیع، تحت تأثیر متغیرهای کمکی نیست برای برخی از داده ها، نامناسب است. لذا در این حالت پارامتر شکل را به صورت زیر با متغیرهای کمکی مرتبط می کنند:

$$\alpha = \exp\left(\alpha_0 + \sum_{i=1}^{p} \alpha_j * X_j\right)$$

2-2-3 توزیع لگ - نرمال[1]

توزیع پارامتری دیگری که معمولا در تحلیل بقا استفاده می شود، لگ – نرمال است. ساده ترین شکل آن به عنوان متغیری است که لگاریتم آن دارای توزیع نرمال باشد. کاربرد این توزیع در علوم زیست و زمان بقای بیماری هایی مانند انواع سرطان و آلزایمر مطرح است. دلیل آن چوله به راست بودن توزیع این داده ها و نرمال بودن لگاریتم زمان های بقاء است.

تعریف 3 : فرض کنید T متغیر تصادفی مثبت است و متغیر تصادفی جدید Y هم به صورت $Y = \log T$ تعریف شده است. اگر Y دارای توزیع نرمال باشد، آنگاه گوییمT دارای توزیع لگ – نرمال است. چگالی توزیع لگ- نرمال به صورت زیر است:

$$f(t;\ \mu,\sigma^2) = \frac{1}{t\sqrt{2\pi\sigma}}\ exp\left[-\frac{1}{2\sigma^2}(\log t - \mu)^2\right]$$

که در آن: $-\infty < \mu < \infty$ و $\sigma > 0$.

برای متغیر تصادفی لگ– نرمال T داریم:

مینگین	$E(T) = \exp\ (\mu + \frac{1}{2\sigma^2})$
واریانس	$Var(T) = \exp(2\mu + 2\sigma^2) - \exp\ (2\mu + \sigma^2)$

جدول 3-2. میانگین واریانس توزیع لگ – نرمال

همچنین اگرT دارای توزیع لگ – نرمال باشد:

$$E[\log(T)] = \mu \quad \text{و} \quad V[\log(T)] = \sigma^2.$$

تابع مخاطره این توزیع ابتدا تا یک نقطه ماکزیمم افزایش یافته و سپس کاهش می یابد. بنابراین توزیع لگ- نرمال برای الگوهای بقاء، که نرخ مخاطره آن ها ابتدا افزایش یافته و سپس کاهش می یابد، مناسب است.

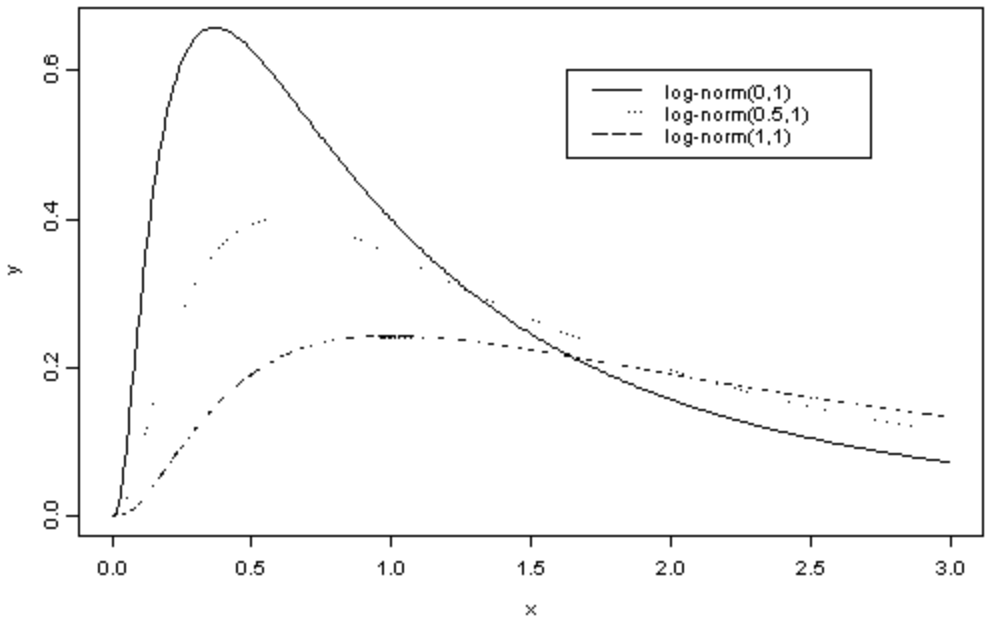
تابع بقای این توزیع به صورت زیر می باشد:

[1] Log – Normal distribution

$$S(t|\,\mu,\sigma) = 1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right)$$

اثبات:

در شکل 3-2، نمودار چگالی توزیع لگ – نرمال به ازای میانگین های مختلف و واریانس های ثابت مشاهده می کنیم:



شکل 3-2. نمودار چگالی توزیع لگ – نرمال به ازای پارامترهای مختلف

2-2-4 توزیع گاما[1]

در حساب دیفرانسیل و انتگرال تابعی را که به تابع گاما موسوم است به صورت زیر تعریف می کنند:

[1] Gamma distribution

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}\, e^{-x}\, dx\,, \qquad \alpha > 0$$

با انتگرال گیری به روش جزء به جزء نتیجه می گیرند، $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$. واضح است که وقتی α صحیح و مثبت باشد، $\Gamma(\alpha) = (\alpha-1)!$ است. $\Gamma(\alpha)$ را می توان به ازای مقادیر گویای α نیز به دست آورد. به ویژه متذکر می شویم که $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

تابع چگالی گاما را می توان به دو صورت تعریف کرد. ما در اینجا هر دو تعریف را می آوریم ولی بنای کارمان را در این مطالعه بر اساس تعریف دوم قرار می دهیم:

تعریف اول:

تعریف 4: متغیر تصادفی X دارای توزیع گاماست وقتی که تابع چگالی آن به صورت زیر باشد:

$$f(x) = \frac{1}{\beta^{\alpha}\,\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-x/\beta}$$

که درآن: $x > 0\,, \alpha > 0\,, \beta > 0$. برای سایر نقاط مقدار تابع چگالی برابر صفر است.

تعریف دوم:

تعریف 5: متغیر تصادفی T دارای توزیع گاماست هرگاه تابع چگالی آن به صورت زیر باشد:

$$f(t;\, \alpha, \beta) = \frac{\beta}{\Gamma(\alpha)}\, (\beta t)^{\alpha-1} e^{-\beta t} \quad ;\ \alpha > 0\,, \beta > 0\,, t > 0.$$

توزیع گاما تعمیمی از توزیع نمایی است. اگر در توزیع گاما مقدار α=1 را قرار دهیم به توزیع نمایی دست می یابیم. نکته دیگری که باید به آن توجه کنیم این است که مجموع متغیرهای تصادفی $(T_i\,; i = 1,2,\dots,n)$ که مستقل و هم توزیع با توزیع نمایی هستند، دارای توزیع گاما می بشد.

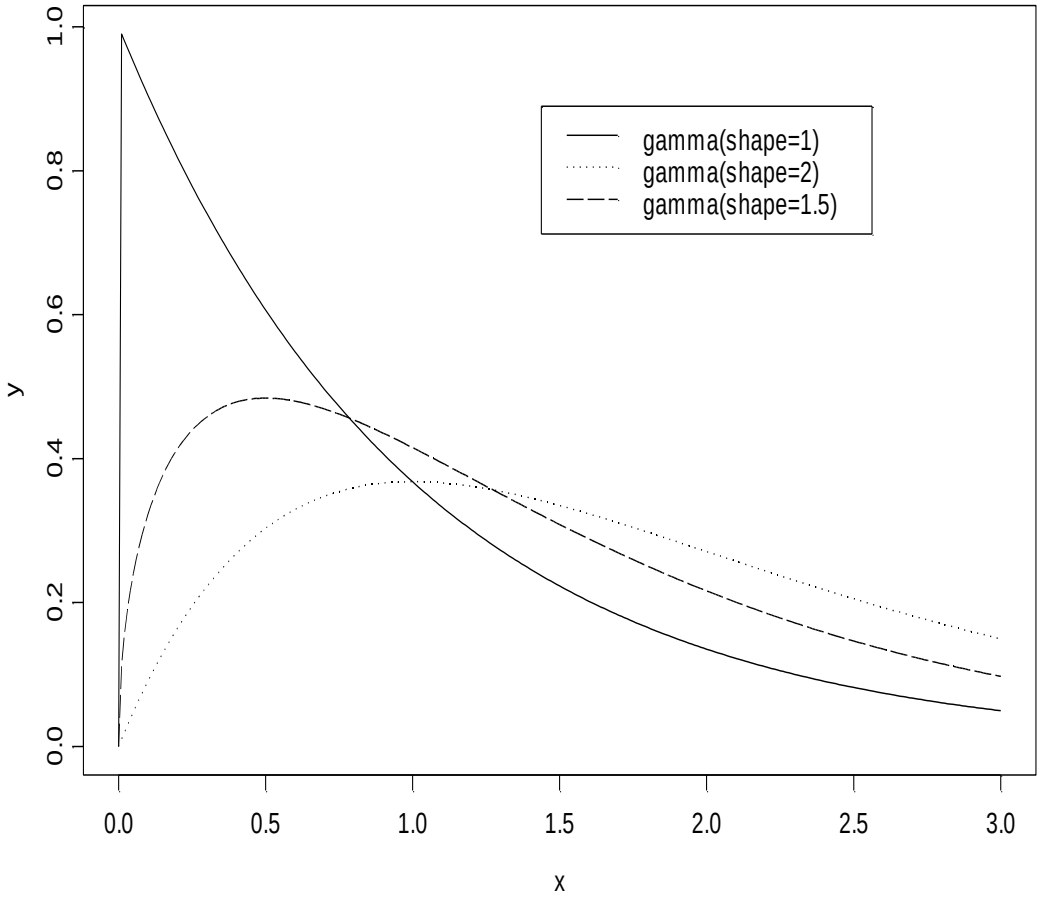
اگرT دارای توزیع گاما با پارامترهای α و β باشد و $\alpha < \beta$ ، آنگاه:

میانگین	$E(T) = \frac{\alpha}{\beta}$
واریانس	$V(T) = \frac{\alpha}{\beta^2}$

تابع مولد گشتاور	$M_t(u) = (\frac{\beta}{\beta - t})^{\alpha}$

جدول 4-2 . ویژگی های توزیع گاما

نمودارهای تابع چگالی گاما به ازای مقادیر مختلف α و βی ثابت، متفاوت اند. در شکل 4-2، چهار نوع خم نمایش این تابع ها را ملاحظه می کنید:



شکل 4-2. نمودار چگالی توزیع گاما به ازای پارامترهای مختلف

در نمودار بالا ملاحظه می شود که منحنی توزیع گاما با مقدار $\alpha = 1$ ، مشابه منحنی توزیع نمایی با پارامتر 1 می باشد. بنابراین توزیع گاما نیز حالت خاصی از توزیع نمایی است. از توزیع گاما همچون توزیع های دیگر معرفی شده در این فصل کاربردهای مهمی در برازش دادن داده های بقاء دارد. از این توزیع در مباحث قابلیت اعتماد در صنعت و توزیع بقای انسان به کار می رود. در توزیع گاما با پارامترهای α و β ، وقتی که $0 < \alpha < 1$ ، نرخ مخاطره به طور یکنوا از بینهایت تا β با افزایش زمان، از صفر تا

بی نهایت، کاهش می یابد. وقتی $\alpha > 1$ نرخ مخاطره به طور یکنوا از صفر تا λ با افزایش زمان، از صفر تا بینهایت، افزایش می یابد. اگر $\alpha = 1$ نرخ مخاطره برابر β است.

قضیه 2: اگر متغیر تصادفی T دارای توزیع گاما با پارامترهای α و β باشد، که در آن α عدد صحیح مثبتی است، آنگاه:

$$F(t) = 1 - \sum_{j=0}^{r-1} \frac{e^{-\beta t}\,(\beta t)^j}{j!}.$$

برای $\beta = 1$ در عبارت بالا، $F(t)$ را **تابع گامای ناقص** گویند. این تابع به طور گسترده ای جدول بندی شده است. با استفاده از این تابع می توانیم صورت کلی تابع بقای توزیع گاما را بنویسیم:

$$S(t) = 1 - \int_0^t f(u)\,du = 1 - \left(\frac{\text{تابع گاما ناقص}}{\text{تابع گاما کامل}}\right)$$

2-3 روش برآورد

برآوردهای توابع بقاء و مخاطره در مطالعات مقدماتی تحلیل بقاء حائز اهمیت هستند. چراکه می توان توزیع داده ها را با دانستن آنها به دست آورد و هم چنین انتخاب مدل برای آنالیز بیشتر را فراهم کرد و از سویی امکان اعتبارسنجی این مدل ها را میسر می سازند. برآورد اثرات متغیرها بر منحنی ماندگاری براساس روش های درست نمایی ماکسیمم[1] صورت می گیرد. با این حال ساخت تابع درست نمایی در مدل های پارامتری نیازمند روش های متفاوتی است.

در مدل های پارامتریک می توان نشان داد که سهم داده سانسور نشده در درست نمایی برابر با مقدار تابع چگالی در زمان شکست می باشد و سهم داده سانسور شده در درست نمایی برابر با مقدار تابع ماندگاری در زمان سانسور شدن می باشد. در نتیجه تابع درست نمایی کامل از حاصل ضرب سهم مستقل از هر یک از رکوردها به دست می آید. سپس برآورد پارامترها از طریق ماکسیمم کردن این تابع یا لگاریتم آن به دست می آید.

