
بسم الله الرحمن الرحیم

دانشکده علوم ریاضی

رشته آمار ریاضی، گرایش آمار ریاضی

پایان‌نامه کارشناسی ارشد

مدل رگرسیون نرخ خطر کاکس جریمه شده

نگارنده: محمد گلپایگانی

استاد راهنما

دکتر محمد آرشی

شهریور ۱۳۹۶

شماره:
تاریخ:

باسمه‌تعالی

مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای محمد گلپایگانی با شماره دانشجویی ۹۴۳۸۸۷۴ رشته آماررياضی گرایش آمار ریاضی تحت عنوان مدل رگرسیون نرخ خطر کاکس جریمه شده که در تاریخ ۱۳۹۶/۶/۱۴ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

قبول (با درجه: عالی) ☑ **مردود** ☐

نوع تحقیق: نظری ☐ **عملی** ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنمای اول	دکتر محمد آرشی	دانشیار	
۲- استادراهنمای دوم			
۳- استاد مشاور			
۴- نماینده تحصیلات تکمیلی	دکتر احمد معتمد نژاد	دانشیار	
۵- استاد ممتحن اول	دکتر حسین باغیشنی	استادیار	
۶ـ استاد ممتحن دوم	دکتر محمدرضا رهبری	استادیار	

نام و نام خانوادگی رئیس دانشکده: دکتر ابراهیم هاشمی

تاریخ و امضاء و مهر دانشکده:

تبصره: در صورتی که کسی مردود شود حداکثر یکبار دیگر (در مدت مجاز تحصیل) می تواند از پایان نامه خود دفاع نماید (دفاع مجدد نباید زودتر از ۴ ماه برگزار شود).

تقدیم به خانواده مهربانم که هر لحظه، وجودم را از چشمه سار پر
از عشق چشمانشان سیراب می کنند....

سپاس‌گزاری...

پروردگارا...

نه می‌توانم مویشان را که در راه عزت من سفید شد، سیاه کنم و نه برای دست‌های پینه بسته‌شان که ثمره تلاش برای افتخار من است، مرهمی دارم. پس توفیقم ده که هر لحظه شکرگزارشان باشم و ثانیه‌های عمرم را در عصای دست بودنشان بگذارم.

در آغاز وظیفه خود می‌دانم از زحمات بی‌دریغ استاد راهنمای محترم جناب آقای دکتر محمد آرشی که علاوه در پژوهش گرد آمده در دوران تحصیل نیز از تجربیات ایشان بهره‌مند گردیده‌ام تشکر و قدردانی نمایم.

همچنین وظیفه خود می‌دانم از تمامی اساتید محترم گروه آمار دانشگاه صنعتی شاهرود، همچنین سرکار خانم مینا نوروزی‌راد که در تمامی مراحل این پایان نامه از دانش ایشان استفاده نموده‌ام قدر دانی به‌عمل بیاورم.

صمیمانه از مهربانان زندگیم، پدر و مادر عزیزم و نامزد مهربانم کمال تشکر را دارم.

در پایان از تمامی دانشجویان آمار و ریاضی دانشکده ریاضی و همچنین از مسئولین محترم آموزش دانشکده به پاس لطف و محبت‌هایشان قدر دانی می‌نمایم.

محمد گلپایگانی
شهریور ۱۳۹۶

تعهد نامه

اینجانب **محمد گلپایگانی** دانشجوی کارشناسی ارشد رشته **آمار ریاضی علوم ریاضی** دانشگاه شاهرود، نویسنده پایان‌نامه با عنوان **مدل رگرسیون نرخ خطر کاکس جریمه شده** ، تحت راهنمایی **محمد آرشی** متعهد می‌شوم:

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

محمد گلپایگانی
شهریور ۱۳۹۶

مالکیت نتایج و حق نشر

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

چکیده

با پیشرفت تکنولوژی در ذخیره‌سازی داده‌ها، اخیرا در مطالعات کارآزمایی بالینی مه‌داده‌های بقا ظاهر شده‌اند. تعیین متغیرهای موثر در پیش‌بینی زمان فوت در یک بیماری منجر به صرفه‌جویی در هزینه و زمان محاسبات می‌شود و از این رو از اهمیت بسزایی برخوردار است. امروزه در بسیاری از مطالعات پزشکی، از تکنولوژی ریزآرایه‌ها استفاده می‌شود تا بتواند ارتباط بین رفتار ژن‌ها و ویژگی‌های مختلف بالینی بیماران را پیدا کند.

در این پایان‌نامه، با به کارگیری توابع جریمه ریج، لاسو و ترکیبی از این دو تابع، الاستیک‌نت، در رگرسیون خطرات متناسب کاکس، مدل مناسب برای پیش‌بینی الگوریتم‌های مختلف معرفی و مورد مقایسه قرار می‌گیرد.

کلمات کلیدی:

برآوردگر الاستیک‌نت، تحلیل بقا، رگرسیون جریمه‌شده، مدل خطرات متناسب کاکس.

لیست مقالات مستخرج از پایان‌نامه

۱. گلپایگانی، م.، آرشی، م.، نوروزی‌راد، م. (۱۳۹۶). کاربرد بهینه سازی در قابلیت اطمینان، دهمین کنفرانس بین‌المللی انجمن ایرانی تحقیق در عملیات، بابلسر.

۲. گلپایگانی، م.، آرشی، م.، نوروزی‌راد، م. (۱۳۹۶). روش لارس برای برآوردگر الاستیک‌نت سازوار در مدل نرخ خطر کاکس، یازدهمین سمینار احتمال و فرآیندهای تصادفی، قزوین.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مقدمات و تعاریف

۱.۱ مقدمه

امروزه در بسیاری از مطالعات پزشکی، از تکنولوژی ریزآرایه‌ها استفاده می‌شود تا بتوان ارتباط بین رفتار ژن‌ها و ویژگی‌های مختلف بالینی بیماران را پیدا کرد. از آن‌جایی که تعداد ژن‌ها بسیار زیاد است، به منظور صرفه‌جویی در هزینه و زمان، تعیین ژن‌های بااهمیت و به‌طور کلی متغیرهای موثر در پیش‌بینی یک بیماری یا زمان فوت یک بیمار از اهمیت بسزایی برخوردار است. در این پایان‌نامه با به‌کار گیری تکنیک جریمه‌شده در رگرسیون این نوع از داده‌ها که به داده‌های بقا معروف هستند، مدل مناسب برای پیش‌بینی الگوریتم‌های مختلف معرفی و مقایسه می‌شود. در ادامه، تعریف مختصری از رگرسیون ارایه شده است و سپس رگرسیون جریمه‌شده و انواع آن در حالت کلی مورد بررسی قرار گرفته است. بررسی و یافتن یک مدل برای زمان‌های ثبت‌شده از وقوع پدیده‌هایی که زمان رخ دادن آن‌ها از اهمیت بسزایی برخوردار است، هدف شاخه‌ای از آمار به نام «تحلیل بقا» است. رگرسیون کاکس (فصل ۲) و همین‌طور رگرسیون کاکس جریمه‌شده (فصل ۳) نیز زیرشاخه‌ای از تحلیل بقا است که در انتهای فصل ۱، مفاهیم

اساسی آن بیان شده است.

