دانشگاه صنعتی اصفهان
دانشکده علوم ریاضی

برآوردگر رگرسیونی داده‌های بقا سانسور شده بر روی فواصل با مدل شکست متناسب

پایان نامه کارشناسی ارشد آمار

محسن ابوالقاسمی

استاد راهنما

دکتر سروش علیمرادی

۱۳۷۹

۱۳۸۰ /۱۱/ ۱۵

دانشگاه صنعتی اصفهان

دانشکده علوم ریاضی

پایان نامه کارشناسی ارشد رشته آمار آقای محسن ابوالقاسمی

تحت عنـــوان

برآوردگر رگرسیونی داده‌های بقا سانسور شده بر روی فواصل با مدل شکست متناسب

در تاریخ ۷۹/۷/۳۰ توسط کمیته تخصصی زیر مورد بررسی و تصویب نهائی قرار گرفت.

۱ ـ استاد راهنمای پایان‌نامه — دکتر سروش علیرادی

۲ ـ استاد مشاور پایان‌نامه — دکتر علی رجالی

۳ ـ استاد داور ۱ — دکتر حمید پزشک

۴ ـ استاد داور ۲ — دکتر سیدمحمود طاهری

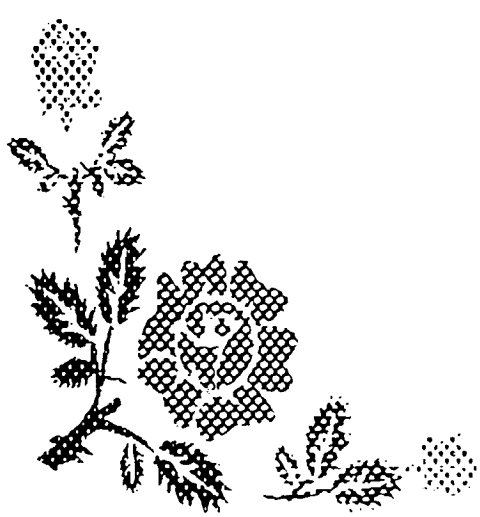
سرپرست تحصیلات تکمیلی دانشکده — دکتر امیر نادری

تقدیم به پیشگاه :

پدر و مادر عزیز و زحمتکشم

همسر مهربان و فداکارم

دوستان گرانقدر دوران تحصیلم



فهرست مطالب

چکیده

ما تحلیل رگرسیون داده های بقا را با مشاهدات سانسور از راست – چپ یا فاصله‌ای بررسی می کنیم. همچنین بـا بـه کارگیری مدل های لگاریتم خطی برای شکستهای سانسور شدهٔ فاصله ای روشهای جدول عمر را برای داده هـای بقـای سانسور شده گسترش می دهیم. الگوریتم EM برای محاسبهٔ برآوردهای حداکثر درستنمائی پارامترها استفاده شده است.

ما فرض می‌کنیم تابع شکست یک تابع پله ای روی فواصل مجزای زمان است ، بنابراین مدل نمایی پارامتری و مدل شکست متناسب کاکس به سادگی ابزاری برای حالات خاص هستند. ما الگوریتم محدودشدهٔ EM را برای آزمون فرض‌ها و بنا کردن فواصل اطمینان برای پارامترهاسازگار می‌کنیم. این روشها در یک تحلیل زمان عود بیماری در معالجهٔ مبتلایان ملانوما به کار برده شده اند.

فصل اول

مقدمه

تحلیل داده‌های بقاء[1] به مفهوم بررسی اطلاعاتی است که با طول عمر موجودات سرو کار دارد.فایدهٔ اصلی این کار، شناسایی روابط بین زمان بقاء و مجموعه‌ای از متغیرهای توضیحی[2] مانند: سن،جنس، رفتار و دیگر مشخصه‌ها است. [۱]

یکی از مهمترین کاربردهای تحلیل بقاء ، در پزشکی است. در این نوع تحلیل ها ارتباط بین طول مدت بقاء بیمار ومجموعه ای از متغیرها مانند: سن ، جنس ، میزان پیشرفت بیماری ، شیوهٔ درمان و سایر خصوصیات بالینی بیمار مورد بررسی قرار می گیرد . [۲]

1-Survival Analysis
2-Covariate Variables

اما مسئلهٔ پیچیده در مورد اطلاعات پزشکی بیماران ، این است که اطلاعات همیشه در دسترس نیست و ممکن است در قسمتی از زمان درمان در دسترس نباشد. اینجاست که مسئلهٔ داده‌های سانسور شده [۱] مطرح می‌گردد.اگر شکست را زمان مرگ بیمار در نظر بگیریم ، در صورتی که زمان دقیق شکست مشخص نباشد و یا اطلاعات جزئی در مورد زمان شکست وجود داشته باشد ، آن‌گاه داده ها سانسور شده اند.

