موضوع: آنالیز بقا

سمینار درس مدل های خطی (۱)

ارائه دهندگان:

نگار دهقان- بهاره اسدی

پاییز ۹۸

فهرست

تحلیل بقا

در بسیاری از پژوهشهای آماری رفتار متغیرهایی تحت عنوان مدت زمان، مانند زمان ابتلا به بیماری تا مرگ بیمار، مدت زمان دورهٔ نقاهت بیماری، مدت تصویب لایحه ای در مجلس، مدت زمان تدوین آئین نامه قوانین، طول عمر یک قطعه و مدت زمان جذب یک دانش آموخته دانشگاه به بازارکار بررسی میشود.
در اینگونه پژوهشها به ویژه در بررسیهای پزشکی، هدف یافتن و مدلبندی منابع تأثیرگذار بر مدت زمان بررسی شده است تا با تغییر منابع خطر، مدت زمان مذکورکنترل شود

. اینگونه دادههای برحسب زمان، دادههای بقا وروشهای تحلیل آنها تحلیل بقاء نامیده میشوند . در تحلیل دادههای بقاء به علت وجود سانسور و چولگی
مدلهای معمول نمی توانند مورد استفاده قرار گیرند . در روشهای تحلیل بقاء مرسومترین مدل در برازش به دادههای بقا مدل کاکس میباشد که بعدا بطور مفصل راجع به این موضوع بحث خواهیم کرد.

آنالیز بقاء ، تحلیل ماندگاری، تجزیه بقاء یا تحلیل بقاء یکی از مباحث علم آمار است که در رشته های مختلفی ازجمله علوم کامپیوتر، اپیدومیولوژی و کشاورزی کاربرد دارد .تحلیل بقاء به مجموعهای از روشهای آماری تحلیل داده گفته میشود که در آنها متغیر مطلوب ، زمان وقوع یک پدیده است. این موضوع در علوم مهندسی نظریه قابلیت اطمینان نامیده میشود.

ویژگی خاص تحلیل بقاء این است که با داده های سانسور شده وفق داشته و از این رو از اطلاعات دادههایی که در زمان ارزیابی هنوز زنده هستند استفاده مینماید. تحلیل بقاء، به عنوان یک روش آماری که اساساً برای تحقیقات زیستی و مهندسی یافته میتواند در آنالیز دادههای طول عمر مورد استفاده قرار گیرد. این روش آماری اطلاعات حاصل از دادههای حذف شده (سانسور شده) و حذف نشده (سانسور نشده) را با یکدیگر ترکیب نموده و تحلیل آماری دادههای سانسور شده را امکانپذیر ساخته و از سویی دیگر خصوصیت غیر خطی دادههای طول عمر را مورد توجه قرار میدهد

مفاهیم تحلیل بقا

تابع چگالی احتمال را با f نشان داده.یکی از موارد مهم استفاده از تابع چگالی احتمال برای تابع توزیع تجمعی است که بصورت زیر نمایش میدهیم.