2-3-1 برآوردگر درست نمایی ماکسیمم[2]

برای تعریف کردن برآوردگر درست نمایی ماکزیمم ابتدا تابع درست نمایی را تعریف می کنیم:

تعریف 6- **تابع درست نمایی**: تابع درست نمایی n متغیر تصادفی $X_1, X_2, \dots, X_n$، با چگالی توأم n متغیر تصادفی مانند $f_{x_1,\dots,x_n}(x_1, \dots, x_n;\ \theta)$ تعریف می شود که به صورت تابعی از θ در نظر گرفته

[1] Maximum likelihood (ML)
[2] Maximum likelihood estimator (MLE)

می شود. به خصوص اگر $X_1, X_2, \dots, X_n$ نمونه ای تصادفی از چگال $f(x;\theta)$ باشد آنگاه تابع درست نمایی برابر $f(x_1;\theta)\ f(x_2;\theta) \dots f(x_n;\theta)$ می باشد (الکساندر و همکاران، 1913).

نمادگذاری: برای تداعی اینکه تابع درست نمایی تابعی از θ است، نماد $L(\theta, x_1, \dots, x_n)$ را برای تابع درست نمایی به کار می بریم. درست نمایی مقدار یک تابع چگالی است.

تعریف 7- **برآوردگر درست نمایی ماکسیمم**: فرض کنید $L(\theta) = L(\theta; x_1, \dots, x_n)$تابع درست نمایی برای متغیرهای تصادفی $X_1, X_2, \dots, X_n$ باشد. اگر $\hat{\theta}$ (که $\hat{\varphi} = \hat{\vartheta}(x_1, x_2, \dots, x_n)$ تابعی از مشاهدات $x_1, x, \dots, x_n$ می باشد) مقدار θ در φ باشد که $L(\theta)$ را ماکسیمم می کند، آنگاه $\hat{\varphi} = \hat{\vartheta}(x_1, x_2, \dots, x_n)$ برآورد درست نمایی ماکسیمم θ است (الکساندر و همکاران، 1913).

بشتر توابع درست نمایی در شرایط نظم صدق می کند، به طوری که برآوردگر درست نمایی ماکسیمم جواب معادله $\frac{dL(\theta)}{d(\theta)} = 0$ می باشد. هم چنین $L(\theta)$ و $\log L(\theta)$ درمقدار یکسانی از θ ماکسیمم می شوند و گاهی اوقات پیدا کردن ماکسیمم لگاریتم درست نمایی آسان تر است.

اینک الگوی برش تصادفی را در نظر می گیریم. (توجه شود که در اینجا برش نوع اول را با فرض $c_i = t_c$ در نظر می گیریم. درست نمایی برای برش نوع دوم، مشابه نوع اول است. با این تفاوت که به علت منظور کردن ترتیب، یک ضریب ثابت وجود دارد).

تابع درست نمایی زوج مرتب (y_i, δ_i) به شرح زیر است:

$$L(y_i, \delta_i) = \begin{cases} f(y_i) & \delta_i = 1 \quad \text{بدون برش} \\ s(y_i) & \delta_i = 0 \quad \text{بریده شده} \end{cases}$$

$$= f(y_i)^{\delta_i} . s(y_i)^{1-\delta_i}$$

تابع درست نمایی کامل نیز به شرح زیر است:

$$L = L(y_1, \dots, y_n, \delta_1, \dots, \delta_n) = \prod_{i=1}^{n} L(y_i, \delta_i)$$

$$= (\prod_u f(y_i)) \left(\prod_c s(y_{i)})\right)$$

در واقع تابع های درست نمایی برای برش تصادفی، به شرح زیر است:

$$L(y_i, \delta_i) = \begin{cases} f(y_i)[1 - G(y_{i)}] & \delta_i = 1 \\ g(y_i)S(y_i) & \delta_i = 0 \end{cases}$$

$$L = (\prod_u f(y_i))\ (\prod_c S(y_i))\ (\prod_c g(y_i))\ (\prod_u [1 - G(y_i)])$$

اگر فرض شود که زمان برش با زمان بقاء ارتباط ندارد، دو حاصل ضرب آخری $\prod_c g(y_i)$ و $\prod_u [1 - G(y_i)]$ ، شامل پارامترهای مجهول نیستند. درنتیجه، این دو را می توان در بیشینه کردن L ثابت فرض کرد.

فرض کنید، $\underline{\theta} = (\theta_1, .., \theta_p)'$ ، بردار پارامتر باشد. محاسبه $max\ L(\underline{\theta})$ با محاسبه جواب $\underline{\hat{\theta}}$ معادلات درست نمایی زیر یکسان است:

$$\frac{\partial}{\partial \theta_j} \log L(\underline{\theta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \log(y_i, \delta_i)$$

$$= \sum_u \frac{\partial}{\partial \theta_j} \log f_{\underline{\theta}}(y_i) + \sum_c \frac{\partial}{\partial \theta_j} \log S_{\underline{\theta}}(y_i) = 0 \qquad j = 1,2, \dots, p$$

معمولا، محاسبه جواب به کمک رایانه و روش های عددی امکان پذیر است. روش های نیوتون – رافسون[1] و روش امتیازی (چوب خطی)[2] از جمله این روش ها هستند. در ادامه با ذکر دو مثال از توزیع های نمایی و وایبل به کاربرد این مبحث در برآورد توزیع های بقاء می پردازیم:

مثال 1- نمایی: تحت برش تصادفی، فرض کنید n_u ، تعداد مشاهدات بریده نشده باشد، در این صورت داریم:

$$L = \lambda^{n_u} \exp\left(-\lambda \sum_u t_i - \lambda \sum_c c_i\right) = \lambda^{n_u} \exp\left(-\lambda \sum_{i=1}^{n} y_i\right)$$

$$\log L = n_u\ \log \lambda - \lambda \sum_{i=1}^{n} y_i$$

$$\frac{\partial}{\partial \lambda} \log L = \frac{n_u}{\lambda} - \sum_{i=1}^{n} y_i$$

$$\hat{\lambda} = \frac{n_u}{\sum_{i=1}^{n} y_i}$$

[1] Newton – Raphson method
[2] Method of scoring

$$\frac{\partial^2}{\partial \lambda^2} \log L = \frac{-n_u}{\lambda^2}$$

$$i(\underline{\lambda}) = \frac{n_u}{\lambda^2}$$

توجه شود، که $\hat{\lambda} = \frac{n_u}{\sum_{i=1}^{n} y_i}$ ، برآورد MLE تحت برش اول ودوم – به خوبی برآورد تحت برش تصادفی- نیز هست.

مثال 2 – وایبل: اگر توزیع وایبل را با پارامتر $\gamma = \beta^{\alpha}$ بنویسیم، می توان مشتقات آن را ساده تر محاسبه کرد.

$$S(t) = e^{-(\beta t)^{\alpha}} = e^{-\gamma t^{\alpha}}$$

$$f(t) = \gamma \alpha t^{\alpha-1}. e^{-\gamma t^{\alpha}}$$

در این صورت، داریم:

$$L = (\gamma\alpha)^{n_u} \left(\prod_u {t_i}^{\alpha-1} \right) exp\left(-\gamma \sum_u {t_i}^{\alpha} \right) exp\left(-\gamma \sum_c {c_i}^{\alpha} \right)$$

$$= (\lambda\alpha)^{n_u} \left(\prod_u {t_i}^{\alpha-1} \right) exp\left(-\gamma \sum_{i=1}^{n} {y_i}^{\alpha} \right)$$

$$\log L = n_u \log \gamma + n_u \log \alpha + (\alpha - 1) \sum_u \log t_i - \gamma \sum_{i=1}^{n} {y_i}^{\alpha}$$

$$\frac{\partial}{\partial \gamma} \log L = \frac{n_u}{\gamma} - \sum_{i=1}^{n} {y_i}^{\alpha}$$

$$\frac{\partial}{\partial \gamma} \log L = \frac{n_u}{\alpha} + \sum_u \log t_i - \gamma \sum_{i=1}^{n} {y_i}^{\alpha} \log y_i$$

بنابراین برآورد درست نمایی ماکسیمم $(\hat{\alpha}, \hat{\gamma})$ به شرح زیر است:

$$\hat{\gamma} = \frac{n_u}{\sum_{i=1}^{n} {y_i}^{\hat{\alpha}}}$$

$$\frac{n_u}{\hat{\alpha}} + \sum_u \log t_i - \hat{\gamma} \sum_{i=1}^{n} y_i^{\hat{\alpha}} \log y_i = 0$$

این معادلات را باید به روش عددی حل کرد.

برآورد $S(t)$: برآورد تابع بقاء، یکی از اهداف اصلی تحلیل بقاء است.

$$S(t) = exp\left(-\int_0^t \lambda(u)du\right)$$

برای مثال، یکی از مقیاس های معالجه سرطان این است که احتمال زنده بودن بیمار را به مدت حداقل پنج سال حساب کنیم. در مطالعات مهندسی تابع بقا را تابع اعتماد نامند و معمولا آن را با $R(t)$ نشان می دهند. برای مثال دانستن قابلیت اعتماد یک مؤلفه در یک دستگاه بعد از هزار ساعت کار آن دستگاه.

باداشتن MLE ، برآورد تابع بقاء در دو حالت نمایی یا وایبل بسیار ساده است.

حالت نمایی $\hat{S}(t) = e^{-\hat{\lambda}t}$

حالت وایبل $\hat{S}(t) = e^{-(\hat{\lambda}t)^{\hat{\alpha}}} = e^{-\hat{\gamma}t^{\hat{\alpha}}}$

همچنین برای هر مقدار ثابت t ، تابع بقاء $\hat{S}(t)$ خود تابعی از $\hat{\lambda}$ یا $(\hat{\gamma}, \hat{\alpha})$ است.

2-4 روش های ناپارامتری

در بررسی داده های بقاء اگر هدف، توصیف زمان بقاء بدون در نظر گرفتن متغیرهای کمکی باشد، از روش تحلیل ناپارامتری مانند جدول طول عمر و کاپلان- میر استفاده می شود.

2-4-1 جدول طول عمر[1]

روش رسمی برآورد $S(t)$ در علوم بیماری های مسری و بیمه گری، همان روش بیمه گری است که به جدول طول عمر بستگی دارد و در زیر آن را شرح می دهیم.

فرض کنید زمان به دنباله های ثابتی از بازه های $I_1, \dots, I_k$ ، افراز شده باشد. این بازه ها تقریبا همیشه (و نه

لزوما) مساوی در نظر گرفته می شوند. برای جمعیت انسان ها این فاصله ها معمولا یک سال است . بنا به

[1] Life table

تعریف، در یک جدول طول عمر موارد زیر فرض می شوند:

n_i = تعداد افراد زنده در شروع بازه I_i .

d_i = تعداد افراد مرده در طول بازه I_i .

l_i = تعداد گم شده ها در طول بازه I_i .

w_i = تعداد خارج شده ها در طول بازه I_i .

p_i = احتمال زنده بودن در طول بازه I_i ، در صورتی که در ابتدای بازه زنده بوده است. $1 - p_i = q_i$.

جدول 5-2، مثالی از یک جدول طول عمر است. I_1 تا I_5 ، هر کدام به مدت یک سال اند. ستون (2) شاملn_i ، ستون (3) شامل d_i ، ستون (4) شامل l_i و ستون (5) شامل w_i است. می خواهیم $S(5)$ را برآورد کنیم.

جدول 5-2. محاسبه نرخ بقای 5 ساله

سال های بعد از تشخیص بیماری (1)	تعداد زنده ها در ابتدای بازه (2)	تعداد مرده ها در طول بازه (3)	گم شده ها (4)	خارج شده ها (5)
1-0	126	47	4	15
2-1	60	5	6	11
3-2	38	2	-	15
4-3	21	2	2	7

5-4	10	-	-	6

ادامه

جدول 5-2

تعداد مؤثر در معرض خطر مرگ (6)	نسبت مرگ و میر (7)	نسبت زنده ها (8)	نسبت تجمعی زنده ها تا پایان بازه (9)
116/5	0/40	0/60	0/60
51/5	0/10	0/90	0/54
30/5	0/07	0/93	0/50
16/5	0/12	0/88	0/44
7/0	0/00	1/00	0/44

2-4-1-1 روش کاهش نمونه[1]

برای برآورد $S(\tau_k)$ ، تنها افرادی را مورد توجه قرار می دهیم که در بازه $[0, \tau_k)$ در معرض خطر قرار دارند (یعنی: در بازه مورد توجه). موارد زیر را داریم:

$$n = n_1 - \sum_{i=1}^{k} l_i - \sum_{i=1}^{k} w_i$$

$$d = \sum_{i=1}^{k} d_i$$

$$\hat{S}(\tau_k) = 1 - \frac{d}{n}$$

[1] Reduced sample method

در مثال

جدول 5-2 ، داریم:

$$n = 126 - 12 - 56 = 60$$

و

$$d = \sum_{i=1}^{5} d_i = 56$$

درنتیجه:

$$\hat{S}(5) = 1 - \frac{56}{60} = 0.067$$

عیب روش کاهش نمونه این است که از اطلاعات موجود در l_i و w_i چشم پوشی می کند. این یک برآورد اریب (نقصانی) از $S(t)$ است.

2-4-1-2 روش بیمه گری[1]

می توان احتمال بقاء $S(\tau_k)$ را به صورت حاصل ضرب چند احتمال تجزیه کرد. داریم:

$$S(\tau_k) = P(T > \tau_k)$$

$$= P(T > \tau_1).P(T > \tau_2|T > \tau_1) \dots P(T > \tau_k|T > \tau_{k-1})$$

$$= P_1.P_2 \dots P_k$$

در این رابطه: $P_i = P(T > \tau_i|T > \tau_{i-1})$.