در مورد اطلاعات پزشکی سه نوع سانسور وجود دارد. [۲]

۱- سانسور از راست [۲] : این سانسور زمانی اتفاق می‌افتد که اطلاعات پزشکی بیماران در انتهای طول دورهٔ درمان از دست رفته است.

۲- سانسور از چپ [۳] : این نوع سانسور به این مفهوم است که،اطلاعات پزشکی بیماران در ابتدای دورهٔ درمان وجود ندارد.

۳- سانسور بر روی فواصل [٤] : این نوع سانسور به این مفهوم است که ، قسمتی از اطلاعات پزشکی بیمار در طول دورهٔ درمان (یک یا چند فاصلهٔ زمانی جدا از هم) از دست رفته باشد.

در صورتی که اطلاعات شامل زمان‌های دقیق شکست [۵] و یا سانسور باشد، برای تحلیل زمان بقاء، چندین روش پارامتری و نیمه پارامتری وجود دارد.

سانسور بر روی فواصل موقعی اتفاق می‌افتد که زمان دقیق شکست مشخص نباشد و شکست در یک فاصلهٔ زمانی پیوسته اتفاق افتاده باشد.یکی از هدفهای مهم مطالعات پزشکی، کوتاه کردن طول فاصله سانسور است. در صورتی که اطلاعات بر روی فواصل سانسور شده باشند،چندین روش آماری برای تحلیل بقاء، وجود دارد.

تر نبال [٦] در سال ۱۹۷٦ ، یک الگوریتم تکرار شونده،برای تحلیل بقاء با داده‌های سانسور شده بر روی فواصل را ابداع نمود [۳] و در ادامه فین کلشتاین [۷] و ولف [۸] در سال ۱۹۸۵ یک مدل رگرسیون نیمه پارامتری [۹] را بر اساس تجزیه توزیع توأم زمان بقاء و متغیرها فرمول بندی نمودند. [٤]

1-Censored Data
2-Right-Censored
3-Left-Censored
4-Interval-Censored
5-Faiure Time
6- Turnball
7- Finkelstein
8- Wolfe
9- Semiparametric Regression

همچنین فین کلشتاین در سال ۱۹۸٦ روشی را برای تحلیل بقـاء در مـدل شکسـت متناسـب بـا داده‌هـای سانسور شده بر روی فواصل ، ابداع کرد. [٥]

در ادامه راخر [1] و مَسّرِر [2] در سال ، ۱۹۸۸ با استفاده از الگوریتم تکرار معرفی شده توسط تر نبال،زمـان بـهبود بیماران لوکمیا [3] را بر آورد نمودند. [٦]

ادل اندرسون [4] و اگوستینو [5] در سال ۱۹۹۲ ، برای تحلیل بقاء با داده‌های سانسور شده بر روی فواصـل، از مدل زمان شکست وایبال [6] استفاده نمودند. . [۷]

همچنیـن کیـم [7] و تیلـور [8] در سـال ، ۱۹۹٤ سـعی کردنـد کـه زمـان تشـخیص بیمـاری ایـدز [9] را بـا اسـتفاده از الگوریتم تکرار معرفی شده توسط ترنبال، دنبال نمایند. [۸]

با وجود این که اکثر اطلاعات پزشکی بیماران بر روی فواصل، سانسور شده می‌باشد، اما با توجه به مقـالات موجود واضح است که تاکنون تحقیقات بسیار کمی در مورد تحلیل بقاء بر روی اطلاعات سانسور شده روی فواصل ، انجام شده است.

دلیل اصلی کمی پیشرفت در تحلیل بقاء اطلاعـات سانسـور شـده روی فواصـل را مـی تـوان پیچیـدگی محاسباتی و همچنین عدم پشتیبانی کامل نرم افزارهای آماری این نوع تحلیل ها دانست.

هدف اصلی این رساله ، معرفی و فرمول بندی یک مدل رگرسیونی در تحلیل بقاء بـا داده هـای سانسـور شده بر روی فواصل و همچنین یک روش نسبتا ساده برای برآورد پارامترهای مدل است.

در فصل دوم ابتدا به معرفی برخی تعاریف و اصطلاحات مورد نیاز مـی پردازیـم ودر ادامـه، پیش‌نیازهـا و روابطی را که در طول رساله مورد نیاز است معرفی می کنیم .