$$F(x) = P(X < x) = \int_0^X f(s)ds$$

تابع بقاء به صورت زیر بدست می آید

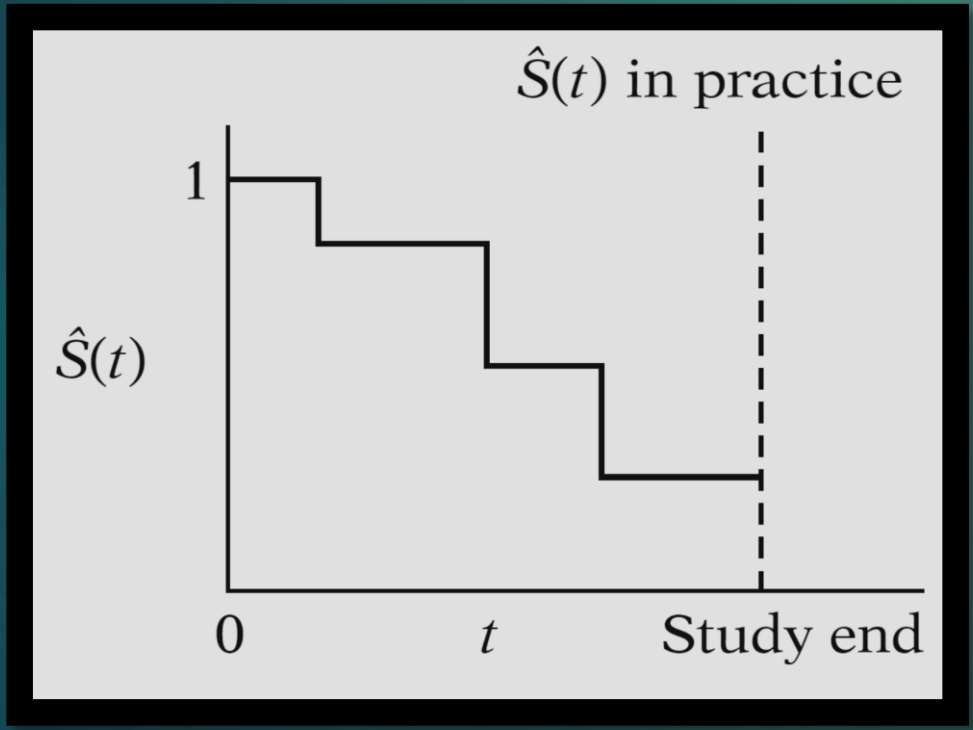
$$S(x) = 1 - F(x) = P(X \geq x) = \int_X^{\infty} f(s)ds$$

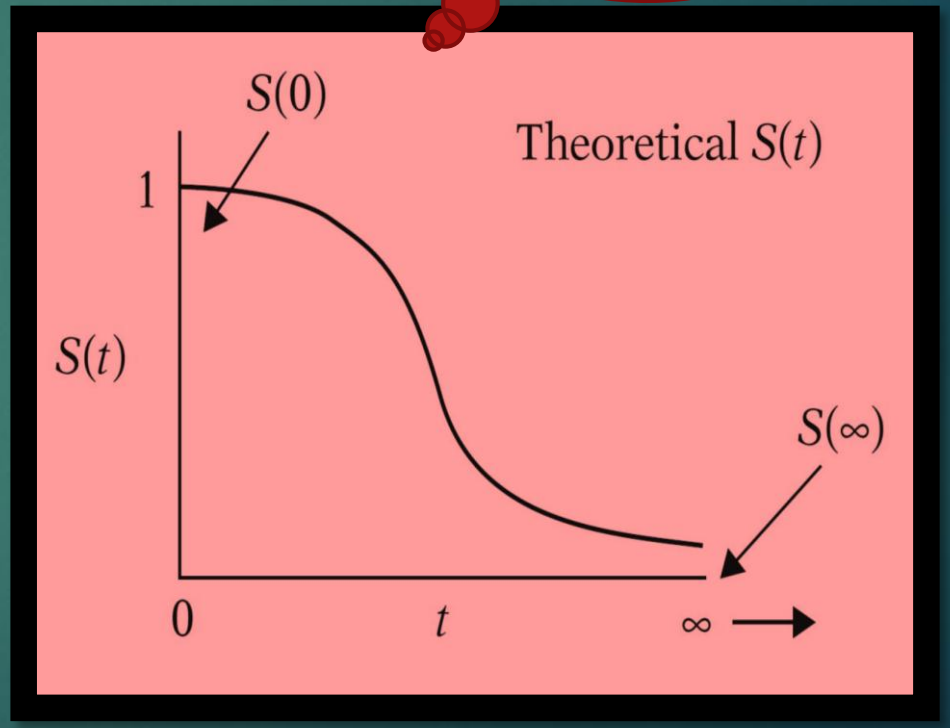
مهمترین موضوع در بقاء ،تابع خطر است که بصورت زیر تعریف می کنیم

$$\lambda(X) = \lim_{\Delta \to \infty} \frac{P(x < X < x + \Delta | X \geq x)}{\Delta} = \frac{f(x)}{1 - F(X)}$$

منحنی تابع بقا

ولی در عمل وقتی از داده ها ی واقعی استفاده می کنیم شکل تابع بقا به صورت پله ای خواهد بود