روش بیمه گری، یک برآوردی جدا برای هر P_i ، ارائه می دهد. سپس، با ضرب این برآورها، برآورد $S(\tau_k)$ نتیجه می شود. در برآورد P_i ، می توان از $1 - \frac{d_i}{n_i}$ ، استفاده کرد. به شرطی که مشاهده ای در I_i گم یا خارج نشده باشد. با این وجود، اگر l_i و w_i صفر نباشند، فرض می کنیم که به طور متوسط، افرادی که در بازه I_i حذف شده اند، در معرض خطر در نصف بازه بوده اند. بنابراین، حجم مؤثر نمونه را

[1] Actuarial method

به صورت زیر تعریف می کنیم:

$$\widehat{p_i} = 1 - \widehat{q_i} \quad ; \quad \widehat{q_i} = \frac{d_i}{\acute{n}_i} \quad ; \quad \acute{n}_i = n_i - \frac{1}{2}(l_i + w_i)$$

در نتیجه، برآورد روش بیمه گری معادل: $\hat{S}(\tau_k) = \prod_{i=1}^{k} P_i$ است. در

Error! Reference source not found.

جدول 5-2، ستون (6) شامل $\acute{n}$، ستون (7) شامل $\widehat{q_i}$، ستون (8) شامل $\widehat{p_i}$ و در ستون (9) دیده می شود که $S(5) = 0.44$ است.

برای بهبودبخشی در پیدا کردن جانشینی برای حجم نمونه مؤثر، تلاش های زیادی به انجام رسیده است. ولی اگر یک برآورد دقیق تری برای $S(t)$ لازم شود، حد حاصل ضرب برآورگر کاپلان – میر ، روش مناسبی خواهد بود.

2-4-1-3 انواع جدول های طول عمر

Error! Reference source not found. ، مثالی از یک جدول طول عمر گروهی[1] است. یک گروه یا دسته، مجموعه ای از افراد است که در جریان بررسی، مورد مطالعه قرار دارند. افراد در معرض خطر در ابتدای بازه ی I_i ، همان افرادی هستند که در بازه ی قبلی (I_{i-1}) زنده مانده اند. (نمرده یا گم یا خسته نشده اند).

نوعی دیگر از جدول طول عمر، جدول طول عمر جاری[2] است. در این جدول گروهی از افراد با سن τ_{i-1} در نظر گرفته می شوند، که در ابتدای بازه $I_i = (\tau_{i-1}, \tau_i]$ در معرض خطر قرار دارند. این نوع از افراد، کاملا متفاوت از افرادی هستند که در بازه قبلی I_{i-1} در معرض خطر بوده اند. نوعا، گروه های سنی مختلف در جمعیت در یک زمان بررسی می شوند (میلر، 2001).

2-4-2 برآورد گر حدی حاصل ضرب کاپلان - میر[3]

برآورگر ناپارامتری استاندارد تابع احتمال بقاء، برآورد کپلان – میر است. این برآوردگر که به

[1] Cohort life table
[2] Current life table
[3] Kaplan – Meier Product-limit estimator

عنوان حد حاصل ضربی نیز شناخته می شود، به صورت زیر تعریف می گردد:

$$S(t) = \begin{cases} 1 & t \le t_1 \\ \prod_{t_i \le t}\left[1 - \frac{d_i}{y_i}\right] & t \ge t_1 \end{cases}$$

که t_1 ، زمان مشاهده اولین شکست و d_i ، نمایانگر تعداد شکست ها و y_i ، تعداد افرادی است که رخداد مورد نظر را تجربه نکرده اند. واضح است که قبل از وقوع اولین شکست احتمال بقاء برابر با یک است و پس از آن مقدار حاصل ضربی کاهش می یابد (روشنی و همکاران، 1390).

برآوردگر حدی حاصل ضرب (PL) ، همانند برآوردگر بیمه ای است با این تفاوت که طول بازه های I_i متغیراند. در واقع τ_i ، انتهای راست بازه ی I_i مشاهده بریده شده یا بریده نشده مرتبه iام است. به خاطر داشته باشید که مشاهدات تکراری وجود ندارند و زوج های (Y_1, δ_1) تا (Y_n, δ_n) را مشاهده می کنیم. فرض کنید $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$ ، آماره های مرتب Y_1 تا Y_n باشند. هم چنین ، فرض کنید: $\delta_{(i)}$ مقدار δ متناظر با $Y_{(i)}$ باشد. یعنی: اگر $Y_{(i)} = Y_j$ ، آنگاه $\delta_{(i)} = \delta_j$ ، باشد. توجه شود که $\delta_{(1)}$ تا $\delta_{(n)}$ مرتب نیستند، فرض کنید $R(t)$ ، تابع خطر در زمان t باشد. یعنی: مجموعه ی افرادی که تا زمان t^- زنده هستند.

هم چنین n_i تعداد اعضای $R(Y_i)$ ، که در زمان $Y_{(i)}^-$ زنده هستند، d_i تعداد افرادی که در زمان $Y_{(i)}$ فوت شده و p_i ، احتمال زنده بودن در طول I_i باشد (به شرطی که در ابتدای بازه I_i زنده باشند). به عبارت معادل:

$$p_i = P(T > \tau_i | T > \tau_{i-1}) \text{ و } q_i = 1 - p_i .$$

برآوردهای $\widehat{p_i}$ و $\widehat{q_i}$ ، به شرح زیراند:

$$\widehat{p_i} = 1 - \widehat{q_i} = \begin{cases} 1 - \frac{1}{n_i} & \delta_{(i)} = 1 \\ 1 & \delta_{(i)} = 0 \end{cases}$$

حال برآورد PL ، در صورت نبود تکرار، به شرح زیر است:

$$\hat{S}(t) = \prod_{y_{(i)} \le t} \widehat{p_i} = \prod_{u; y_{(i)} \le t} \left(1 - \frac{1}{n_i}\right) = \prod_{y_{(i)} \le t} \left(1 - \frac{1}{n_i}\right)^{\delta_{(i)}}$$

$$= \prod_{y_{(i)} \le t} \left(1 - \frac{1}{n-i+1}\right)^{\delta_{(i)}} = \prod_{y_{(i)} \le t} \left(\frac{n-i}{n-i+1}\right)^{\delta_{(i)}} .$$ (کاپلان و میر، 1958)

مثال: مطالعه درمان AML : یک آزمایش بالینی به منظور ارزشیابی یک روش شیمیایی روی

بیماری (AML) صورت گرفته است. بعد از رسیدن به مرحله ی خاصی، بیماران به تصادف به دو گروه تقسیم شده اند. گروه اول تحت درمان شیمیایی قرار گرفته اند و گروه دوم یا شاهد، بدون درمان مورد مطالعه قرار گرفته اند. هدف این است که آیا درمان شیمیایی زمان شفا یافتن را به تأخیر می اندازد؟ به عبارت دیگر، آیا سبب افزایش زمان بهبودی می شود؟

برای یک تحلیل اولیه در طول دوره ی آزمایش، داده ها (برحسب 10/74) به شرح زیر بوده است. طول بهبودی کامل برحسب هفته است.

گروه تیمار $9, 13, 13^+, 18, 23, 28^+, 31, 34, 45^+, 48, 161^+$

گروه شاهد $5, 5, 8, 8, 12, 16^+, 23, 27, 30, 33, 43, 45$

برآوردگر PL (کاپلان – میر) برای گروه تیمار، به صورت زیر محاسبه می شود:

$$\hat{S}(0) = 1$$

$$\hat{S}(9) = \hat{S}(0) \times \frac{10}{11} = 0.91$$

$$\hat{S}(13) = \hat{S}(9) \times \frac{9}{10} = 0.82$$

$$\hat{S}(18) = \hat{S}(13) \times \frac{7}{8} = 0.72$$

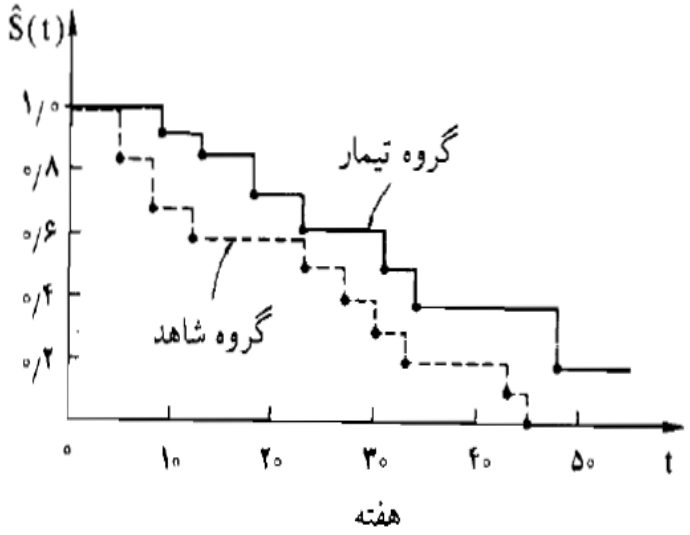
$$\hat{S}(23) = \hat{S}(18) \times \frac{6}{7} = 0.61$$

$$\hat{S}(31) = \hat{S}(23) \times \frac{4}{5} = 0.49$$

$$\hat{S}(34) = \hat{S}(31) \times \frac{3}{4} = 0.37$$

$$\hat{S}(48) = \hat{S}(34) \times \frac{1}{2} = 0.18$$

درنمودار زیر، برآوردگرهای PL برای گروه تیمار و گروه شاهد نشان داده شده است (امبوری، 1971).



شکل 5-2 . نمودار برآورد تابع بقا برای مطالعه درمان AML

2-5 برآوردگرهای تابع مخاطره

می دانیم تابع مخاطره به صورت زیر تعریف می شود:

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

همچنین، مشکلات برآورد $\lambda(t)$ معادل برآورد تابع چگالی است. حالت ساده تر، برآورد تابع مخاطره تجمعی[1] است. این تابع را به صورت زیر در نظر می گیریم:

$$\Lambda(t) = \int_0^t \lambda(u)du$$

توابع Λ و S با رابطه $S(t) = e^{-\Lambda(t)}$ با هم ارتباط دارند. برای سادگی فرض می کنیم تکرار وجود ندارد. نلسون (1969)، $\Lambda(t)$ را به صورت زیر برآورد می کند:

$$\hat{\Lambda}(t) = \hat{\Lambda}_2(t) = \sum_{y_{(i)} \le t} \frac{\delta_{(i)}}{n - i + 1}$$

برآورد پترسن (1977) Λ ، به شرح زیر است:

$$\widehat{\Lambda_1}(t) = \sum_{y_{(i)} \le t} -\log\left(1 - \frac{\delta_{(i)}}{n - i + 1}\right)$$

[1] Cumulative hazard function

دو برآوردگر بالا بسیار نزدیک به هم اند، زیرا برای مقادیر کوچک x ، $\log(1-x) \cong -x$ است. نکته ای که باید به آن اشاره کنیم این است که برآوردگر پترسن متناظر برآوردگر PL (کاپلان – میر) تابع بقاء است. داریم:

$$\hat{S}_2(t) = e^{-\hat{\Lambda}_2(t)}$$

فلمینگ و هارینگتن (1979) ، $\hat{S}_2(t)$ را به عنوان یک برآوردگر دیگر تابع بقاء پیشنهاد می کنند و نشان می دهند که در بعضی از مواقع دارای میانگین مربع خطای کوچکتری است.

2-6 برآورد تجربی[1] تابع بقاء

تابع بقای تجربی ما را از توزیع زمان های بقاء آگاه می سازد. تابع توزیع تجربی را می توان از طریق فرمول کاپلان – میر محاسبه کرد:

$$\hat{S}_{KM}(t) = \prod_{K \mid T_{[K]<t}} \left(\frac{n_k - d_k}{n_k}\right).$$

در این فرمول، $\hat{S}_{KM}(t)$ مقدار تابع بقاء در زمان t و $T_{[K]}$ نمایانگر زمان های شکست، به ترتیب از کوچکترین به بزرگترین و d_k تعداد افرادی است که در زمان T_{K} دچار شکست شده اند. $\hat{S}_{KM}(t)$ ، برآورد حد حاصل ضرب یا برآورد کاپلان – میر تابع بقاء نامیده می شود. هم چنین $\hat{S}_{KM}(t)$ برآورد درست نمایی ماکسیمم $S(t)$ شامل همه توزیع های ممکن است. در ضمن از $\hat{S}_{KM}(t)$ ، تابع مخاطره تجربی، یعنی $\lambda(y)$ ، برآورد می شود.

[1] Tentativ estimate

3 فصل سوم

مدل رگرسیون کاکس

3-1 مقدمه

به طور کلی دو مدل رگرسیونی برای بررسی داده های بقاء وجود دارد. یکی مدل شتاب دار زمان شکست است که به عنوان مدل های پارامتریک شناخته می شود و در فصل قبل، توزیع های پارامتری مربوط به این مبحث مورد بررسی قرار گرفت. مدل رگرسیونی دیگر، مدل کاکس[1] می باشد که مدلی نیمه پارامتریک است. اغلب پژوهشگران در زمینه پزشکی، متمایل به استفاده از مدل های نیمه پارامتریک همچون کاکس هستند. مدل کاکس علی رغم داشتن برخی محدودیت ها به عنوان رایج ترین مدل به منظور مدل سازی عوامل مؤثر بر بقاء به کار می رود (ترنیو و گرامبسچ، 2000).