۲.۱ رگرسیون

واژه‌ی «رگرسیون[۱]» از نظر لغوی به معنی *پسروی*، *برگشت* و *بازگشت* است اما از دید آمار و ریاضی، از این واژه جهت رساندن مفهوم «بازگشت به یک مقدار متوسط یا میانگین یا شاخص دیگر از توزیع» به‌کار می‌رود. بدین معنی که برخی از پدیده‌ها با گذشت زمان، از نظر کمی به طرف یک مقدار متوسط میل می‌کنند.

در سال ۱۸۷۷، فرانسیس گالتون[۲] در مقاله‌ای که درباره‌ی بازگشت به میانگین منتشر کرد، بیان کرده بود که متوسط قد پسران دارای پدران قد بلند، کمتر از قد پدرانشان می‌باشد. همین‌طور متوسط قد پسران دارای پدران کوتاه‌قد، بیشتر از قد پدرانشان است. به این ترتیب، گالتون پدیده‌ی بازگشت به میانگین را در داده‌های قد پدران و پسران تایید کرد. گالتون، مفهومی زیست‌شناختی به رگرسیون داده بود اما کارل پیرسون[۳] کارهای گالتون را برای مفاهیم آماری توسعه داد.

گالتون، برای تاکید بر پدیده‌ی «بازگشت به میانگین» از تحلیل رگرسیون استفاده کرد، اما به هر حال امروزه تحلیل رگرسیونی فن و تکنیک آماری برای بررسی و مدل‌سازی ارتباط بین متغیرها است.

یک مدل رگرسیونی، متغیرهای زیر را دارد:

- پارامترهای ناشناخته که معمولا با β نمایش داده می‌شود.
- متغیر یا متغیرهای مستقل یا پیش‌بین $\boldsymbol{X}$
- متغیر وابسته یا متغیر پاسخ $\boldsymbol{Y}$

یک مدل رگرسیونی، $\boldsymbol{Y}$ را به تابعی از $\boldsymbol{X}$ و $\boldsymbol{\beta}$، $f(\boldsymbol{X},\boldsymbol{\beta})$، مرتبط می‌کند.

فرض کنید، بردار $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_n)^\top$، p بعدی و n مشاهده‌ی $(\boldsymbol{X}_i, Y_i)$ در اختیار باشد

[۱] Regression
[۲] Francis Galton
[۳] Karl Pearson

که ماتریس $n \times p$ بعدی $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ که در آن $\boldsymbol{X}_i \in \mathbb{R}^p$ و بردار nـبعدی $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ را می‌سازند. تحلیل رگرسیون با یکی از حالت‌های زیر روبرو می‌شود:

- $n < p$ در این حالت تعداد مشاهدات از تعداد پارامترها کمتر است و از آن‌جایی که سیستم معادلات تعریف‌شده برای مدل رگرسیون قابل برآورد نیست و داده‌ی کافی برای بازیابی $\boldsymbol{\beta}$ وجود ندارد، دیدگاه کلاسیک برای این تحلیل نمی‌تواند استفاده شود.

- $n = p$ در این حالت، تعداد مشاهدات با تعداد پارامترها برابر است. اگر تابع f، یک تابع خطی باشد، معادلات $\boldsymbol{Y} = f(\boldsymbol{X}, \boldsymbol{\beta})$ یک جواب دقیق خواهند داشت. تعداد محاسبات به یک مجموعه‌ی n معادله با n مجهول (همان تعداد عناصر $\boldsymbol{\beta}$) کاهش می‌یابد و در صورتی‌که $\boldsymbol{X}$ متغیرهای مستقل خطی باشند، یک جواب یکتا خواهد داشت.

- $n > p$ در این حالت، تعداد مشاهدات از تعداد پارامترها بیشتر است. در این صورت اطلاعات کافی در داده‌ها برای تخمین مقدار یکتا برای $\boldsymbol{\beta}$ وجود دارد.

حالت سوم، بسیار متداول است و در این مورد، رگرسیون، ابزاری است که به دنبال تخمین پارامترهای مجهول $\boldsymbol{\beta}$ است که فاصله بین مقادیر پیش‌بینی و اندازه‌گیری‌شده از متغیر پاسخ Y را کمینه می‌کند و نیز، با در نظر گرفتن پذیره‌های آماری مشخصی، تحلیل رگرسیون از اطلاعات زیادی برای تعیین اطلاعات آماری درباره‌ی پارامترهای مجهول $\boldsymbol{\beta}$ و مقادیر پیش‌بینی متغیر Y استفاده می‌کند.

شیوه‌های مهم تحلیل‌های رگرسیونی به‌صورت زیر هستند:

- رگرسیون خطی ساده[۴]
- رگرسیون خطی چندگانه[۵]
- رگرسیون خطی جریمه‌شده[۶]

[۴] Simple linear regression
[۵] Multiple linear regression
[۶] Penalized linear regression

۱.۲.۱ رگرسیون خطی ساده

در این رگرسیون، متغیر وابسته‌ی Y، ترکیب خطی از پارامترها است (لزومی ندارد نسبت به متغیرهای مستقل، خطی باشد). وقتی از رگرسیون متغیر Y بر متغیر X صحبت می‌شود، به دلیل همبستگی بین دو متغیر است.

همبستگی بین متغیرهای X و Y، بدین معنی است که تغییر در متغیر Y، چقدر بر X تاثیر می‌گذارد. به عبارت دیگر، هماهنگی بین متغیرهای X و Y را نشان می‌دهد. برای مثال، می‌توان به هماهنگی قد و وزن اشاره کرد. بدیهی است همبستگی، مثبت است زیرا افرادی که قدر بلندتری دارند، وزن بیشتری نیز دارند. معمولا همبستگی را با ضریب همبستگی پیرسن[۷] اندازه‌گیری می‌کنند. هر چه قدر مطلق مقدار همبستگی به عدد یک نزدیکتر باشد، همبستگی بین دو متغیر بیشتر است و هر چه به عدد صفر نزدیک‌تر باشد، به معنی عدم همبستگی است. همبستگی برابر یک، یعنی وجود یک رابطه‌ی خطی و کامل. همبستگی می تواند مثبت (ارتباط مستقیم) یا منفی (ارتباط غیرمستقیم) باشد. با رسم نمودار پراکنش، می‌توان شدت همبستگی بین دو متغیر را مشاهده کرد.