در فصل سوم، مدل‌های رگرسیونی زمان شکست را بر اسـاس توزیـع هـای وایبـال و نمـایی ، معرفـی و در ادامـه ، پارامترهـای مـدل را بـر اسـاس داده هـای سانسـور شـده، بـرآورد می‌کنیـم [۹] و در پایـان ، مدل شکست متناسب [10] با داده‌های سانسور شده بر روی فواصل را معرفـی و پارامترهـای مـدل را بـا اسـتفاده از الگوریتم نیوتن رافسون [11] برآورد می‌کنیم. [۱۰]

در فصل چهارم ، ابتدا الگوریتم E M [12] را معرفی و در ادامه از این الگوریتم برای برآورد پارامترهای مدل(به روش حداکثر درستنمایی) با داده‌های گم شده،استفاده می‌کنیم .در ادامه از الگوریتـم محـدود EM در برآورد پارامترها با در نظر گرفتن محدودیت خطی $\underline{A}\,\underline{\theta}=\underline{a}$ استفاده می‌کنیم. ($\underline{A}$ یک مـاتریس معلـوم

1-Rücker
2-Messerer
3-Levkemia patients
4-Odell
5-Agostino
6-Weibull failure time
7-Kim
8-Taylor
9-AIDS
10-Proportional hazard model
11-Newton-Raphson algorithm
12- EM-algorithm

$q\times p$ بعدی و $\underline{a}$ بردار معلوم $q\times 1$ بعدی و $\underline{\theta}$ بردار پارامتری $P\times 1$ بعدی است.) اگر θ_j، عنصر j ام از بردار پارامتری $\underline{\theta}$ باشد، در آن صورت از الگوریتم محدود EM برای آزمون $H_0 : \theta_j = \theta_0$ (θ_0 مقدار ثابت) و ساختن فواصل اطمینان برای پارامترهای θ_j ($j=1,2,\ldots,p$) استفاده می‌کنیم. [۱۱]

در فصل پنجم، یک روش تقریباً ساده برای تخمین زمان بقاء با داده‌های سانسور شده روی فواصل مطرح و یک مدل رگرسیونی برای داده‌های بقاء سانسور شده بر روی فواصل را پایه ریزی می‌کنیم. در ادامه از الگوریتم EM که از مشخصه‌های آن سادگی، ثبات و خاصیت همگرایی مطلوب است، برای برآورد پارامترهای مدل رگرسیونی استفاده می‌کنیم. سپس از الگوریتم محدود EM برای آزمون فرض و ساختن فواصل اطمینان روی پارامترهای مدل استفاده می‌کنیم. در انتهای فصل نیز به عنوان مثال کاربردی، به تحلیل بقاء دربیماری ملانوما[1] می‌پردازیم. [۱۲]

در پایان رساله نیز به معرفی اجمالی از الگوی خطی تعمیم یافته پرداخته و برنامه های کامپیوتری مورد استفاده در رساله را می آوریم.

1- Melanoma

فصل دوم

تعاریف و مفاهیم اولیه

۲ . ۱ . مقدمه

در این فصل ابتدا به معرفی برخی اصطلاحات آماری و پیش‌نیازهایی که در طول رساله از آنها استفاده شده،می‌پردازیم، سپس روابطی را که در تجزیه و تحلیل بقا از آنها استفاده شده ، به دست می‌آوریم.

۲.۲. تعاریف و نمادها

همان طور که در فصل اول ذکر شد، تجزیه و تحلیل داده‌های بقا به مفهوم بررسی داده‌هایی است که با طول عمر موجودات جاندار سرو کار دارد.اما قبل از پرداختن به تحلیل رگرسیونی بر روی داده های بقا ابتدا به ذکر تعاریف لازم می پردازیم .

۲.۲ . ۱. تعریف قابلیت اعتماد

قابلیت اعتماد[1] به عنوان مشخصه‌ای از یک موجود، عبارت از احتمال کارکرد (زندگی) موجود تحت شرایط معین، در فاصلهٔ زمانی معین و بدون خرابی برای انجام یک ماموریت خاص است. [۱۳]

اگر قابلیت اعتماد را با R(t) نشان دهیم، داریم:

$R(t)=P$ (موجود در فاصلهٔ $(0,t)$ تحت شرایط لازم از انجام کار معین باز نماند) (۲-۱)

۲.۲ . ۲. تعریف تابع بقا

تابع بقا[2] به مفهوم احتمال زنده بودن یک موجود زنده در زمان معین است و آن را با S (t) نشان می‌دهیم. [۳]

$S(t)=P$ (موجود در فاصلهٔ $(0,t)$ زنده باشد) (۲-۲)

در رابطه با اطلاعات پزشکی بیماران ، از اصطلاح شکست برای بیمارانی که شیوهٔ درمان در بهبود وضعیت بالینی آنها مؤثر نبوده،استفاده می گردد.

عدم شکست نیز به این مفهوم است که یا شیوهٔ درمان در بهبود وضعیت بالینی موثر بوده و یا بیمار تا زمان t زنده می‌باشد و اطلاعات پزشکی بیمار از زما ن t به بعد در دسترس نیست.یعنی این اطلاعات در زمان t از راست سانسور شده است.

1- Reliability
2- Survival function