الزامی نبودن یک توزیع احتمالی برای زمان های بقاء یکی از مزایای مدل مخاطره متناسب کاکس می باشد ولی یک پیش فرض مهم و اساسی در این مدل وجود دارد و آن ، فرض متناسب بودن خطر برای تمامی متغیرهای مستقل موجود در مدل نهایی می باشد. در صورت برقراری این فرض تفسیر مدل به دست آمده ساده تر از مدل های پارامتری خواهد بود (رجائی فرد و همکاران، 1388).

به هر حال این مدل یک مدل نیمه پارامتری می باشد و نیاز به برآورد پارمترهای مدل توزیع های زمان شکست (بقاء و مخاطره) لازم است. کاکس از روش درست نمایی شرطی[2] برای برآورد استفاده می کند. مهم ترین ویژگی این روش این است که تنها به ترتیب رخ دادن حوادث بستگی دارد نه به زمان دقیق وقوع آنها.

3-2 مدل رگرسیونی

در بسیاری از حالات، خصوصیات اصلی توابع بقاء یا تابع چگالی شناخته شده نبوده اما برخی اطلاعات در مورد تغییرات میزان شکست $\lambda(t)$ در دسترس می باشد. بنابراین برای تعیین عوامل پیشگویی وابسته به زمان بقاء، مدل های تحلیل بقاء، معمولا از روی تابع مخاطره که شکست در زمان t را نشان می دهد، ساخته می شوند. فرض کلی در این حالت این است که تابع مخاطره برای همه داده های بقاء دارای فرم پایه مشترکی است و مفهوم میانگین کل را دارد.

---

[1] Cox model
[2] Conditional likelihood

از آنجا که $\lambda(t)$ یک تابع مثبت است، تابع مخاطره به صورت تابعی خطی از متغیرهای توضیحی بیان می شود:

$$\lambda(t) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$

که در آن، $\beta_i ; i = 1,2, \dots, p$ها به عنوان ضرائب رگرسیونی مدل معرفی می شوند. در این مدل فرض می شود تابع مخاطره به زمان بستگی ندارد.

مدل رگرسیونی عمومی توابع بقاء یا ماندگاری را می توان به شکل زیر نوشت:

$$\lambda(t, x_i) = \lambda_0(t) \exp(x\beta).$$

که $\lambda(t, x_i)$ بیانگر مخاطره برای متغیر iام بوده، $\lambda_0(t)$ عبارت از تابع مخاطره پایه[1] برای هر متغیر بوده و تنها به زمان وابسته است.

3-3 مدل مخاطره نسبی و بسط آن

مدل مخاطره نسبی یکی از مدل هایی است که در تحلیل بقاء به خصوص در تحلیل طول عمر به طور وسیعی مورد استفاده قرار می گیرد. برقرار بودن مخاطره نسبی به این معناست که در این مدل، مخاطره دو فرد A و B (بعنوان مثال) در طی زمان، با میزان مخاطره $exp(x_A - x_B)\beta$ ثابت می ماند. به عبارت دیگر، نسبت خطر برای یک نفر در زمان مشخصی چون t به خطر برای فرد دیگر در همان زمان، مقدار ثابتی داشته باشد. در صورت برقراری این فرض تفسیر مدل به دست آمده ساده تر و راحت تر از مدل های پارامتری خواهد بود. با این وجود در برخی حالات فرض مخاطره نسبی همواره برقرار نیست. چرا که مخاطره برای هریک از افراد مورد نظر در طی زمان بسته به شاخص های مخاطره تغییر می کند.

خصوصیت دیگر مدل های مخاطره نسبی این است که در این مدل ها می توان متغیرهایی که در این مدل مسئول تغییر در سیاست حذف در طی زمان هستند، را در مدل وارد کرد. در این حالت محور زمان به فواصل زمانی متعددی تفکیک می شود و به طوری که در هر فاصله زمانی فرض مخاطره نسبی برقرار بوده اما میزان مخاطره در هر فاصله زمانی متفاوت است.

3-4 روش های بررسی متناسب بودن خطرات

بنابراین مدل رگرسیونی مناسب برای تحلیل بهتر داده های مربوط به زمان بقاء مدلی است که این فرض در آن برقرار باشد. روش های گوناگونی برای ارزیابی متناسب بودن خطرات برای مدل ها ی رگرسیونی وجود دارد. به عنوان مثال، باقیمانده های اسکور[2]، اسکنفیلد[3] و لگاریتم خطر تجمعی[1] برای

[1] Basis hazard function
[2] Score
[3] Schoenfeld

ارزیابی برقرار بودن متناسب بودن خطر در مدل مخاطره متناسب کاکس به کار می روند (حسینی و همکاران، 1386). به اختصار دو مورد آخر را توضیح می دهیم:

3-4-1 روش باقیمانده های اسکنفلد

فرض کنیم که p متغیر توضیحی و n مشاهده مستقل از زمان و δ_i نماد سانسور شدن باشد به طوری که برای داده های سانسور نشده مقدار یک، و برای داده های سانسور شده مقدار صفر را به خود بگیرد. iامین باقیمانده اسکنفلد برای jامین متغیر توضیحی، یعنی x_j ، که در آن ؛ $j = 1,2, \dots, p$ می باشد، به صورت زیر به دست می آید:

$$r_{s_{ji}} = \delta_i\{x_{ji} - a_{ji}\}$$

که در آن x_{ji} ، مقدار jامین متغیر توضیحی برای iامین فرد مورد مطالعه است.

$$a_{ji} = \frac{\sum_{l\varepsilon R(t_i)} x_j \left(\exp\left(\hat{\beta} x_l\right)\right)}{\sum_{l\varepsilon R(t_i)} \exp\left(\hat{\beta}\, x_l\right)}$$

در رابطه بالا $R(t_i)$ مجموعه افراد در معرض خطر در زمان t_i می باشد. باید توجه داشت که مقادیر غیرصفر این باقیمانده فقط برای مشاهدات سانسور نشده آورده می شود.

باقیمانده های اسکنفلد یک برآوردی از iامین مؤلفه رتبه اثر برای jامین پارامتر مدل ارائه می دهد. موقعی که مدل خطرات متناسب برقرار باشد، نمودار باقیمانده های اسکنفلد به صورت تصادفی در طول زمان مطالعه پراکنده می باشند و این نمودار در مقابل مقادیر متغیر پیشگوکننده (متغیر توضیحی) مورد نظر دارای شکل و روند مشخص نمی باشد (حسینی، 1390).

3-4-2 روش گرافیکی لگاریتم خطر تجمعی

منحنی بقا (log-log) یک منحنی برآوردی است که در نتیجه دوبار لگاریتم گیری از تابع بقاء برآوردی ،$\hat{S}$ ، حاصل می شود و در واقع انتقالی از $\hat{S}$ می باشد و آن را به طور خلاصه به صورت $\ln(-\ln(s))$ نشان می دهند. در حالت کلی برای بررسی فرض متناسب بودن خطرات برای دو یا چند گروه زمان بقاء می توان منحنی لگاریتم خطر تجمعی را برای گروه های تحت بررسی ترسیم نمود، چنانچه منحنی ها با هم موازی باشند، نشان دهنده متناسب بودن خطرات است و در غیر این صورت در متناسب بودن خطرات باید شک نمود (حسینی، 1390).

تا کنون مطالعات زیادی در رابطه با به کارگیری مدل رگرسیون کاکس انجام شده ولی براساس یک

[1] Cumulative Hazard Log

مطالعه سیستماتیک تنها در 5 درصد این مطالعات فرض متناسب بودن خطرات مورد بررسی قرار گرفته شده است (رجائی فرد و همکاران، 1388).

3-5 مدل خطرات متناسب کاکس

یکی از عمومی ترین مدل ها در تحلیل داده های بقاء، مدل نیمه پارامتری کاکس است. علت نیمه پارامتری بودن این است که فرضیه ای درباره فرم پارامتری بودن تابع مخاطره پایه آن وجود ندارد. این مدل اولین بار توسط کاکس، آمار شناس معروف انگلیسی در سال 1972 به منظور بررسی اثرات متغیرهای توضیحی (مستقل) تأثیرگذار بر زمان بقاء ارائه شد که به صورت زیر است:

$$\lambda_i(t) = \lambda_0(t)exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

که در آن، تابع $\lambda_0(t)$ تابع خطر پایه در زمان t است. $exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$ را تابع خطر نسبی و β_iها پارامترهای رگرسیون می باشند. و اما تفسیر این مدل از روی مدل رگرسیونی عمومی، که در ابتدای فصل معرفی شد، این گونه است که برای زمان بقای T_i و برای واحدهای مطالعاتی iام فرض می شود که تابع خطر شرطی که وابسته به متغیر کمکی $X_i(t)$ است، در شرایط زیر صادق است:

$$\lambda_i\big(t \mid X(t)\big) = \lambda_0(t)\ \exp\ (\beta X_i(t))$$

که $\lambda_0(t)$ ، تابع مخاطره پایه برای فردی است که متغیر توضیحی اش برابر صفر باشد یعنی X=0 . این تابع پایه برای همه واحدهای مطالعه یکسان است. ضریب نامعلوم β وابستگی بین T_i و متغیرهای کمکی $X_i(t)$ را بازگو می کند. یک فرضیه مهم این است که شکست واحدهای مختلف، مستقل از یکدیگر رخ می دهند و مقدار تابع متغیر کمکی برای یک واحد بر روی بقاء هیچکدام از واحدها اثر نمی گذارند (روشنی و همکاران، 1390).

از زمانی که کاکس این مدل را ارائه کرده است، نقش اساسی را در تحلیل بقاء ایفا کرده است. این مدل فرض می کند که میزان مخاطره حاصل ضربی از تابع نامشخص از زمان مشترک برای همه واحدهای مطالعه در تابع پیوند معلوم از ترکیب خطی متغیرهای کمکی است.

مدل کاکس یک مدل استوار است. بنابراین با وجود نامشخص بودن توزیع تابع خطر پایه، برآوردهای ضرایب رگرسیونی و سایر نتایج به دست آمده از این مدل، بسیار نزدیک به مدل های پارامتری می باشد. بنابراین حتی زمانی که توزیع تابع خطر پایه در دست اما مشکوک است، استفاده از رگرسیون کاکس منطقی و قابل اطمینان است.

مدل رگرسیون کاکس را، مدل خطرات متناسب هم می نامند و علت آن را نیز این گونه بیان می کنند

که در این مدل هیچ فرضی درباره توزیع تابع خطر پایه در نظر گرفته نمی شود و میزان خطر برای یک فرد، نسبت ثابت و مشخصی را با میزان خطر فرد دیگری دارد. به عبارتی دیگر فرض کنید که دو فرد i و j ، دارای مقادیر xهای متفاوت بوده و پیشگویی های آن ها خطی به صورت زیر باشند:

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

$$\eta_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk}$$

بنابراین نسبت خطر برای این دو نفر به شکل زیر است:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)\, e^{\eta_i}}{\lambda_0(t)\, e^{\eta_j}} = \frac{e^{\eta_i}}{e^{\eta_j}} = e^{\eta_i - \eta_j}$$

این نسبت خطرات در همه زمان ها مقداری ثابت است. به عبارت دیگر به زمان بستگی ندارد و این یکی از مفروضات مهم مدل کاکس است (حسینی، 1390).

تابع بقای متناسب با این مدل به صورت زیر خواهد بود:

$$S(t,x) = exp\left(-exp(X\beta)\int_0^t \lambda_0(u)\, du\right)$$

این انتگرال تابع تجمعی مخاطره پایه نامیده می شود. روش های گوناگونی برای برآورد این تابع وجود دارد.

مدل کاکس، کاربردی ترین روش برای یافتن ارتباط متغیرهای توضیحی با متغیر پاسخ بقاء یا هر متغیر پاسخ دیگری است که از راست سانسور شده باشد. «شرط استفاده از مدل کاکس در حالت داده های بقاء این است که صرف نظر از متغیرهای توضیحی، بین زمان بقاء ناهماهنگی وجود داشته باشد. به این معنی که به شرط وجود نداشتن مشاهده سانسور شده، تمام اعضایی که دارای متغیرهای توضیحی یکسان هستند، دارای توزیع بقای یکسانی خواهند بود (قدیمی و همکاران، 1389)».

نکته 1: خانواده ای از توزیع ها را « خانواده توزیع های لهمن» گوییم، هرگاه تابع توزیعی مانند F_0 وجود داشته باشد، به گونه ای که در ازای هر F در این خانواده رابطه زیر برقرار باشد:

$$1 - F = (1 - F_0)^{\gamma}$$

در این رابطه γ یک عدد مثبت است. می توان رابطه را برحسب تابع بقاء و به صورت $S = S_0{}^{\gamma}$ ، نیز نوشت. از الگوی مخاطره های متناسب نتیجه می شود که توابع توزیع آنها تشکیل یک خانواده توزیع های لهمن می دهد.

3-6 تحلیل درست نمایی شرطی[1]

ساخت تابع درست نمایی برای برآورد در مدل نیمه پارامتری کاکس، نسبت به مدل های دیگر نیازمند روش های متفاوتی است. برآورد β از طریق ماکسیمم کردن تابع درست نمایی و بدون ایجاد هیچ فرضی در مورد $\lambda(t)$ صورت می گیرد. کاکس اصول روش درست نمایی جزئی[2] را برای اولین بار ارائه داد (کاکس، 1972). تابع درست نمایی جزئی بخشی از تابع درست نمایی کامل[3] بوده و به $\lambda_0(t)$ وابسته نیست. این تابع از طریق درست نمایی حاشیه ای رتبه ای زمان های شکست به دست می آید. پس از حداکثر نمودن تابع درست نمایی جزئی، تابع مخاطره پایه و یا تابع بقاء را می توان با استفاده از برآورد های انجام شده برای ضرائب رگرسیون، محاسبه کرد.