در یک تحلیل رگرسیونی، در ابتدا فرض می‌شود بین دو متغیر ارتباطی وجود دارد. در رگرسیون خطی، این ارتباط باید به‌صورت یک خط بین دو متغیر وجود داشته باشد. با رسم نمودار پراکنش، این ارتباط بررسی می‌شود. در صورتی که نمودار نشان‌دهنده‌ی این باشد که داده‌ها تقریبا (نه لزوما دقیق) در امتداد یک خط مستقیم پراکنده شده‌اند، حدس اولیه تایید می‌شود و این ارتباط به صورت $Y = \beta_0 + \beta_1 X$ نشان داده می‌شود که در آن، β_0، عرض از مبدا و β_1، شیب خط است. از آن‌جایی که ممکن است به خاطر اندازه‌گیری متغیرها یا شرایط محیط یا تفاوت‌های محیطی و ... مقداری خطا وجود داشته باشد، معادله‌ی اولیه به‌صورت زیر اصلاح می‌شود:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (۱.۱)$$

معادله‌ی (۱.۱) یک معادله‌ی رگرسیون خطی ساده نامیده می‌شود که ε خطای تصادفی است. معمولا فرض می‌شود که (۱) خطاها یکدیگر را خنثی می‌کنند به‌عبارت دیگر، مجموع خطاها

[۷] Pearson's coefficient

و در نتیجه میانگین آن‌ها، صفر است. (۲) خطای موجود در یک مشاهده رابطه‌ای با دیگر خطاها ندارد، به‌عبارت دیگر، خطاها مستقل از یکدیگر هستند. (۳) تغییرات بین خطاها ثابت در نظر گرفته می‌شود. به عبارت‌دیگر، واریانس خطاها ثابت است. چنانچه $E(\varepsilon) = ۰$ باشد، می‌توان معادله‌ی (۱.۱) را به منزله یک رگرسیون میانگین دانست که در آن میانگین شرطی پاسخ به‌صورت یک رابطه خطی مدل می‌شود. به‌عبارت دقیق‌تر میانگین شرطی پاسخ از طریق یک خط راست با متغیرهای پیش‌بین در ارتباط است و داریم:

$$E(Y|X) = \beta_۰ + \beta_۱ X$$

برای مثال، فرض کنید متغیر پاسخ Y، مرگ و میر (در هر ۱۰ میلیون نفر) ناشی از سرطان پوست و متغیر مستقل X، عرض جغرافیایی (درجه شمالی) مرکز هر یک از ۴۹ ایالت کشور آمریکا باشد (این مجموعه داده، در اواخر دهه‌ی ۱۹۵۰ گردآوری شده است، بنابراین آلاسکا و هاوایی جزو ایالت‌های آمریکا محسوب نشده‌اند و واشینگتون دی. سی. نیز در شمار ایالت‌ها آورده شده است). شکل ۱.۱ این ارتباط را نشان می‌دهد که به‌صورت یک رابطه‌ی خطی معکوس می‌باشد.

تا این مرحله، مدل رگرسیون خطی ساده معرفی شده و کافی است پارامترهای مجهول مدل (در اینجا $\beta_۰$ و $\beta_۱$) برآورد شوند. برآورد پارامترها به روش‌های مختلفی انجام می‌شود که از مهم‌ترین آن‌ها می‌توان به روش کمترین توان‌های دوم خطا و درستنمایی ماکزیمم اشاره کرد.

فرض کنید مجموعه داده‌های $\{(X_i, Y_i); i = ۱, ۲, \ldots, n\}$ در دسترس باشد. با داشتن مجموعه‌ای از نقاط می‌توان مدل را به‌صورت زیر به‌دست آورد:

$$Y_i = \hat{\beta}_۰ + \hat{\beta}_۱ X_۱ + e_i = \hat{Y}_i + e_i \qquad (۲.۱)$$

که در آن، $e_i = Y_i - \hat{Y}_i$ مانده نام دارد. از متداول‌ترین روش‌ها برای به‌دست آوردن پارامترها، کمینه کردن مجموع توان‌های دوم مانده‌ها[۸]، SSE می‌باشد. یعنی

$$\text{SSE} = \sum_{i=۱}^{n} e_i^۲ = \sum_{i=۱}^{n} (Y_i - \hat{Y}_i)^۲ = \sum_{i=۱}^{n} (Y_i - \hat{\beta}_۰ - \hat{\beta}_۱ X_i)^۲ \qquad (۳.۱)$$

[۸] Sum of square errors

در مورد رگرسیون خطی ساده، پارامترها با این روش برابر خواهند بود با:

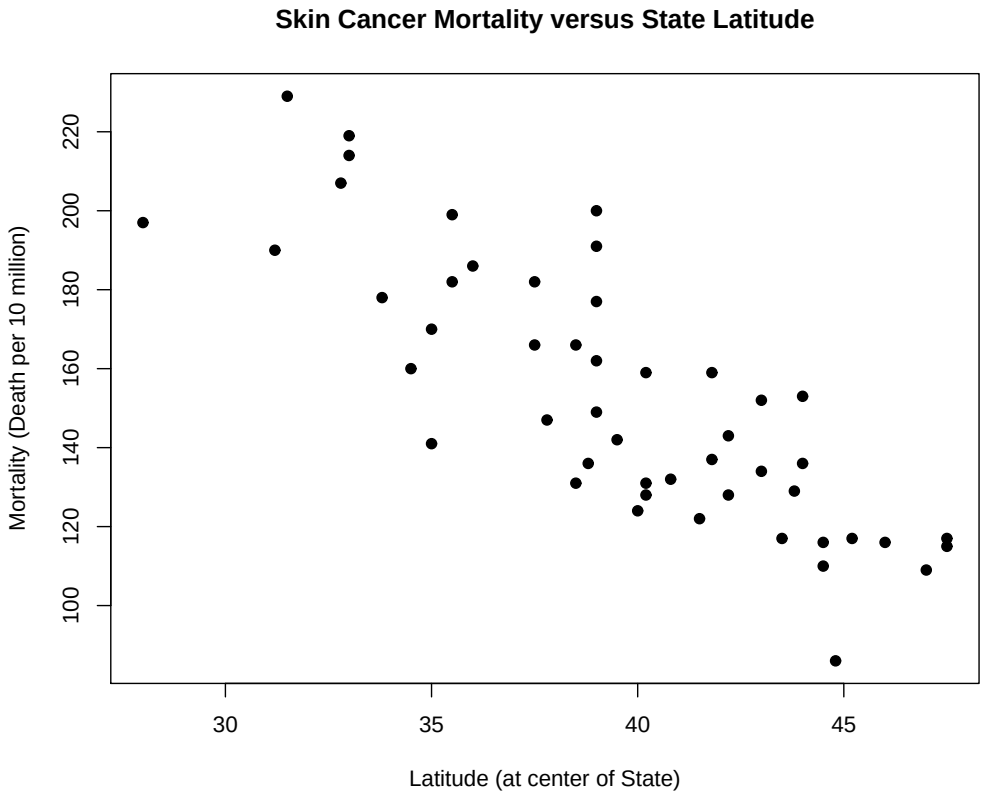
$$\begin{aligned}\hat{\beta}_۱ &= \frac{\sum_{i=۱}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=۱}^{n}(X_i-\bar{X})^۲}\\ \hat{\beta}_۰ &= \bar{Y}-\hat{\beta}_۱\bar{X}\end{aligned}$$

که در آن $\bar{X}$ و $\bar{Y}$ میانگین X و Y هستند.

برآورد پارامترها در مجموعه داده‌های شکل ۱.۱ با روش کمترین توان‌های دوم به‌صورت $\hat{\beta}_۰ = ۳۸۹/۱۹$ و $\hat{\beta}_۱ = -۵/۹۸$ به‌دست آمد.