در مواقعی که در دادههای بقا از یک توزیع معلوم با تکیه گاه مثبت مانند وایبل پیروی میکنند ، برای تحلیل داده های بقا می توان از مدلهای پارامتری بهره برد ، که در آنها رابطه بین تابع بقاء و تابع خطر با تابع چگالی زمان های بقاء به صورت:

$$f(t)=h(t).S(t)$$

تابع نرخ خطر می تواند صعودی ، نزولی یا ثابت باشد

- یک تابع نرخ خطر صعودی در زمان † نشان می دهد که احتمال خرابی مولفه در افزایش زمان بعدی بیشتر از خرابی آن در زمان حاضر است یعنی با گذشت زمان مولفه فرسوده می شود.به عنوان مثال لاستیک اتومبیل ها دارای نرخ خطر صعودی است.
- همچنین یک تابع نرخ خطر نزولی بدین معنا است که مولفه با گذشت زمان رو به بهبود است.به عنوان مثال قالی کرمان که با گذشت زمان ارزش آن بیشتر می شود.
- تابع نرخ خطر ثابت ، که برای توزیع نمایی است و ویژگی بدون حافظه بودن این توزیع را بیان می کند.
- باید توجه داشت که تابع خطر را باید همواره برای یک فرد مطالعه در نظر گرفت،نه برای یک جامعه و یا یک نمونه چون تابع خطر هرگز مستقیما مشاهده نمی شود و یک کمیت غیر قابل مشاهده است همواره می توان برآوردی از آن را داشت.

سانسور کردن

در آمار، مهندسی، اقتصاد و تحقیقات پزشکی، منظور از «سانسور کردن» (Censoring)، ثبت و اندازه‌گیری بخشی از اطلاعات مربوط به مشاهدات یا متغیرها است. برای مثال فرض کنید که قرار است اثر یک دارو روی نرخ مرگ و میر اندازه‌گیری شود. گفتنی است که این دارو به گروهی از افراد داده شده و می‌دانیم که یکی از آنها در سن ۷۵ سالگی از بررسی‌های پزشکی انصراف داده است. اگر این فرد از داده‌های آزمایشگاهی خارج شود اطلاعاتی که توسط او تولید شده، از بین می‌رود. از آنجایی که می‌دانیم که در هنگام خروج از آزمایش پزشکی ۷۵ ساله بوده، می‌توان این اطلاع را کسب کرد که سن مرگ او با توجه به مصرف دارو بیشتر از ۷۵ سال است. استفاده از داده های سانسور شده و اطلاعات حاصل از آن‌ها در استنباط آماری و یا برآورد پارامترهای مربوط به متوسط سن فوت برای این گونه افراد باعث افزایش دقت برآوردها خواهد شد. از داده های سانسور شده بیشتر برای بررسی طول عمر در مباحث «قابلیت اعتماد» (Reliability) استفاده می‌شود؛ این مباحث به بررسی زمان خرابی یا طول عمر قطعات و دستگاه‌ها می‌پردازند.

تابع درستنمایی برای داده های سانسورشده

تابع درستنمایی را می توان بر حسب تابع چگالی f(t) و تابع بقا s(t) برای T به صورت زیر نوشت:

مشاهدات بدون سانسور

$$l(b) = \log L = \sum \log(f(t)) + \sum \log(s(t)) + \sum \log(1 - s(t)) + \sum \log(s(v) - s(t))$$

مشاهدات سانسور فاصله ای مشاهدات سانسور چپ مشاهدات سانسور راست

احتمال بقا را می توان با دو روش زیر محاسبه و برآورد نموداین روش ها نا پارامتری است.

روش کاپلان-مه یر

روش طول عمر

کاپلان مه یر

یکی از روشهای تحلیل بقا در داده ها ازمشاهدات سانسور شده استفاده از براوردگر کاپلان میر است که مدل ناپارامتری است.

یکی از ویزگی های مدل ناپارامتری اینست که توانایی روبه رو شدن با توزیعی که هیچگونه پیش فرضی را داراست اما از نظر هزینه پرداخت گران است مدل ناپارامتری نیاز به تعداد زیادی داده دارند تا به نتیجه مطلوب برسند.و بدست اوردن پارامتر از تابع خطر غالبا اطلاعات مناسبی را داراست اما بسیار مشکل است.براوردگر کاپلان میر بصورت زیر میباشد

$$\hat{S}(t) = \prod_{i:T(i)<t} \left(1 - \frac{\Delta(i)}{n - i + 1}\right)$$

روش طول عمر

وقتی که تعداد مشاهدات و بیماران مورد بررسی زیاد باشند احتمال وجود زمان های بقای یکسان قوت می گیرد.بدین معنا که ممکن است بیش از یک حادثه در هر زمان رخ دهد.در این صورت روش KM جداول بسیار طولانی را موجب می شود که ارائه و تفسیر آن خیلی مطلوب نبوده و وقت گیر خواهد شد.