در نوشته های کاکس داریم: فرض کنید $\lambda_0(t)$ اختیاری باشد. هیچ اطلاعی درباره β از فاصله های زمانی، که در آنها هیچ شکستی رخ نداده، به دست نمی آید. زیرا، امکان دارد، $\lambda_0(t)$ به طور قابل درکی با صفر متحد باشد. بنابراین، روی آن لحظه هایی که در آن ها شکست رخ می دهد، شرطی می کنیم. در زمان گسسته نیز روی مشاهدات تکراری شرطی می کنیم. به هر حال نیاز به روشی داریم. که تمام $\lambda_0(t)$ را تحلیل نماید. در نظر گرفتن این توزیع شرطی اجباری به نظر می رسد.

فرض کنید تکرار وجود نداشته باشد. زمان های مشاهده ای را به صورت $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$ ، مرتب می کنیم. فرض کنید، $\delta_{(i)}$ تابع نشانگر و $x_{(i)}$ ، متغیر وابسته به $y_{(i)}$ باشد، می نویسیم:

$\Re_{(i)} = \Re(y_{(i)}-)$. برای هر زمان بریده نشده $y_{(i)}$ ، داریم:

$$P\{[y_{(i)}, y_{(i)} + \Delta y[\text{ یک مرگ در بازه} \mid \mathcal{R}_{(i)}\} \cong \sum_{j \in \mathcal{R}_{(i)}} e^{\hat{\beta} x} \lambda_0(y_{(i)})\Delta y$$

$$P\{y_{(i)} \text{ در زمان } (i) \text{ مرگ} \mid y_{(i)} \text{ در زمان } \mathcal{R}_{(i)} \text{ یک مرگ در }\} = e^{\hat{\beta} x_{(i)}} \Big/ \sum_{j \in \mathcal{R}_{(i)}} e^{\hat{\beta} x_j}$$

اگر حاصل ضرب سه احتمال شرطی را حساب کنیم، درست نمایی شرطی به دست می آید.

$$L_c(\beta) = \prod_u \frac{e^{\hat{\beta} x_{(i)}}}{\sum_{j \in \mathcal{R}_{(i)}} e^{\hat{\beta} x_j}}$$

کاکس پیشنهاد می کند، که از درست نمایی شرطی به جای درست نمایی معمولی استفاده کنیم. به ویژه

[1] Conditional likelihood analysis
[2] Partial likelihood
[3] Full likelihood

پیدا کردن برآورد حداکثر درست نمایی از بردار امتیاز[1] و ماتریس اطلاعات نمونه[2]، به شرح زیر استفاده شود:

$$\frac{\partial}{\partial\beta}\log L_c(\beta) = \left(\frac{\partial}{\partial\beta_1}\log L_c(\beta), \dots, \frac{\partial}{\partial\beta_p}\log L_c(\beta)\right)'$$

$$i(\beta) = -\frac{\partial^2}{\partial\beta^2}\log L_c(\beta) = -\begin{bmatrix} \frac{\partial^2}{\partial\beta_1\partial\beta_1}\log L_c(\beta) & \cdots & \frac{\partial^2}{\partial\beta_1\partial\beta_p}\log L_c(\beta) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\beta_p\partial\beta_1}\log L_c(\beta) & \cdots & \frac{\partial^2}{\partial\beta_p\partial\beta_p}\log L_c(\beta) \end{bmatrix}$$

ماتریس اطلاع نمونه عبارت است از : مجموع ماتریس های کوواریانس برای مجموعه های نرخ شکست مشاهدات سانسور نشده.

می خواهیم معادلات زیر را، که معمولا به روش تکرار منجر می شود، حل کنیم:

$$\frac{\partial}{\partial\beta}\log L_c(\beta) = 0$$

فرض کنید، $\hat{\beta}^0$ یک حدس اولیه باشد، داریم:

$$\hat{\beta}^1 = \hat{\beta}^0 + \mathrm{i}^{-1}(\hat{\beta}^0)\frac{\partial}{\partial\beta}\log \mathrm{L_c}\left(\hat{\beta}^0\right)$$

اگر $\hat{\beta}$ جواب معادله باشد، داریم:

$$\hat{\beta} \sim^a N\left(\beta, \mathrm{i}^{-1}(\beta)\right)$$

یعنی بطور مجانبی دارای توزیع نرمال است (میلر، 2001).

3-7 بررسی درست نمایی شرطی

3-7-1 درست نمایی حاشیه ای برای رتبه ها

مجددا فرض کنید تکرار وجود ندارد. با وجود تکرار استدلال زیر صحیح نیست.

فرض کنید داده ها سانسور نشده اند و Y_1 تا Y_n مستقل و Y_i دارای توزیع F_i و چگالی f_i باشد. دو

[1] Score vector
[2] Sample information matrix

نماد $Y = (Y_1, \dots, Y_n)$ و $R = (R_1, \dots, R_n)$ ، که در آن R_i ، رتبه مشاهده شده Y_i است را در نظر می گیریم.

در این صورت احتمال بردار رتبه r به صورت زیر است:

$$P(r) = \int \dots \int \prod_{i=1}^{n} f_{(i)}\,(u_i)\, du_1 \dots du_n \quad ;\ u_1 < \dots < u_n\ , i = 1 \dots n$$

در این رابطه $f_{(i)}$ چگالی متناظر $y_{(i)}$ است. برای مثال با $n = 3$ و $r = (3,1,2)$ ، داریم:

$$P(r) = P(R_1 = 3, R_2 = 1, R_3 = 2) = \iiint f_2(u_1) f_3(u_2) f_1(u_3)\, du_1\, du_2\, du_3$$

که در آن $u_1 < u_2 < u_3$ (میلر، 2001).

3-7-2 درست نمایی جزئی

دنباله کمیت های تصادفی زیر را در نظر بگیرید:

$$(X_1, S_1;\ X_2, S_2; \dots;\ X_m, S_m)$$

در رگرسیون با برش، فرض کنید: $y_{u(i)}$ ، مشاهده iام بریده نشده باشد. هم چنین فرض کنید: متغیر X_i دارای تمام اطلاعات بریده شده در $[\ y_{u(i-1)}, y_{u(i)})$ ، با این اطلاع که یک شکست در زمان $y_{u(i)}$ رخ داده باشد. S_i را متغیری در نظر می گیریم که شامل مشاهده خاصی باشد، که با همبستگی $x_{u(i)}$ در زمان $y_{u(i)}$ شکست می خورد.

درست نمایی حاشیه ای S_1 تا S_m به شرح زیر است:

$$P(S_1\ , \dots, S_m \mid \beta) = \prod_{i=1}^{m} P(S_i \mid\ S_1, \dots, S_{i-1}\ ;\ \beta)$$

و درست نمایی شرطی S_1 تا S_m با فرض X_1 تا X_m نیز مطابق زیر است:

$$P(S_1\ , \dots, S_m | X_1\ , \dots, X_m\ ;\ \beta) = \prod_{i=1}^{m} P(S_i \mid\ S_1, \dots, S_{i-1}\ ; X_1\ , \dots, X_m;\ \beta)$$

و بالاخره درست نمایی کامل به شرح زیر است:

$$P(X_1, \dots, X_m;\ S_1, \dots, S_m|\ \beta) = \prod_{i=1}^{m} P(X_i, S_i|X_1, \dots, X_{i-1}; S_1, \dots, S_{i-1}; \beta)$$

$$= \prod_{i=1}^{m} P(X_i|X_1, \dots, X_{i-1}; S_1, \dots, S_{i-1}; \beta) \times \prod_{i=1}^{m} P(S_i|X_1, \dots, X_{i-1}; S_1, \dots, S_{i-1}; \beta)$$

کاکس (1975)، عبارت دوم حاصل ضرب بالا را درست نمایی جزئی نامیده است. یعنی عبارت:

$$\prod_{i=1}^{m} P(S_i|X_1, \dots, X_{i-1}; S_1, \dots, S_{i-1}; \beta)$$

در رگرسیون با داده های سانسور شده (بریده شده) ، درست نمایی جزئی با L_c برابر است، که ما آن را درست نمایی شرطی گوییم. از مقایسه تعاریف درست نمایی های شرطی و جزئی، معلوم می شود که، درست نمایی جزئی، یک درست نمایی شرطی یا حاشیه ای واقعی نیست. کاکس مدعی است که: درست نمایی جزئی شامل بیشترین اطلاعات درباره β برای رگرسیون با داده های سانسور شده است. لذا، می توان از جمله اول حاصل ضرب بالا، یعنی عبارت زیر چشم پوشی کرد (میلر، 2001).

$$\prod_{i=1}^{m} P(X_i|X_1, \dots, X_{i-1}; S_1, \dots, S_{i-1}; \beta)$$

3-8 برآورد تابع بقا

تحت مدل مخاطره کاکس، داریم:

$$S(t; x) = exp\left(-e^{\widehat{\beta}x} \int_0^t \lambda_0(u)\, du\right) = exp\left(-e^{\widehat{\beta}x} \Lambda_0(t)\right) = S_0(t)^{-e^{\widehat{\beta}x}}$$

که در آن $S_0(t) = e^{-\Lambda_0(t)}$ است.

برای برآورد $S(t; x)$ ، به جای β از $\hat{\beta}$ استفاده می کنیم. ولی $\Lambda_0(t)$ یا $S_0(t)$ را چگونه برآورد کنیم؟

برسلو (1972) فرض می کند که $\lambda_0(t)$ بین مشاهدات سانسور نشده ثابت است، داریم:

$$\hat{\lambda}_{0,B}(t) = \frac{1}{\left(y_{u(i)} - y_{u(i-1)}\right) \sum_{j \in \Re_{u(i)}} e^{\widehat{\beta}x_j}} \quad , y_{u(i-1)} < t < y_{u(i)}$$

و $S_0(t)$ را به صورت زیر برآورد می کنیم:

$$\hat{S}_{0,B}(t) = \prod_{y_{(i)} \le t} \left(1 - \frac{\delta_{(i)}}{\sum_{j \in \Re_{u(i)}} e^{\hat{\beta} x_j}}\right)$$

توجه شود که $\hat{\Lambda}_{0,B}(t) = \int_0^t \hat{\lambda}_{0,B}(u)\, du$ و $\hat{S}_{0,B}(t)$ ، حتی به ازای $t = y_{(i)}$ ، سازگار نیستند، یعنی:

$$\hat{S}_{0,B}(t) \neq e^{-\hat{\Lambda}_0(t)}$$

هم چنین $\hat{S}_{0,B}(t)$ می تواند مقادیر منفی اختیار کند. تسیاتیس (1978) از مقدار زیر استفاده می کند:

$$\hat{S}_{0,T}(t) = e^{-\hat{\Lambda}_{0,T}(t)}$$

که در آن $\hat{\Lambda}_{0,B}(t) = \sum_{y_{(i)} \le t} \frac{\delta_{(i)}}{\sum_{j \in \Re_{(i)}} e^{\hat{\beta} x_j}}$ است. ولی هنگامی که $\hat{\beta} = 0$ باشد، $\hat{S}_{0,T}(t)$ به برآوردگر PL (کاپلان – میر) ساده نمی شود. توجه نمایید که $\hat{\Lambda}_0(t)$ یک تابع پله ای است.

لینک (1979) یک صورت خطی $\hat{\Lambda}_{0,B}(t) = \sum_{y_{(i)} \le t} \frac{\delta_{(i)}}{\sum_{j \in \Re_{(i)}} e^{\hat{\beta} x_j}}$ را با تعریف زیر به کار می برد، که معادل انتگرال برسلوی $\lambda_0(t)$ است.

$$\hat{S}_{0,L}(t) = e^{-\hat{\Lambda}_{0,L}(t)}$$

برآوردگرهای دیگر $S(t; x)$ ، که از نظر محاسباتی پیچیده ترند توسط کاکس و افرون پیشنهاد شده است.(میلر، 2001)

3-9 نیکویی برازش[1]

در این بخش اعتبار سنجی مدل خطرات متناسب کاکس را با روش ترسیمی همراه با یک مثال را مورد بررسی قرار می دهیم.

روش ترسیمی: به راحتی می توان با نگاه کردن، خط را از منحنی تمییز داد. بنابراین برای روش رسم نمودار از اصل زیر استفاده می کنیم:

اصل اساسی: مقیاس محورهای مختصات را به گونه ای اختیار می کنیم، که اگر مدل برقرار باشد،

[1] Goodnece of fit

نمودار به شکل خط مستقیم و در غیر این صورت به شکل منحنی درآید.

معمولا، دو نوع نمودار بقاء و نرخ شکست (مخاطره) مورد استفاده قرار می گیرند، این دو مورد بسیار نزدیک هم اند و در هر حالت مناسب ترین انتخاب می شود.

نمودارهای بقاء: در این نمودارها یا $\hat{S}(y_{u(i)})$ را در مقابل $y_{u(i)}$ و یا $\hat{S}(t)$ را در مقابل t رسم می کنند. این نوع، حالت خاصی از نمودار Q-Q یا نمودارهای احتمالی است (ویلک وگنانادسیکان، 1968).