۲.۲.۱ رگرسیون خطی چندگانه

رگرسیون خطی چندگانه تعمیمی از رگرسیون خطی ساده است. در این رگرسیون، مقادیر متغیر وابسته‌ی Y از روی مقادیر دو یا چند متغیر دیگر (متغیرهای مستقل $X_۱$، $X_۲$، ... و X_p) برآورد



شکل ۱.۱: ارتباط بین سرطان پوست با عرض جغرافیایی مرکز هر ایالت در آمریکا

می‌شوند. معادله‌ی رگرسیون خطی چندگانه برای مشاهده‌ی iام به‌صورت زیر است:

$$Y_i = \beta_۰ + \beta_۱ X_{i۱} + \cdots + \beta_p X_{ip} + \varepsilon_i, \qquad i = ۱, \cdots n \tag{۴.۱}$$

که در آن، پارامترهای $\beta_۱$، $\beta_۲$، ... و β_p ضرایب رگرسیون جزیی هستند و $\beta_۰$ مقدار ثابت رگرسیون است. β_iها ضرایب رگرسیون جزیی نامیده می‌شوند زیرا این ضرایب، میزان افزایش متغیر وابسته را به ازای یک واحد افزایش متغیر مستقل موردنظر با ثابت نگه داشتن سایر متغیرها نشان می‌دهد.

می‌توان مدل (۴.۱) را براساس نماد ماتریسی به‌صورت زیر نوشت:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{۵.۱}$$

که در آن $\boldsymbol{Y} = (Y_۱, \ldots, Y_n)^\top$ یک بردار n مولفه‌ای شامل متغیر پاسخ و Y_i نمایش‌دهنده iامین پاسخ مشاهده‌شده، $\boldsymbol{X} = (\boldsymbol{X}_۱, \ldots, \boldsymbol{X}_n)^\top$ ماتریس $n \times p$ که در آن $\boldsymbol{X}_i \in \mathbb{R}^p$. فرض کنید ماتریس $\boldsymbol{X}$ دارای رتبه کامل ستونی باشد، $\boldsymbol{\beta} = (\boldsymbol{\beta}_۱, \ldots, \boldsymbol{\beta}_n)^\top$ یک بردار p بعدی شامل ضرایب رگرسیونی و $\varepsilon = (\varepsilon_۱, \ldots, \varepsilon_n)^\top$ بردار n مولفه‌ای خطای تصادفی است که فرض کنید $E(\epsilon) = ۰$ و $E(\epsilon\epsilon^\top) = \sigma^۲ I_n$.

برآورد کمترین توان‌های دوم معمولی[۹] (OLS) برآوردگری است که SSE در زیر را کمینه کند:

$$\begin{aligned}
\mathrm{SSE}(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} &= \sum_{i=۱}^{n} \varepsilon_i^۲ \\
&= \sum_{i=۱}^{n} \left(Y_i - \hat{\boldsymbol{\beta}}_۰ - \sum_{j=۱}^{p} \hat{\boldsymbol{\beta}}_j \boldsymbol{X}_{ij} \right)^۲ \\
&= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= \boldsymbol{Y}^\top \boldsymbol{Y} - ۲\boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}
\end{aligned} \tag{۶.۱}$$

[۹] Ordinary least squares

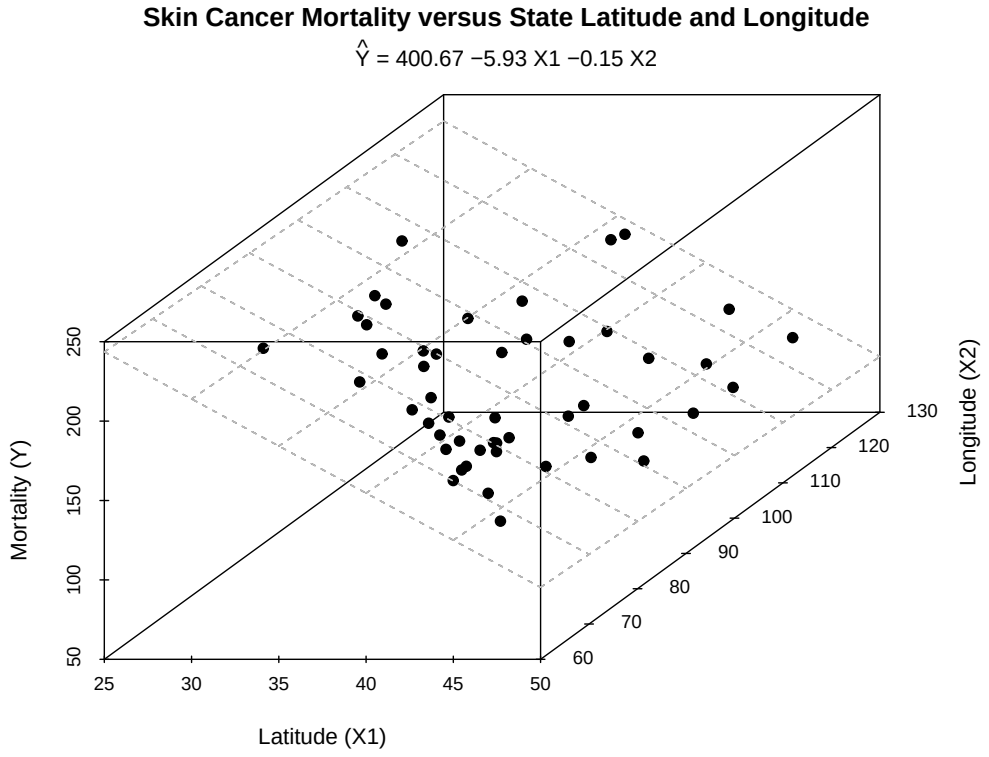
با مشتق‌گیری از $\text{SSE}(\boldsymbol{\beta})$ نسبت به $\boldsymbol{\beta}$ نتیجه می‌شود $-۲\boldsymbol{X}^\top\boldsymbol{Y}+۲\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}=۰$ از آن‌جایی که رتبه‌ی ماتریس $\boldsymbol{X}$، p بوده و p ستون مستقل خطی وجود دارد، بنابراین ماتریس p بعدی $\boldsymbol{X}^\top\boldsymbol{X}$ معکوس‌پذیر است. بنابراین، برآوردگر ضرایب مدل به‌صورت زیر به‌دست می‌آید:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \text{argmin}_{\boldsymbol{\beta}}\{\text{SSE}(\boldsymbol{\beta})\}\\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-۱}\boldsymbol{X}^\top\boldsymbol{Y}.\end{aligned}$$

در مجموعه‌ی داده‌ی شکل ۱.۱، فرض کنید تعداد مرگ و میر به طول جغرافیایی مرکز هر ایالت نیز وابسته باشد. در این صورت اگر عرض و طول جغرافیایی به ترتیب با متغیرهای $X_۱$ و $X_۲$ نشان داده شود، آن‌گاه مدل رگرسیونی به‌صورت $Y=\beta_۰+\beta_۱X_۱+\beta_۲X_۲+\varepsilon$ خواهد بود که برآورد پارامترها به روش کمترین توان‌های دوم عبارتند از: $\hat{\beta}_۰=۴۰۰/۶۷$، $\hat{\beta}_۱=-۵/۹۳$ و $\hat{\beta}_۲=-۰/۱۵$. به‌کارگیری صفت خطی بودن در این مدل رگرسیونی اشاره براین دارد که مدل، برحسب پارامترهای $\beta_۰$، ... و β_p خطی است و یک ابرصفحه در فضای p بعدی از متغیرهای مستقل X_i ایجاد می‌کند که شکل ۲.۱ این موضوع را به‌خوبی نشان می‌دهد.