بنابراین روش دیگری را به نام روش طول عمر استفاده می کنند که در آن زمان وقوع حوادث را به صورت فواصل زمانی تقسیم بندی می کنند.این فاصله بندی به دلخواه می تواند کم و زیاد گردد و الزامی به تساوی فاصله گروه ها نمی باشد.هر چند مساوی بودن آن ها معمول و متداول است.این روش می تواند نمودار ها و برآوردهایی از توابع بقا و خطر را ارائه دهد.

مدل خطرهای متناسب

در سال ۱۹۷۲میلادی کاکس،آمار شناس معروف و معاصر انگلیسی مدلی را ارائه نمود که از آن زمان تا کنون بصورت استاندارد وگسترده ای مورد استفاده قرار می گیرد.وقتی که تحقیق بررسی اثرات چند متغیر بر روی زمان بقا به طور همزمان مد نظر باشد مدل کاکس کاربرد فراوانی دارد.

اما چرا این مدل نام «مدل خطرات متناسب» را به خود گرفته است؟

زیرا خطر هر فرد یک نسبت ثابت و مشخصی را با خطر هر فرد دیگردارد.

در روش های تحلیل بقا مرسوم ترین مدل در برازش به داده های بقا مدل کاکس است که مستلزم برقراری فرض استقلال داده های بقا است

یکی از ویژگی های مدل کاکس بدون در نظر گرفتن هیچ فرضی توزیعی درباره تابع خطر پایه میتوان آن را به داده های بقا برازش داد از طریق تابع خطر پایه مدل های پارامتری و نیمه پارامتری را به داده های بقا برازش داد

مدل خطر متناسب بصورت زیر است:

$$h_i(t)=h_0(t).\exp\{\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_k x_{ik}\}$$

مدل کاکس، یک مدل نیمه پارامتری است و به توزیع خاصی برای زمان های بقا نیاز ندارد. اما دو فرض بسیار حساس و محکم در این مدل وجود دارد.

اول این که اثر متغیرهای مختلف بر بقا و در طی زمان، یکنواخت و ثابت است.

دیگر این که این اثرات طبق قانون جمع پذیری به مقیاس خاصی افزوده می شوند. با استفاده از مدل خطرات متناسب می توان خطر تجمعی را نیز برآورد کرد. بدین ترتیب که خطر مرگ در فاصله زمانی ۰ تا t متوالیاً به هم افزوده شده تا خطر متناسب تجمعی $H(t)$ بدست آید.

$$H_i(t)= H_0(t).\exp\{ \beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_k x_{ik} \}$$

در این جا $H_0(t)$ تابع خطر تجمعی مبنا است و مانند $h_0(t)$ قابل محاسبه است و ارتباط تنگاتنگی با تابع بقاء، $s(t)$ دارد از روی این مدل می توان احتمال بقاء برای هر فرد در مطالعه را به صورت $S(t)=\exp\{- H(t)\}$ نیز برآورد نمود.

مدل زمان شکست شتابیده AFT

تحلیل بقا را می توان با استفاده از مدل زمان شکست شتابیده مورد تحلیل قرارداد.
در مدل زمان شکست شتابیده برای زمان بقا فرض می شود که لگاریتم زمان بقا T به صورت خطی با متغیر های کمکی در ارتباط است که به صورت زیر میتوان نشان داد:

$$\log T_i = a_0 + \sum_{j=1}^{p} a_j x_j + \sigma\varepsilon = \mu_i + \sigma\varepsilon$$

که x_j و j=1,2,...,p متغیر های کمکی ، a_j ضرایب رگرسیونی ، σ پارامتر مقیاس مجهول است.و ε عبارت خطا ، یک متغیر تصادفی با فرم معلوم تابع چگالی $g(\varepsilon,d)$ وتابع بقای $\sigma(\varepsilon,d)$ با پارامتر نامعلوم d می باشد.