نمودارهای نرخ شکست (مخاطره): در این نمودارها یا $\hat{\Lambda}(y_{u(i)})$ درمقابل $y_{u(i)}$ و یا $\hat{\Lambda}(t)$ در مقابل t رسم می شود. برای این کار از رابطه نلسون (فصل دوم، بخش 5 را ببینید) به صورت:

$\hat{\Lambda}_2(t) = \sum_{y_{(i)} \le t} \frac{\delta_{(i)}}{n-i+1}$ و یا رابطه: $\hat{\Lambda}_1(t) = -\log \hat{S}(t)$ ، استفاده می شود (نلسون، 1969).

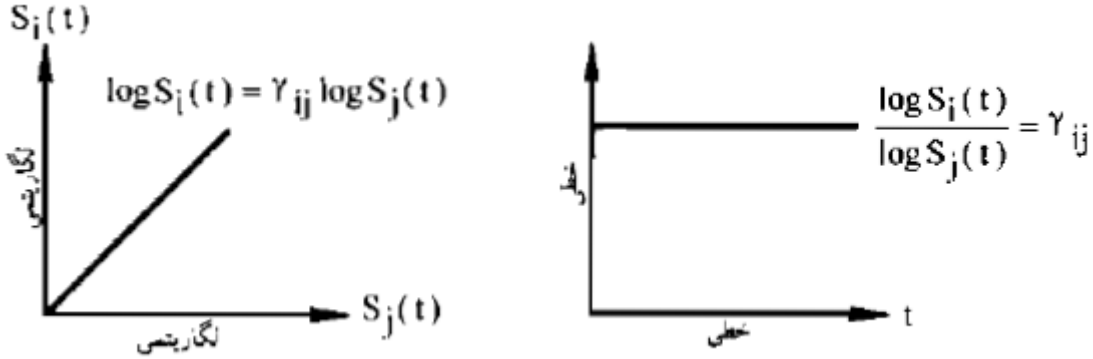
فرض کنید می خواهیم برای K نمونه، اعتبار مدل نرخ شکست متناسب کاکس را بررسی کنیم. تحت الگوی $S_i(t) = S_j(t)^{\gamma_{ij}}$ برای چند مقدار از γ_{ij} ، داریم:

$$\log S_i(t) = \gamma_{ij} \log S_j(t)$$

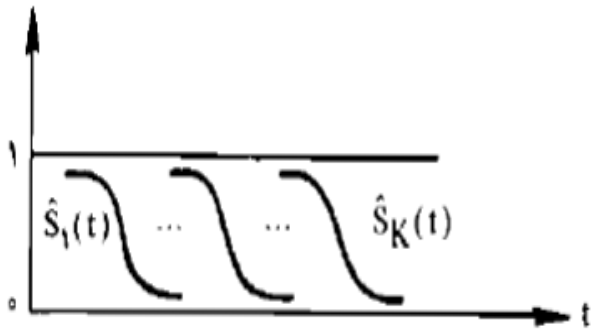
یا

$$\frac{\log S_i(t)}{\log S_j(t)} = \gamma_{ij}$$

برآوردهای جدای PL هریک از $\hat{S}_1(t)$ تا $\hat{S}_K(t)$ را محاسبه و یکی از نمودارهای زیر را می سازیم:

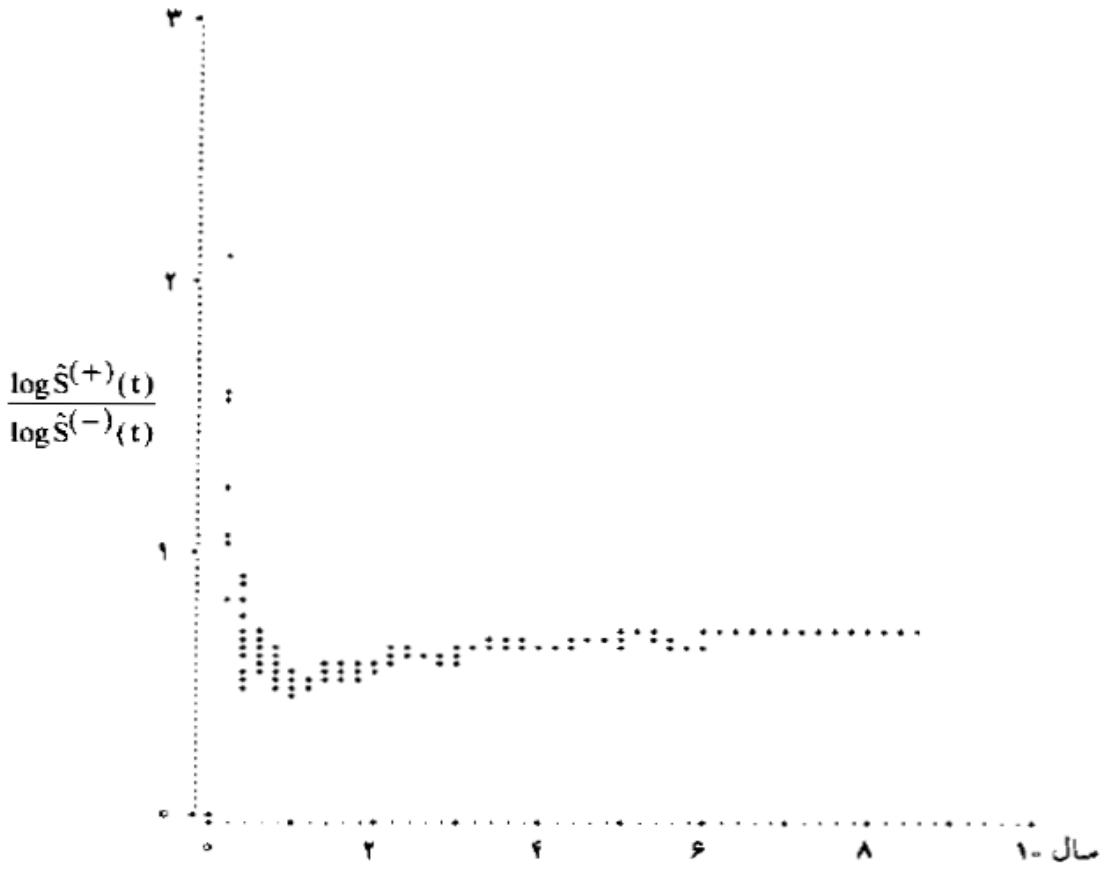


برای وارسی کردن الگوی خطی، برآورد PL را به طور جداگانه برای هریک از K نمونه محاسبه و آن را رسم می کنیم. تغییر مکان ها توسط انتقال وارسی می کنیم.



مثال: **مطالعه DNCB :** بیماران مرض هادکین، مورد حساسیت قرار گرفته و سپس بـه طـور مسـتمر در معرض شیمی درمانی دینی ترو کلروبنزن (DNCB) قرار گرفته اند. جامعه (+) شامل آن بیمارانی است که واکنش مثبـت بـه DNCB نشـان داده و جامعـه (-) شـامل آن بیمـارانی اسـت کـه واکـنش نشـان نـداده انـد. بیماران می توانند در میان جمعیت ها نقل مکان کنند، زمان بقاء را تا زمان جایگذاری در نظر می گیریم.

آیا بیماران جامعه (+) بیش از جامعه بیماران (-) عمر می کنند؟ مدل نرخ شکسـت متناسـب کـاکس بـه کار می رود. رسم $\log \hat{S}^{(+)} / \log \hat{S}^{(-)}$ در نمودار زیر نشان می دهد کـه بـرای زمـان t نزدیـک بـه صـفر، نسبت لگاریتم به طور مستدلی ثابت است، که نشان دهنده برقراری مدل است (گنگ، 1980).



نمودار 1-3 . نمایش $\log \hat{S}^{(+)}(t) \Big/ \log \hat{S}^{(-)}(t)$ **توسط رایانه**

4 فصل چهارم

تحلیل مدل رگرسیون کاکس در نرم افزار s-plus

4-1 مقدمه

یکی ازکاربردهای مهم مدل رگرسیون کاکس[1] استفاده ازآن در نرم افزارهای آماری می باشد. به طوریکه پس ازشناخت کامل این مدل به صورت تئوری در فصل گذشته، نیاز به نرم افزار خاصی داریم تا بتوانیم بحث های مربوط به این مدل را که صرفا به صورت تحقیقی مورد بررسی قرارگرفت را در قالب مثال های کاربردی وارد آن کنیم و نتیجه ای عملی از این بحث را استخراج کنیم. +S یکی از این نرم افزارها است که می تواند نیازهای کاربردی این مدل را با استفاده ازتوابع خاص و مورد نظر این بحث، برطرف سازد. با استفاده از این برنامه آماری می توان فنون و تکنیک های محاسباتی را وارد سیستم کاربردی نموده و مزایای آن را عینا در رابطه با حل مشکلات اجرایی مربوط به این بحث، مشاهده کنیم. استفاده از این نرم افزار نسبتا ساده بوده و نیاز به نوشتن برنامه های طولانی و پیچیده ندارد. در این راستا می توان محاسبات پیچیده و زمان بر را به سادگی و با دقت زیاد انجام داد.

در این فصل با ذکر یک مثال روش تجزیه و تحلیل مدل رگرسیون کاکس با استفاده از این برنامه آماری شرح داده می شود. به این ترتیب که با برازش مدل به داده های مورد نظر به متغیر پاسخ دست می یابیم و از آن برای پیش بینی لازم استفاده می کنیم. سپس با روش های گرافیکی به آزمون نمودارهای باقیمانده ها می پردازیم و با تقسیم داده ها به دو گروه، منحنی بقای آنها را بررسی می کنیم. این روش ما را قادر مي سازد که وجود تفاوت معني دار بين دو گروه را با ذکر مقدار P-value گزارش کنيم.

برای تجزیه و تحلیل مدل رگرسیون کاکس نیاز به توابع خاصی داریم که در این قسمت یکی از این توابع را معرفی می کنیم:

4-2 تابع coxreg

مدل رگرسیون کاکس را می توان به وسیله تابع ( ) coxreg به داده ها برازش[2] داد. شناسه مورد نیاز این تابع یک بردار شامل زمان های بقاء، یک بردار شامل اطلاع در مورد موقعیت و وضعیت متغیر مورد نظر و یک بردار یا یک ماتریس شامل مقادیر متغیرهای توضیحی می باشد (نواب پور و همکاران، 1382).

4-2-1 شناسه های لازم

شناسه time : این شناسه بردار مقادیرزمان می باشد. تمام مقادیری که این شناسه می پذیرد باید بزرگتر یا مساوی صفرباشند.

شناسه status : این شناسه به صورت بردار بوده و نشان دهنده موقعیت و وضعیت اختصاصی متغیر مورد نظر می باشد. به عنوان مثال می توان مقادیر صفر و یک را اینگونه در نظر بگیریم که صفر

[1] Cox regression model
[2] Fit

نشان دهنده مقادیر سانسورشده و یک نشان دهنده مقادیر سانسور نشده باشد. هم چنین این مقادیرمی توانند 1 و 2 باشند که 1 از همه مقادیر تفریق شده باشد. طول این بردار باید با طول بردار time یکسان باشد.

شناسه x : این شناسه بردار یا ماتریسی از متغیرهای توضیحی می باشد. سطرها مشاهدات را نشان می دهند و ستون ها متغیرها را.

4-2-2 شناسه های اختیاری

شناسه strata : برداری عددی یا داده هایی طبقه بندی شده از هر یک از مشاهدات مربوط به آنالیز طبقه ای داده ها می باشد. اما قرارداد این است که آن را طبقه فرض کنیم. همچنین طول این بردار باید با طول بردار time یکسان باشد.

شناسه resid : از این شناسه برای انجام محاسبات روی باقیمانده های مدل استفاده می شود که شامل انواع مختلفی هستند؛ همچون **باقیمانده انحراف**[1] و **باقیمانده مارتینگل**[2].

این تابع شناسه های اختیاری دیگری نیز دارد که کاربردهای ویژه خودشان را دارند و در مباحث مختلف استفاده می شوند. مانند شناسه های init ، man.iter ، eps ، table و ratio.Inf .

4-3 مثال

برای شرح بیشتر تحلیل داده های زیستی با استفاده از مدل رگرسیون کاکس از داده های **Error! Reference source not found.** استفاده می کنیم.این داده ها مربوط به بیماران سرطانی است. در اینجا به زمان بقای بیماران توجه بیشتری می شود و سوال اصلی این است که آیا می توان نتیجه مراقبت از بیماران را پیش گویی کرد؟

فرض کنید که داده های **Error! Reference source not found.** در یک فایل متنی به نام cancer.dat ذخیره شده است (نواب پور و همکاران، 1382).