۳.۱ مدل رگرسیون چندگانه جریمه‌شده

معمولا برآوردگر OLS، متداول‌ترین روش در برآورد پارامتر رگرسیون، برآوردگر خوبی از $\boldsymbol{\beta}$ است. از مهم‌ترین ویژگی این روش، محاسبات نسبتا ساده آن است. علاوه بر آن، قضیه‌ی معروف گوس‌ـ مارکوف نیز در این راستا وجود دارد که ثابت می‌کند بهترین برآوردگر خطی نااریب پارامتر $\boldsymbol{\beta}$، برآوردگر OLS است. اما موقعیت‌های متفاوتی وجود دارند که در برخی از آن‌ها، این برآوردگر به خوبی رفتار نمی‌کند. هنگامی‌که خطاها توزیع دم سنگین دارند، داده‌های پرت می‌توانند بر برآوردگر OLS تاثیر زیادی بگذارند. وقتی که تعداد پارامترها زیاد است یا وقتی ستون‌های ماتریس X به‌شدت همبسته هستند یا به عبارت دیگر، در داده‌ها هم‌خطی وجود دارد، واریانس برآوردگر OLS افزایش می‌یابد. این بدان معنی است که قدرمطلق برآوردهای کمترین توان‌های دوم خیلی بزرگ می‌شود و در نتیجه برآوردهای ناپایداری تولید می‌کند. یعنی با ارایه‌ی نمونه‌های متفاوت اندازه‌ی این برآوردگرها به‌طور قابل توجهی تغییر می‌کند و در نتیجه درستی پیش‌گویی‌ها کاهش می‌یابد (رستا و همکاران، ۱۳۸۹). این برآوردها، اغلب دارای اریبی کم



شکل ۲.۱: صفحه رگرسیون سرطان پوست با عرض و طول جغرافیایی مرکز هر ایالت در آمریکا

ولی واریانس زیاد هستند.

یک روش برای مواجهه با این مسایل، استفاده از تکنیک جریمه است. این روش، به جای کمینه کردن تابع مجموع توان دوم خطا که در رابطه‌ی (۶.۱) تعریف شده است، تابع SSE جریمه‌شده‌ی زیر را کمینه می‌کند:

$$\mathrm{SSE}_{(\beta)} + P_\lambda(\boldsymbol{\beta}) \qquad (۷.۱)$$

که در آن، $P_\lambda(\boldsymbol{\beta})$ تابعی است که مقادیر بزرگ‌تر پارامترهای مجهول را جریمه می‌کند. روش‌های جریمه‌شده همواره با حداقل یک پارامتر تنظیم‌کننده[۱۰] (در اینجا، λ) ظاهر می‌شود که موازنه بین تابع SSE و تابع جریمه را کنترل می‌کند.

مثال ۱.۳.۱. به عنوان یک مثال، فرض کنید مدل رگرسیونی به‌صورت $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 +$

[۱۰] Tuning parameter

ε باشد و نیز متغیرهای $X_۱$ و $X_۲$ به‌شدت همبسته باشند. بدین منظور، ۲۰ عدد از توزیع نرمال استاندارد به عنوان متغیر $X_۱$ و ۲۰ عدد از توزیع نرمال با میانگین $X_۱$ و انحراف‌معیار ۰/۰۱ تولید شده است. همبستگی بین این دو متغیر، ۰/۹۹ به‌دست می‌آید که نشان می‌دهد به‌شدت همبسته هستند. فرض کنید مقدار واقعی بردار پارامترهای $\boldsymbol{\beta}$ به صورت $(۳, ۱, ۱)$ باشد. با روش کمترین توان‌های دوم، برآورد ضرایب به‌صورت $(۳/۰۹, ۲۲/۶۰, -۲۰/۳۹)$ به‌دست می‌آید. همان‌طور که ملاحظه می‌شود برآوردهای به‌دست آمده با مقدار واقعی پارامتر اختلاف زیادی دارند که به‌دلیل همبسته بودن متغیرهای $X_۱$ و $X_۲$ است. مقدار تابع SSE به‌ازای مقدار واقعی پارامتر $\boldsymbol{\beta}$ برابر با ۱۲/۸۴ و به‌ازای برآوردگر OLS برابر با ۱۰/۷۲ است که اختلاف زیادی مشاهده نمی‌شود. همین‌طور $\beta_۱ + \beta_۲ = ۲$ و مجموع مولفه‌های دوم و سوم برآوردگر OLS برابر با ۲/۲۲ می‌باشد که بسیار به هم نزدیک می‌باشند. اما $\beta_۱^۲ + \beta_۲^۲ = ۱^۲ + ۱^۲ = ۲$ عدد شدیدا کوچکتری از $(-۲۰/۳۹)^۲ + (۲۲/۶۰)^۲ = ۹۲۶/۶۲۹۸$ است. بنابراین از روش‌های موثر در این مسئله این است که تابع جریمه $\lambda \sum_{j=۱}^{p} \beta_j^۲$ در رابطه‌ی (۷.۱) در نظر گرفته شود. بردار $(۱, ۱)$ به‌عنوان پاسخ، جریمه‌ی کمتری را منجر می‌شود و نسبت به $(۲۲/۶۰, -۲۰/۳۹)$ از علاقه‌مندی بیشتری برخوردار است. این رهیافت (اضافه کردن تابع جریمه‌ی $\sum_{j=۱}^{p} \beta_j^۲$ به تابع SSE)، رگرسیون ریج نام دارد که به‌ازای مقدار دلخواه $\lambda = ۱$، بردار $\boldsymbol{\beta}$ را به‌صورت $(۳/۰۹, ۱/۱۲, ۱/۰۸)$ برآورد می‌کند که مشاهده می‌شود به مقادیر واقعی پارامتر نزدیک می‌باشد.

فرض کنید تعدادی متغیر پیش‌بین مهم وجود دارند اما برخی از آن‌ها با متغیر پاسخ نامرتبط هستند. مدل‌هایی با تعداد متغیرهای کمتر همواره دلخواه تحلیل‌گران آماری بوده است چرا که این گونه مدل‌ها (۱) روابط بین متغیرهای علمی را به‌صورت ساده بیان می‌کنند و قابلیت تفسیر بیشتری دارند؛ (۲) هزینه‌های محاسباتی و صرف زمان زیاد، روش‌های انتخاب متغیر سنتی و کلاسیک را با چالش بسیار جدی در زمینه‌ی استفاده از داده‌ها با بعد بالا رو به رو کرده است؛ (۳) انتخاب یک مدل آماری خوب، مدلی است که استفاده از آن آسان، قابل فهم و قابل توضیح برای دیگران باشد.