یک حالت ساده در نظر میگیریم که فقط یک متغیر کمکی x با
دو حالت x=1 , x=0 داشته باشیم آنگاه :

$$\log t = a_0 + a_1 x + \sigma\varepsilon$$

اگر T_0 و T_1 نشان دهنده زمان بقا برای دو مولفه (فرد)با x=0و x=1 باشند آنگاه:

$$x = 0 \Rightarrow \log T_0 = a_0 + \sigma\varepsilon \Rightarrow T_0 = e^{a_0 + \sigma\varepsilon}$$

$$x = 1 \Rightarrow \log T_1 = a_0 + a_1 + \sigma\varepsilon \Rightarrow T_1 = e^{a_0 + a_1 + \sigma\varepsilon}$$

$$\frac{T_1}{T_0} = e^{a_1} \Rightarrow T_1 = e^{a_1} T_0$$

پس اگر

$a_1 > 0$ پس $T_1 > T_0$ خواهد بود

$a_1 < 0$ پس $T_1 < T_0$ خواهد بود

یعنی متغیر کمکی x زمان بقا(زمان شکست) را شتاب می دهد و یا کند میکندبه همین دلیل این مدل ها مدل های زمان شکست شتابیده نامیده می شوند.

بطور کلی دو مدل رگرسیونی برای بررسی داده های سانسور راست وجود دارد . مدل خطرات متناسب کاکس به عنوان یک مدل نیمه پارامتری و مدل های زمان شکست شتابیده به عنوان مدل های پارامتریک. بسیاری از مدل های استاندارد پارامتریک مانند وایبل نمایی و لگ نرمال در این گروه از مدل ها قرار میگیرند .

اگرچه رگرسیون کاکس کاربردی ترین مدل در تحلیل بقاست مدل پارامتریک در برخی شرایط می توانند مناسب تر باشند .

اغلب پژوهشگران در فیلد پزشکی متمایل به استفاده از مدل های نیمه پارامتریک چون کاکس هستند زیرا این مدل ها به پیش فرض های کمتری در مقایسه با مدل پارامتریک نیازمندند .

در برخی شرایط مدل های پارامتریک نسبت به کاکس نتایج بهتری دارند . در مدل های پارامتریک معمولا از روش درستنمایی ماکزیمم برای براورد پارامتر مجهول استفاده می شود که بسیار برای پژوهشگران آسان تر است


منابع

منابع فارسی

[۱] آبیار ، آمنه و محمدزاده ، محسن و مترجم ، کیومرث ، (۱۳۹۵) ، مدل های بقای کاکس و شکنندگی برای تحلیل داده های سرطان مری .

[۲]سمنانی ، ش . ، عربعلی ، ع . ، کشت کار ، ع . ، بهنام‌پور ، ن . ، بشارت ، س . و روشن دل ، غ (۱۳۸۸) ، نیترات و نیتریت منابع آب آشامیدنی مناطق شهری استان گلستان و بروز سرطان‌های مری و معده ، مجله دانشگاه علوم پزشکی کرمان ، ۱۶ ، ۲۹۱ – ۲۸۱ .

[۳] مترجم ، کیومرث و محمدزاده ، محسن و آبیار ، آمنه ، (۱۳۹۴) ، مدل های شکنندگی و خطرهای متناسب برای تحلیل داده‌های بقای فضایی.

[۴]محمدزاده م . (۱۳۹۱) ، آمار فضایی و کاربردهای آن ، مرکز نشر آثار علمی دانشگاه تربیت مدرس ، تهران ، ایران.

[۵] نورافکن ، ز . ، یاوری ، پ . ، روشندل ، غ . ، خلیلی ، د . ، بهنام‌پور ، ن . و زائری ، ف . (۱۳۹۲) ، برآورد میزان بقای مبتلایان به سرطان مری و برخی عوامل مرتبط با آن در استان گلستان در سال ۱۳۸۷ ، مجله اپیدمیولوژی ایران ، ۹ ، ۱۸ – ۱۱ .

منابع لاتین

[6] Aalen, O.O. and Tretli, S. (1999). Analyzing incidence of tests cancer by means of a frailty model. Cancer Causes Control, 10, 285-92.

[7] Andersen, P.K., Gill, R.D. (1982) Cox's regression model for counting processes: a large sample study. Annals of Statistics 10, 1100 – 1120.

[8] Bender, R., Augustin, T. and Blettner, M. (2005), Generating Survival Times to Simulate Cox Proportional Hazards Models, *Statistics in Medicine*, **24**, 1713-1723.


ممنون از توجه شما
abartazeha.com