4-3-1 برازش مدل کاکس

همانطور که قبلا اشاره شد برای برازش مدل کاکس به داده ها از تابع ( ) coxreg استفاده می شود. با استفاده از عبارت های زیر مدلی شامل هر شش متغیر چاچوب اطلاعاتی cancer .data برازنده می شود:

```
> Time <- cancer.dat [,"Time"] ↵

> Status <- 1- cancer.dat [,"Status"] ↵
```

[1] Deviance residual
[2] Martingel Residual

```
> cancer.cox <- coxreg (Time, Status, cancer.dat [, c("Age", ↲
+ "Smear","Infill","Index","Blasts","Temp")]) ↲
> cancer.cox ↲
```

```
 Alive Dead Deleted
     6   45       0

           coef exp(coef) se(coef)      z       p
   Age  0.03200     1.033  0.01027  3.114 0.00184
 Smear  0.00918     1.009  0.01417  0.648 0.51730
Infill -0.01686     0.983  0.01246 -1.353 0.17620
 Index -0.06923     0.933  0.03841 -1.802 0.07148
Blasts  0.00238     1.002  0.00552  0.432 0.66590
  Temp  0.01463     1.015  0.01098  1.333 0.18266

       exp(coef) exp(-coef) lower .95 upper .95
   Age     1.033      0.969     1.012      1.05
 Smear     1.009      0.991     0.982      1.04
Infill     0.983      1.017     0.960      1.01
 Index     0.933      1.072     0.865      1.01
Blasts     1.002      0.998     0.992      1.01
  Temp     1.015      0.985     0.993      1.04

Likelihood ratio test= 17.6  on 6 df,   p=0.0072
Efficient score test = 17.8  on 6 df,   p=0.00668
```

تنها متغیر مهمی که در خروجی بالا برای پیش بینی بیشتر به چشم می آید، متغیر سن است. بنابراین به وسیله عبارت زیر مدلی شامل متغیر سن برازنده می شود:

```
> coxreg (Time, Status, cancer.dat[,"Age"]) ↲
```

```
 Alive Dead Deleted
     6   45       0

       coef exp(coef) se(coef)    z       p
[1,] 0.0311      1.03  0.00955 3.25 0.00115

     exp(coef) exp(-coef) lower .95 upper .95
[1,]      1.03      0.969      1.01      1.05

Likelihood ratio test= 10.8  on 1 df,   p=0.00101
Efficient score test = 11.2  on 1 df,   p=0.000839
```

تفسیر این مدل همانند تفسیر برآورد ضرائب استاندارد نشده در رگرسیون چند متغیره است. در اینجا نتیجه می گیریم که افزایش به میزان یک سال در سال های زندگی، لگاریتم تابع مخاطره را به میزان 0/0311 افزایش می دهد.

4-3-2 تحلیل گرافیکی باقیمانده ها

برای اطمینان از برازش مناسب مدل، نیازمند آن هستیم که ملاکی برای بهتر بودن مدل برازش شده داشته باشیم. همان طور که در مدل رگرسیون خطی در نهایت این باقیمانده ها هستند که به ما کمک می کنند تا مدل را ارزیابی کنیم، در آنالیز بقاء نیز می توانیم به کمک باقیمانده ها مدل را مورد ارزیابی قرار دهیم.

در آنالیز بقاء مانند مدل رگرسیون خطی با انواع مختلف باقیمانده ها سر و کار داریم که هر یک برای منظور خاصی مورد استفاده قرار می گیرند.

هر چند در استفاده از باقیمانده ها نظرات متفاوت است و بستگی به نوع اطلاعات ، تجربه و نظر محقق دارد ولی به طور مرسوم، باقیماندهای مارتینگل در ارزیابی شکل تابعی که در برگیرنده متغیرهای کمکی است و همچنین در سنجش عدم توانایی مدل در برازش مناسب به داده ها مورد استفاده قرار می گیرد.

بنابراین پس از برازش مدل کاکس به داده ها لازم است با انجام تحلیل گرافیکی باقیمانده ها، مناسب بودن این مدل مورد بررسی قرار گیرد. برای این منظور از نمودارهای متفاوتی استفاده می شود که در اینجا از نمودار باقیمانده ها در مقابل مقادیر پیش بینی شده متغیر پاسخ استفاده شده است.

این نمودارها برای یافتن نقاط دور افتاده[1] ، نقاط مؤثر[2] و ... مفید است. برای رسم این نمودارها از دو نوع باقیمانده استفاده می شود. همانطورکه در قسمت معرفی تابع ( ) coxreg اشاره شد، این باقیمانده ها

[1] Outliers
[2] Influential points

عبارتند از :

- **باقیمانده مارتینگل** : که برای یافتن شکل تابعی صحیح برای متغیر کمکی مفید است.
- **باقیمانده انحراف** : که برای تشخیص پیش بینی های ضعیف مناسب است.

4-3-3 باقیمانده های مارتینگل

با قیمانده های مارتینگل را به صورت زیر تعریف می کنند:

$$\widehat{M}_i = N_i(t) - \widehat{E}_i(t)$$

که در آن ؛

$N_i(t)$ تعداد پیشامد هایی است که تا زمان t برای عنصر iام رخ داده است و $\widehat{E}_i(t)$ هم خطر تجمعی برآورد شده برای عنصر iام تا زمان t می باشد. اختلاف بین $N_i(t)$ و $\widehat{E}_i(t)$ ، درحقیقت اختلاف واقعی بین تعداد مشاهده شده و تعداد مورد انتظار پیشامدها می باشد (حسینی و همکاران، 1390).

اگر مدلی که به داده های زمان بقاء برازنده شده است درست باشد شکل نمودار باقیمانده ها یک الگو و روند مشخص را نمایش نمی دهد و نقاط روی نمودار کاملا به صورت تصادفی پراکنده خواهند بود.

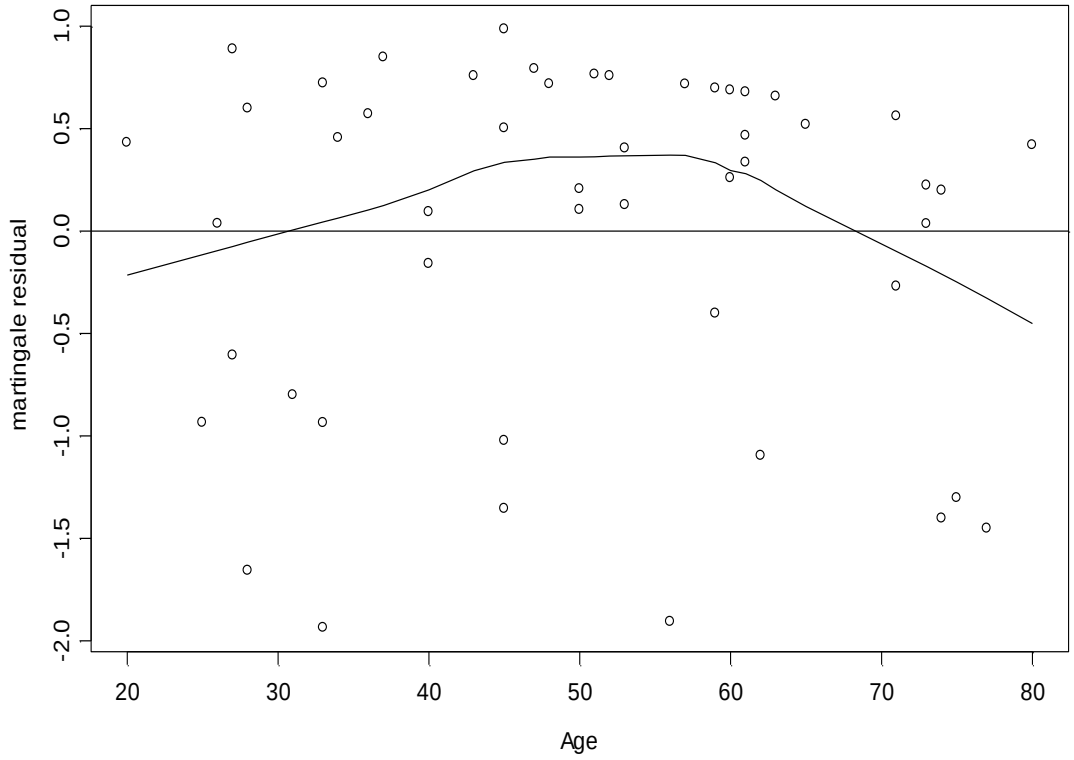
ابتدا با استفاده از عبارت های زیر نمودار باقیمانده های مارتینگل در مقابل سن رسم می شود:

```
> cancer.cox1<- coxreg (Time, Status, ↲
+ cancer.dat [,"Age"], resid = "mart") ↲
> plot (cancer.dat [,"Age"], cancer.cox1 $ resid,↲
+ xlab ="Age", ylab ="martingal residual") ↲
> lines (lowees (cancer.dat [,"Age"], cancer.cox1 $ resid))↲
> abline (h=0)↲
```

با اضافه کردن خط رگرسیون موضعی موزون[1] به نمودار باقیمانده ها، خطی بودن مدل بررسی می شود. در عبارت های بالا از تابع ( ) lowess برای برازش خط رگرسیون موزون موضعی استفاده شده

[1] Local weighted

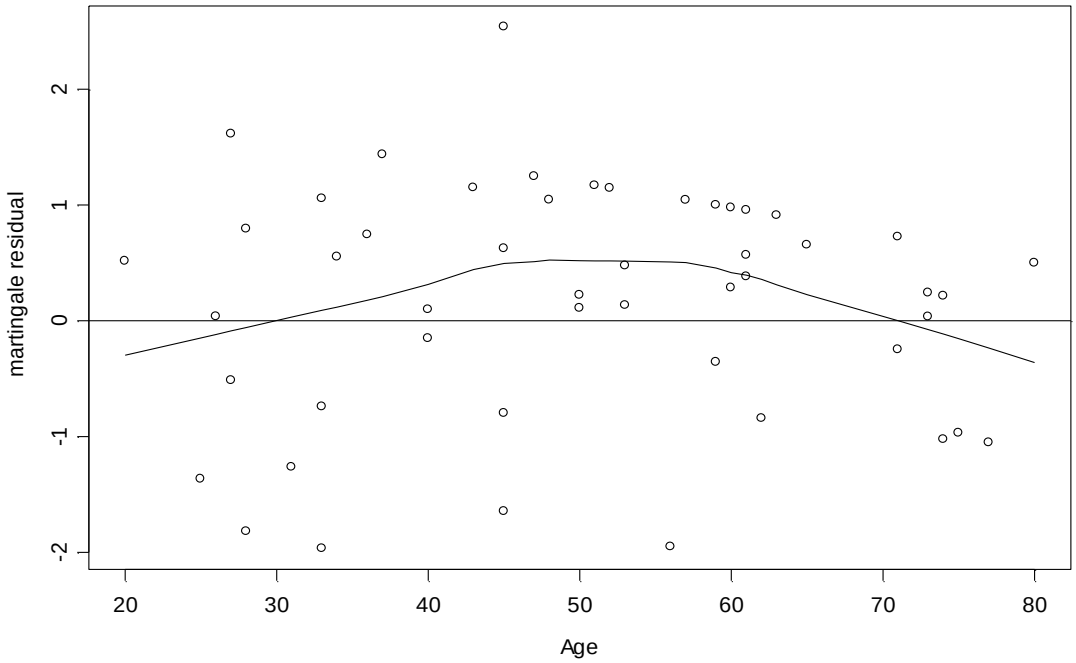
است. خط رگرسیون موضعی برازنده شده به داده ها بیانگر این است که مناسب تر است به جای متغیر سن، یک فرم درجه دوم از سن به مدل اضافه شود.



نمودار 1-4 . نمودار سن در مقابل باقیمانده های مارتینگل به همراه خط رگرسیون موزون موضعی

نمودار سن در مقابل باقیمانده های انحراف با استفاده از عبارت های بالا رسم می شود. با این تفاوت که این بار در تابع ( ) coxreg از شناسه "resid = "dev استفاده می شود. در نمودار 2-4 مشاهده 22 غیر طبیعی به نظر می رسد. برای رسم این نمودار داریم:

```
> cancer.cox1<- coxreg (Time, Status, ↲
+ cancer.dat [,"Age"], resid = "dev") ↲
> plot (cancer.dat [,"Age"], cancer.cox1 $ resid, ↲
+ xlab ="Age", ylab ="martingal residual") ↲
> lines (lowees (cancer.dat [,"Age"], cancer.cox1 $ resid)) ↲
> abline (h=0) ↲
```



نمودار 2-4 . نمودار سن در مقابل باقیمانده های انحراف

4-3-4 تفسیر منحنی بقای مدل کاکس

تابع ( ) surv.fit : این تابع برآوردهای کاپلان – میر منحنی بقاء را محاسبه می کند. ازجمله شناسه های لازم آن time و status می باشد که در ابتدای این فصل تعریف شده است. هم چنین این تابع دارای شناسه های اختیاری strata ، na.strata ، type و ... می باشد.