از آن‌جایی که کمینه کردن تابع SSE مدل کامل، برآوردگری با واریانس زیاد را تولید می‌کند، پیشنهاد شده است یک زیرمجموعه از متغیرها بر اساس معنی‌داری آن‌ها در مدل یا برخی از

معیارهای انتخاب مدل، انتخاب شود. مشکل اساسی که در این رهیافت وجود دارد این است که معیارها زمانی قابل استفاده هستند که همه‌ی زیرمدل‌های ممکن شناخته شده باشند اما تعداد همه‌ی زیرمجموعه‌های ممکن به‌صورت نمایی با افزایش p، زیاد می‌شود و با رایانه‌های امروزی، عملا برای pهای بزرگتر از ۴۰ یا ۵۰ محاسبات ممکن نیست. بدین دلیل روش‌های اصلاح‌شده‌ای مانند انتخاب پیش‌رو، پس‌رو و گام به گام، ترکیبی از انتخاب‌های پیش‌رو و پس‌رو، پیشنهاد شده‌اند. مشکل دیگری که می‌توان به آن اشاره کرد، گسسته بودن این روش است. بنابراین با تغییر بی‌نهایت کوچکی در داده‌ها، برآوردهای کاملا متفاوتی به‌دست می‌آیند. در نتیجه، رهیافت انتخاب زیرمجموعه[۱۱] اغلب ناپایدار بوده و با تغییرپذیری زیاد همراه است.

ایده‌ی جریمه کردن در انتخاب متغیر و کاهش تعداد متغیرها در مدل‌سازی‌های آماری می‌تواند بسیار مفید باشد. روش‌های جریمه کردن را می‌توان به دو گروه کلی تقسیم کرد: گروه اول، روش‌های مبتنی بر جریمه کردن اندازه‌ی مدل مانند معیار اطلاع آکاییک[۱۲] (AIC) و معیار اطلاع بیزی[۱۳] (BIC) که به ترتیب توسط آکاییک (۱۹۷۳) و شوارتز[۱۴] (۱۹۷۸) معرفی شدند. گروه دوم، روش‌های مبتنی بر جریمه کردن تک تک متغیرهای پیش‌بین که معروف به روش‌های کمترین توان‌های دوم جریمه‌شده است که این روش‌ها در انتخاب متغیر، پایدارتر، پیوسته و با روش محاسباتی کارامد است.

تعداد زیادی از معیارهای انتخاب مدل بر اساس نظریه‌ی اطلاع پدید آمده‌اند. یکی از پرکاربردترین این معیارها AIC است که به‌صورت زیر تعریف می‌شود:

$$\text{AIC} = -۲l + ۲p \qquad (۸.۱)$$

که در آن، l لگاریتم تابع درستنمایی و p تعداد ضرایب رگرسیونی موجود در مدل می‌باشد. هر مدل زیرمجموعه با AIC کمتر، بر سایر مدل‌های زیرمجموعه برتری دارد. پس از این معیار، معیارهای دیگری پدید آمده‌اند که معروف‌ترین آن‌ها BIC است و به‌صورت زیر تعریف می‌شود:

$$\text{BIC} = -۲l + p\ln(n) \qquad (۹.۱)$$

[۱۱] Subset selection
[۱۲] Akaike Information Criterion
[۱۳] Bayesian Information Criterion
[۱۴] Schwarz

که در آن، n حجم نمونه است. برای آشنایی با سایر معیارهای اطلاع به بورنهام و اندرسون[۱۵] (۲۰۰۲) مراجعه کنید.

از دیگر موارد به کارگیری رگرسیون جریمه‌شده، زمانی است که تعداد متغیرها از تعداد مشاهدات خیلی بیشتر باشد. امروزه در بسیاری از مطالعات پزشکی و بیولوژیکی، چنین داده‌هایی دیده می‌شود. مثلا فرض کنید محققی می‌خواهد یک متغیر پاسخ (فشار خون، وضعیت بیماری، پاسخ به درمان و ...) را بر اساس ویژگی‌های مولکولی (مانند ژن‌ها، الگوهای DNA، متابولیک‌ها و ...) پیش‌بینی کند.

فن و لی[۱۶] (۲۰۰۱) سه اصل را ارایه کردند که یک تابع جریمه خوب باید ویژگی‌های زیر را داشته باشد:

نااریبی : به منظور جلوگیری از اریبی برآوردگر، باید پارامتر برآوردشده به مقادیر بزرگ پارامتر مجهول نزدیک باشد.

تنکی [۱۷]: برآوردگر باید یک حد آستانه داشته باشد که به‌صورت خودکار پارامترهای برآوردی کوچک را صفر در نظر بگیرد تا بدین وسیله از پیچیدگی مدل کاسته شود.

پیوستگی : برآوردگر به‌دست آمده باید پیوسته باشد تا پیش‌بینی‌های مدل پایدار باشد.

برای به‌کار بردن تکنیک جریمه، بهتر است ابتدا متغیرهای پیش‌بین استاندارد شوند به‌طوری که میانگین آن‌ها صفر و واریانس آن‌ها، یک باشد. متغیر پاسخ نیز مرکزی می‌شود به‌گونه‌ای که میانگین آن برابر صفر گردد. به عبارت دیگر،

$$\bar{Y} = ۰, \qquad \bar{\boldsymbol{X}}_j = ۰, \qquad \boldsymbol{X}_j^\top \boldsymbol{X}_j = n, \quad \forall j \tag{۱۰.۱}$$

برای مثال، فرض کنید متغیر $X_۱$ از ۰ تا ۱ و متغیر $X_۲$ از صفر تا یک میلیون تغییر کند. واضح است که یک واحد تغییر در متغیرها برای هر دوی این متغیرها به یک معنی نیست. وقتی داده‌ها استاندارد شده باشند، در این صورت، عرض از مبدا در مدل وجود نخواهد داشت.

[۱۵] Burnham and Anderson
[۱۶] Fan and Li
[۱۷] Sparse

لازم به توضیح است تغییرات در مقیاس را بعد از برازش مدل، می‌توان به‌صورت زیر بازیابی کرد:

$$X_{ij}\beta_j = \frac{X_{ij}}{a}a\beta_j = \tilde{X}_{ij}\tilde{\beta}_j, \qquad \tilde{X}_{ij} = \frac{X_{ij}}{a}, \qquad \tilde{\beta}_j = a\beta_j \tag{۱۱.۱}$$

به عبارت دیگر، اگر متغیر X_j بر a تقسیم شود، به‌سادگی می‌توان پاسخ $\tilde{\beta}_j$ را بر a تقسیم کرد تا β_j اولیه را به‌دست آورد.

در این پایان نامه برای حل مشکل همخطی و انتخاب متغیر، استفاده از مدل‌های رگرسیونی ریج، لاسو و الاستیک‌نت که از کاربردی‌ترین مدل‌های رگرسیونی جریمه‌شده هستند، مورد استفاده قرار می‌گیرند.

۱.۳.۱ برآورد ریج

روش‌های متعددی برای غلبه بر مشکل همخطی در مدل‌های خطی ارائه شده‌اند. یکی از موثرترین روش‌ها برای حل مشکل همخطی، رگرسیون ریج است که توسط هورل و کنارد[۱۸] (۱۹۷۰) معرفی شد. برآورد ریج به‌صورت زیر به‌دست می‌آید:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\text{ridge}} &= \text{argmin}_{\boldsymbol{\beta}}\left\{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}\beta_j^2\right\} \\ &= (\boldsymbol{X}^\top\boldsymbol{X}+\lambda I)^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \end{aligned} \tag{۱۲.۱}$$

که در آن λ پارامتر تنظیم‌کننده است. اگر $\lambda = 0$ باشد، برآورد ریج به برآورد OLS تبدیل می‌شود و اگر λ بزرگ شود، برآورد ریج به صفر می‌رسد. هم‌چنین برای هر λ رگرسیون ریج یک مجموعه متفاوت از برآورد ضرایب دارد.

در حالتی که $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{I}_p$، برآوردگر ریج به‌صورت $\frac{1}{1+\lambda}\hat{\boldsymbol{\beta}}^{\text{OLS}}$ است که ویژگی اساسی رگرسیون ریج یعنی انقباض را نشان می‌دهد. به‌کارگیری جریمه‌ی ریج، برآوردها را به سمت صفر منقبض می‌کند که باعث به وجود آمدن اریبی می‌شود اما واریانس برآوردها را کاهش می‌دهد. از طرف دیگر، این تابع جریمه شرط تنکی را فراهم نمی‌سازد.

از طرفی، برآوردگر OLS همیشه وجود ندارد. اگر ماتریس $\boldsymbol{X}^\top\boldsymbol{X}$ رتبه‌ی کامل نباشد، $\boldsymbol{X}^\top\boldsymbol{X}$ معکوس‌پذیر نیست و پاسخ یکتا برای OLS وجود ندارد. اما برآوردگر ریج همیشه وجود دارد

[۱۸] Hoerl and Kennard

زیرا برای هر ماتریس $\boldsymbol{X}$، کمیت $\boldsymbol{X}^\top \boldsymbol{X} + \lambda I_p$ همیشه معکوس‌پذیر است. همین‌طور برای هر $\boldsymbol{\beta}$، همیشه λای وجود دارد که MSE برآوردگر ریج کمتر از MSE برآوردگر OLS است که این نتیجه بسیار شگفت‌انگیز است زیرا بیان می‌کند که حتی اگر مدل برازش‌شده کاملا درست باشد و از توزیع دقیق مشخص‌شده پیروی کند، همیشه می‌توان برآوردگر بهتری را با منقبض کردن به سمت صفر به دست آورد.

اگر چه که برآوردگر ریج کمک می‌کند تا برآوردهایی با تغییرپذیری کمتر به‌دست آید، اما دارای نقص‌های زیر است:

- ضرایب با مقادیر واقعی بزرگتر را نیز به سمت صفر منقبض می‌کند.
- مشکل تفسیرپذیری مدل: ضرایب بی‌اهمیت مدل ممکن است به صفر منقبض شوند اما آن‌ها هنوز در مدل وجود دارند.

۲.۳.۱ برآورد لاسو

تیبشیرانی (۱۹۹۶) پیشنهاد کرد از تابع جریمه‌ی $\lambda \sum_{j=۱}^{p} |\beta_j|$ در رابطه‌ی (۷.۱) استفاده شود. در نتیجه، این برآوردگر به‌صورت زیر به‌دست می‌آید:

$$\hat{\beta}^{LASSO} = \text{argmin}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=۱}^{p} |\beta_j| \right\} \qquad (۱۳.۱)$$

تنها تفاوت برآوردگر لاسو با برآوردگر ریج، استفاده از تابع قدرمطلق به جای تابع توان دوم به‌عنوان جریمه است. اگر چه این تغییر بسیار جزیی است، اما تاثیر چشم‌گیری بر برآوردگر نتیجه‌شده دارد.

مشابه برآوردگر ریج، جریمه کردن مقادیر قدرمطلق ضرایب، آن‌ها را به سمت صفر منقبض می‌کند اما با این وجود، بر خلاف برآوردگر ریج، برخی از ضرایب را دقیقا برابر صفر قرار می‌دهد. لذا این تابع جریمه، انتخاب متغیر نیز انجام می‌دهد. برآوردگر به‌دست آمده، لاسو[۱۹] نامیده شده و با نماد $\hat{\beta}^{\text{LASSO}}$ نشان داده می‌شود. برخلاف روش ریج، لاسو صورت بسته‌ای ندارد و برای هر مقدار λ بردار ضرایب به صورت عددی به دست می‌آید.

[۱۹] LASSO:least absolute shrinkage and selction operator

روش لاسو به لحاظ پایداری و دقت پیش‌بینی برآوردگرها عملکرد قابل قبولی از خود نشان داده است و نسبت به رگرسیون ریج اکثر مواقع دقت پیش‌بینی بالایی دارد و هم‌چنین، منجر به تولید جواب‌های تنک می‌شود. اما این روش، محاسبات بسیار پیچیده‌ای دارد که افرون [۲۰] (۲۰۰۴) رگرسیون کمترین زاویه[۲۱] را برای محاسبه‌ی برآوردگر لاسو مطرح کردند. در روش لاسو، ضرایب بزرگ نیز جریمه می‌شوند و به سمت صفر منقبض می‌شوند و ضرایب کوچک‌تر فرصت حضور در مدل را پیدا می‌کنند که باعث بیش‌برآوردی می‌شود و در نتیجه برآوردگر لاسو، دارای ویژگی نااریبی نیست. در این راستا، فن و لی (۲۰۰۱) تابع جریمه‌ی اسکد[۲۲] را معرفی کردند. اگر تعداد واقعی متغیرها بیشتر از حجم نمونه باشد، برآوردگر لاسو، اجازه‌ی حضور حداکثر n متغیر را در مدل نهایی می‌دهد که این مشکل به‌وسیله‌ی ژو و هیستی[۲۳] (۲۰۰۵) رفع شده است. مشکل دیگر برآوردگر لاسو، ناسازگاری آن است. ژو (۲۰۰۶) شرط لازم برای سازگاری این برآوردگر را بیان و صورت اصلاح‌شده‌ای از آن را ارایه نمود.