اکنون داده های مثال قبل را گروه بندی می کنیم و برای هر گروه منحنی بقای جداگانه را رسم و سپس تابع های زمان بقای حاصل از هر کدام را مورد بررسی قرار می دهیم:

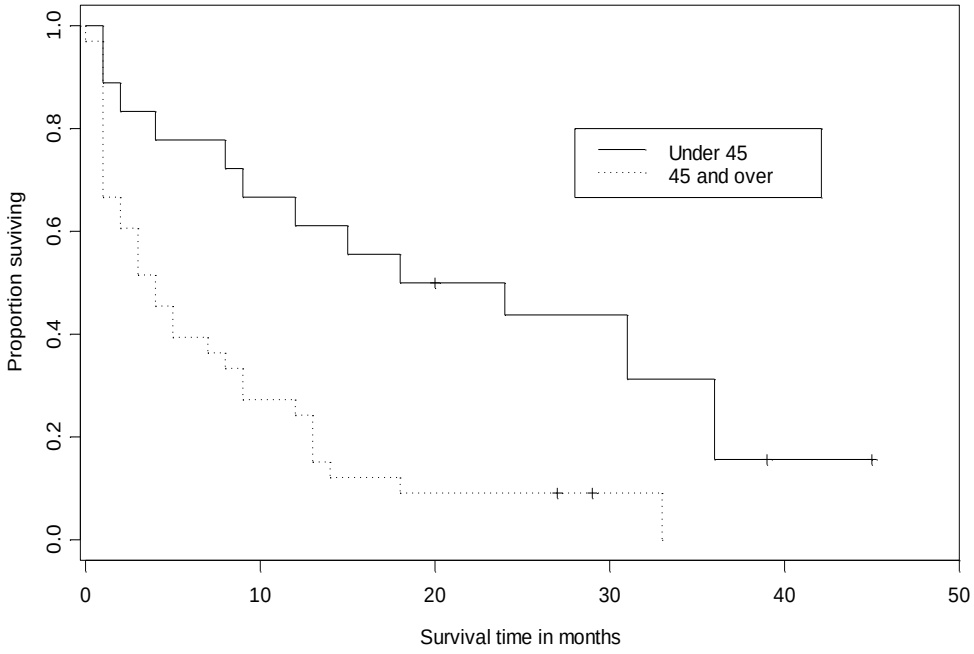
عبارت های زیر بیماران را به دو گروه بالای 45 سال و زیر 45 سال تقسیم کرده و منحنی بقای مربوط به هر گروه از بیماران را رسم می کند :

```
> agroup <- cancer.dat [,"Age"] – 45↵
> agroup [agroup >=0] <- 1↵
> agroup [agroup <0] <- 0↵
> plot (surv.fit (Time, Status, agroup),↵
+ xlab = "survival time in months",↵
```

+ ylab = "proportion surviving", lty = 1:2)↵

> legend (28,0.8, c("under 45","45 and over"),↵

+ lty = 1;2)↵



نمودار 3-4 . منحنی های بقاء برای دو گروه بیماران بالای 45 سال و زیر 45 سال

نمودار 3-4 نشان دهنده اختلاف زیادی در تابع بقای دو گروه سنی است. برای انعکاس تفاوت این دو منحنی به آزمون اختلاف تابع بقاء می پردازیم.

آزمون اختلاف با استفاده از آزمون لگ- رتبه[1]:

برای آزمون اختلاف تابع بقاء دو گروه می توان از تابع ( ) surv.diff استفاده کرد. این تابع، آزمون اختلاف را با استفاده از آزمون لگ – رتبه (که توسط پتو و پتو (1972) ارائه شد) انجام می دهد.

> Surv.diff(Time, Status, agroup)↵

[1] Log – rank test

```
   N Observed Expected (O-E)^2/E
0 18       14    23.82     4.049
1 33       31    21.18     4.553

 Chisq= 11  on 1 degrees of freedom, p= 0.0009261
```

مقدار کـی – دو حاصـل از اجـرای عبـارت بـالا برابـر 11/00 و درجـه آزادی آن 1 اسـت. مقـدار p-value متناظر برابر 0/00093 است. بنابراین تفاوت آشکاری بین منحنی های بقای دو گروه وجود دارد.

اگر رگرسیون کاکس با استفاده از متغیر توضیحی[1] agroup بـه کـار رود آنگـاه فاصـله اطمینـان 95٪ برای ضریب به صـورت نمـایی برابـر بـا (5/95 و 1/46) اسـت. بنـابراین بـرای افـراد بـالای 45 سـال تـابع مخاطره بین 46٪ تا 95٪ افزایش می یابد (نواب پور و همکاران، 1382).

> coxreg(Time, Status, agroup)↵

```
  Alive Dead Deleted
      6   45       0

      coef exp(coef) se(coef)    z       p
[1,] 1.08      2.95    0.358 3.02 0.00251

      exp(coef) exp(-coef) lower .95 upper .95
[1,]      2.95      0.339      1.46      5.95

Likelihood ratio test= 10.2  on 1 df,   p=0.00139
Efficient score test = 9.83  on 1 df,   p=0.00172
```

9	8	7	6	5	4	3	2	**1**
0	18	1	990	0/6	7	39	78	20
1	31	1	1030	35	16	61	64	25
0	31	1	982	7/5	12	55	61	26

[1] Explanatory variable

0	31	1	1000	21/0	16	64	64	26
0	36	1	980	7/5	6	95	95	27
0	1	0	1010	0/6	8	64	80	27
0	9	1	986	4/8	20	88	88	28
1	39	1	1010	10/0	14	70	70	28
1	20	1	988	2/5	5	72	72	31
0	4	0	986	5/7	7	58	58	33
1	45	1	980	2/6	5	92	92	33
0	36	1	984	2/5	12	38	42	33
0	12	0	982	7/0	7	26	26	34
0	8	1	986	4/5	14	55	55	36
0	1	0	1020	4/4	15	71	71	37
0	15	1	986	35/0	9	91	91	40
0	24	1	988	2/1	12	49	52	40
0	2	0	986	0/1	4	63	74	43
0	33	1	980	4/2	14	47	78	45
1	29	1	992	0/6	10	36	60	45
0	7	0	1016	28/1	10	32	82	45
0	0	0	1030	1/1	4	79	79	45
0	1	0	990	0/9	2	28	56	47
0	2	0	1002	2/2	10	54	60	48
0	12	1	996	11/6	19	66	83	50
0	9	1	992	4/5	14	32	36	50
0	1	0	982	0/5	8	70	88	51
0	1	0	986	10/3	7	87	87	52

9	8	7	6	5	4	3	2	**1**
0	9	1	980	2/3	13	68	75	53
0	5	0	982	2/3	6	65	65	53
0	27	1	992	16/0	10	92	97	56
1	1	0	1020	21/6	19	83	87	57

0	13	0	999	1/1	8	45	45	59
0	1	0	1038	0/0	5	34	36	59
0	5	0	988	0/9	7	33	39	60
0	1	0	982	0/4	12	53	76	60
0	3	0	1006	1/4	4	37	46	61
0	4	0	990	0/3	8	8	39	61
0	1	0	990	9/9	1	90	90	61
0	18	1	1020	11/5	19	84	84	62
0	1	0	1014	0/3	5	27	42	63
0	2	0	1004	20/0	10	75	75	65
0	1	0	990	0/3	6	22	44	71
0	8	1	986	10/0	11	63	63	71
0	3	0	1010	0/5	4	33	33	73
0	4	0	1020	38/0	6	84	93	73
0	14	1	1002	2/4	10	58	58	74
0	3	0	988	6/7	16	30	32	74
0	13	1	990	8/2	17	60	60	75
0	13	1	986	1/5	9	69	69	77
0	1	0	986	1/5	7	73	73	80

منابع

1- *اسفندیاری، هادی. اسلمی نژاد، علی اصغر.طهمورث پور، مجتبی. 1389. « کاربرد روش های آماری تحلیل بقاء در آنالیز طول عمر گاوهای شیری ». دهمین کنفرانس آمار ایران. تبریز. ایران.*

2- *یوسف نژاد، کیوان. شعبانی، بیژن و ... . بهمن 1390. « تحلیل داده های بقاء در بیماران قلبی عروقی پس از آنژیوپلاستی به کمک مدل رگرسیون کاکس ». مجله دانشگاه علوم پزشکی مازندران. دوره بیست و یکم. شماره 86 . ص 106-101.*

3- *رجائی فرد، عبدالرضا. مقیمی دهکردی، بیژن و ... . تابستان 1388. « کاربرد مدل های پارامتری*

در تحلیل بقاء در سرطان معده ». فصلنامه علمی پژوهشی فیض. دوره سیزدهم. شماره 2.

4- حسینی تشنیزی، سعید. زارع، شهرام. تذهیبی، مهدی. زمستان 1390. « ارزیابی مدل های خطرات متناسب کاکس و وایبل و کاربرد آنها در شناسایی عوامل مؤثر ب زمان بقای بیماران لوسمی حاد ». مجله پزشکی هرمزگان. سال پانزدهم. شماره 4. ص 278-269.

5- علی حسینی، فرشته (دانشجوی کارشناسی ارشد آمار اقتصادی اجتماعی). « توزیع وایبل تعمیم یافته با نرخ مخاطره U شکل ». دانشکده علوم ریاضی دانشگاه صنعتی اصفهان.....عنوان مجله یا پایان نامه.

6- صانعی، سید حسن. 1380. « تجزیه و تحلیل داده های بقاء ». انتشارات اندیشمند. تهران.

7- روشنی، دائم. آزادی، نمامعلی و همکاران (1390). « کاربرد رویکردهای پارامتری، نیمه پارامتری و ناپارامتری در تحلیل بقاء بیماران مبتلا به سکته قلبی ». مجله دانشگاه علوم پزشکی خراسان شمالی. دوره سوم.

8- صادقی، سکینه. کاظمی، ایرج. تابستان 1387. « بررسی مدل های مختلف برای داده های سانسورشده با اثرات تصادفی ». نهمین کنفرانس آمار ایران.

9- حسینی، مصطفی. محمد، کاظم و ... . بهار 1386. « مقایسه مدل های مختلف بقاء در مطالعه طول مدت شیردهی ». مجله پژوهشی حکیم. دوره هشتم. شماره 1.

10- قدیمی، محمود رضا. محمودی، محمود و ... . تابستان 1389. « تحلیل بقای بیماران مبتلا به سرطان دستگاه گوارش و مقایسه آن در مدل های پارامتری و مدل کاکس ». مجله دانشگاه بهداشت و انستیتو تحقیقات بهداشتی. دوره هشتم. شماره 2. ص 14-1.

11- نواب پور، حمید رضا. یزدانی، افسانه. ایزدی، غلامرضا. « محاسبات آماری با کامپیوتر ». انتشارات تهران، دانشگاه پیام نور. 1382.

12- *Pisani P, Bray F, Parkin D.M. Estimatores of the world-wide prevalence of cacer for 25 sites in the Adult Pupolation. Int J cancer 2002; 97: 72-81.*

13- *Leavaitt and Olshen, unpoblished report (1974), give the insurance example.*

14- *Kllein JP, Moeschberger ML. Survival analysis: thechniques for censored and truncated data. New York: Springer-Verlag Press; 1997.*

15- *Dawson B, Trapp R. Basic and Clinical Biostutistics. England, Appleton and Lange; 2001.*

16- *Millear, R.G .Survival analysis. Translated by Abolghasem Bozorgnia – Hojjat Rezaee Pazhand. Ferdowsi University Press. 2001.*

17- *Alexander M. Mood – Franklin A. Graybill Duanec. Hoes. 1913. Introduction to theory of statistics. Translated by Dr.Ali Meshkani. Ferdowsi University*

Press. Publication No, 235.

18- *Therneau, T. Grambsch, P. 2000. Modeling survival data. New York Univ. Wisconsin Tech Report No, (1978).*

19- *Wilk and Gnanadesikan. Biometriku. 1968.*

20- *Nelson, J. Qual. Tech. 1969.*

21- *Gong. Stanford Univ. Report No, 57. 1980.*

واژه نامه

Actuarial method	روش بیمه گری
Breslow test	آزمون برسلو
Basis hazard function	تابع مخاطره پایه
Censored	سانسور شده
Cumulative	تجمعی
Cumulative hazard function	تابع خطر تجمعی
Censoring	سانسور (برش)
Conditional probbility	احتمال شرطی
Cohort life table	جدول طول عمر گروهی
Corrent life table	جدول طول عمر جاری
Cox regression model	مدل رگرسیون کاکس
Conditional likelihood	درست نمایی شرطی
Cumulative hazard logarithm	لگاریتم خطر تجمعی
Conditional likelihood analysis	تحلیل درست نمایی شرطی
Density function	تابع چگالی
Deviance residual	باقیمانده انحراف

Epidemiology	اپیدومیولوژی
Event rate	نرخ رویداد
Expected future lifetime	طول عمرآینده مورد انتظار
Explanatory variable	متغیر توضیحی
Exponential distribution	توزیع نمایی
Future lifetime	طول عمر آینده
Full likelihood	درست نمایی کامل
Gehan test	آزمون گهان
Gamma distribution	توزیع گاما
Goodness of fit	نیکویی برازش
Hazard function	تابع مخاطره
Hazard rate	نرخ شکست
Interval censoring	برش فاصله ای
Influential points	نقاط مؤثر
Kaplan – meier	کاپلان – میر
Lifetime table	جدول طول عمر
Log – rank test	آزمون لگ – رتبه
Lifetime distribution	توزیع طول عمر
Left censoring	برش چپ
Log – normal distribution	توزیع لگ – نرمال
Local weighted	موزون موضعی
Mean time to failure	زمان متوسط تا شکست
Mean residual lifetime	طول عمر باقیمانده متوسط
Mean	میانگین
Maximum likelihood (ML)	درست نمایی ماکسیمم
Maximum likelihood estimator (MLE)	برآوردگر درست نمایی ماکسیمم
Method of scoring	روش امتیازی (چوب خطی)
Martingale residual	باقیمانده های مارتینگل
Newton – raphson method	روش نیوتون – رافسون
Non – parametric method	روش ناپارامتری
Outliers points	نقاط دور افتاده

Probability density function	تابع مولد گشتاور
Partial likelihood	درست نمایی جزئی
Parametric method	روش پارامتری
Rrliability function	تابع قابلیت اطمینان
Right censoring	برش راست
Random variable	متغیر تصادفی
Reduced sample method	روش کاهش نمونه
Survival analysis	آنالیز بقاء
Survival model	مدل بقاء
Survival curve	منحنی بقاء
Step plot	نمودار پلکانی
Survival function	تابع بقاء
Sorvivor function	تابع بازمانده
Score residual	باقیمانده اسکور
Schoenfeld residual	باقیمانده اسکنفیلد
Score vector	بردار امتیاز
Sample information matrix	ماتریس اطلاعات نمونه
Survival data	داده های بقاء
Truncation	قطع مشاهدات
Tentativ estimate	برآورد تجربی
Variance	واریانس
Weibull distribution	توزیع وایبل