۳.۳.۱ برآورد الاستیک‌نت

همان‌طور که اشاره شد ژو و هیستی (۲۰۰۵) برآوردگر الاستیک‌نت[۲۴] را در راستای برطرف کردن مشکل عدم حضور بیش از n متغیر در مدل مطرح کردند. برآورد الاستیک‌نت خام[۲۵] (NEN) به‌صورت زیر محاسبه می‌شود:

$$\hat{\boldsymbol{\beta}}^{\text{NEN}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\} \qquad (۱۴.۱)$$

که در آن λ_1 و λ_2 مقادیری نامنفی اختیار می‌کنند. با در نظر گرفتن $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$ می‌توان برآورد الاستیک‌نت اولیه را به صورت زیر نوشت:

$$\hat{\boldsymbol{\beta}}^{\text{NEN}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + (1 - \alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2 \right\}. \qquad (۱۵.۱)$$

ناحیه‌ی جریمه در این روش، ترکیب محدبی از نواحی جریمه‌ی روش‌های لاسو و ریج است. وقتی $\alpha = 1$ الاستیک‌نت تبدیل به رگرسیون ریج ساده می‌شود. همین‌طور، رگرسیون لاسو

[۲۰] Efron.
[۲۱] Least angle regression
[۲۲] SCAD: Smoothly clipped absolute deviation
[۲۳] Zou and Hastie
[۲۴] ElasticNet
[۲۵] Naive elastic net estimator

به‌ازای $\alpha = ۰$ به‌دست می‌آید. این تابع جریمه، قادر است برخی از ضرایب را صفر برآورد کند. هم‌چنین، می‌تواند متغیرهایی را که بر متغیر پاسخ اثر یکسانی دارند یا به شدت همبسته هستند، برابر برآورد کند. بنابراین از این روش، برای گروه‌بندی متغیرهای مستقل به‌خصوص زمانی که تعداد آن‌ها بزرگ است، استفاده شده است. برآورد اصلاح‌شده‌ی الاستیک‌نت، $\hat{\beta}^{\text{EN}}$، که دارای دقت پیش‌بینی بالاتری نسبت به برآورد خام است، به‌صورت زیر تعریف می‌شود:

$$\hat{\beta}^{\text{EN}} = (۱ + \lambda_۲)\hat{\beta}^{\text{NEN}}. \qquad (۱۶.۱)$$

این روش، از نظر دقت پیش‌بینی، انتخاب مدل صحیح و تمایز بین متغیرهای موثر و غیرموثر همبسته، از دیگر روش‌های بیان‌شده بهتر عمل می‌کند. برای مدل‌هایی با تعداد متغیرهای متوسط یا کم، روش الاستیک‌نت در حالتی که همبستگی بین پیش‌بین‌ها زیاد باشد و هم در حالتی که همبستگی چندان قابل توجه نباشد، عملکرد قابل قبولی از خود نشان داده است. البته در حالتی که همبستگی بین متغیرهای مستقل قوی نباشد، می‌توان از روش‌های دیگری مانند لاسو استفاده کرد.

۴.۳.۱ انتخاب پارامتر تنظیم‌کننده در رگرسیون جریمه‌شده

بدیهی است که روش‌های جریمه‌شده، شدیدا به پارامتر تنظیم‌کننده‌ی λ وابسته هستند. این پارامتر، موازنه بین برازش و جریمه را کنترل می‌کند.

- وقتی λ بسیار کوچک باشد، داده‌ها بیش‌برازش می‌شوند و در نتیجه مدلی تولید می‌شود که واریانس بزرگی دارد.
- وقتی λ خیلی بزرگ باشد، داده‌ها کم‌برازش می‌شوند و منجر به مدل‌هایی می‌شود که اریبی بزرگی دارند.

بنابراین پارامتر λ یک پارامتر تنظیم‌کننده است که تغییر پارامتر تنظیم‌کننده اجازه می‌دهد تا موازنه‌ی بین اریبی و واریانس به تعادل برسد. بدیهی است که انتخاب این پارامتر، اولین قدم در استفاده از هر تابع جریمه‌ای است.

بر خلاف معیارهای اطلاع، جریمه‌ی بعد پارامترها، یک مقدار ثابت و از پیش تعیین‌شده است، مقدار پارامتر تنظیم‌کننده در رگرسیون جریمه‌شده بر اساس مشاهدات به‌دست می‌آید.

روش‌های مختلفی برای یافتن این پارامتر وجود دارند. اما روشی که بیشترین مورد استفاده را دارد، بر مبنای این است که پیش‌بینی‌ها بر اساس $\hat{\beta}_\lambda$ (برآوردگر به ازای λ) تا چه اندازه می‌تواند مقدار واقعی Y را بازیابی کند. اگر از داده‌ها هم برای برازش مدل و هم برای برآورد دقت پیش‌گویی استفاده شود، آن‌گاه منجر به بیش‌برازشی خواهد شد. یک ایده به این صورت مطرح می‌شود که مجموعه‌ی داده‌ها به دو قسمت تقسیم شود، یک قسمت از آن برای برازش $\hat{\beta}$ و دیگری برای ارزیابی اینکه $X\hat{\beta}$ پیش‌گویی‌شده تا چه‌اندازه می‌تواند مشاهدات را در قسمت دوم پیش‌گویی کند. مسئله اساسی در این روش آن است که به‌ندرت تعداد مشاهدات آن قدر زیاد است که بتوان به راحتی آن‌ها را به دو قسمت تقسیم کرد که بتوان از نیمی از آن‌ها به‌تنهایی برای برآورد λ استفاده کرد.

روش پیشنهادی جایگزین، اعتبارسنجی متقابل[۲۶] است که داده‌ها را به K قسمت تقسیم می‌کند؛ مدل به $K-1$ زیرمجموعه برازش داده می‌شود و سپس مخاطره روی قسمتی که کنار گذاشته شده است، محاسبه می‌گردد. این رویه برای هر قسمت تکرار می‌شود و در نهایت، میانگین مخاطره‌ها روی همه‌ی این نتایج محاسبه می‌شود. معمولا فرض می‌شود K برابر با ۵، ۱۰ یا n است. اگر $K=n$ ، بدین معنی است که تنها یکی از مشاهدات کنار گذاشته می‌شود. در این حالت، این معیار به‌صورت زیر تعریف می‌شود:

$$\mathrm{CV}(\lambda)=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_\lambda^{-i})^2 \qquad (۱۷.۱)$$

که در آن، $\hat{Y}_\lambda^{-i}$ بیانگر مقدار برازش‌شده پس از کنار گذاشتن iامین مشاهده است. در این صورت، λ مورد نظر متناظر با λیی خواهد بود که این معیار را کمینه کند.

محاسبه‌ی $\mathrm{CV}(\lambda)$ بسیار وقت‌گیر است. کروان و واهبا[۲۷] (۱۹۷۹) معیار اعتبارسنجی متقابل تعمیم‌یافته[۲۸] (GCV) را به عنوان جایگزینی برای $\mathrm{CV}(\lambda)$ به‌صورت زیر معرفی کردند:

$$\mathrm{GCV}(\lambda)=\frac{(\boldsymbol{Y}-\hat{\boldsymbol{Y}}_\lambda)^\top(\boldsymbol{Y}-\hat{\boldsymbol{Y}}_\lambda)}{n\left(1-\frac{\mathrm{tr}(\boldsymbol{A}(\lambda))}{n}\right)^2} \qquad (۱۸.۱)$$

که در آن، $\hat{Y}_\lambda$ بردار مقادیر برازش شده تحت ماتریس تصویر[۲۹] $\boldsymbol{A}(\lambda)$ است.

[۲۶] Cross validation
[۲۷] Craven and Wahba
[۲۸] Generalized cross validation (GCV)
[۲۹] Projection matrix

رگرسیون تقریبا در هر زمینه‌ای از جمله مهندسی، فیزیک، اقتصاد، مدیریت، علوم زیستی، بیولوژی و علوم اجتماعی برای برآورد و پیش‌بینی مورد نیاز است. معروفترین رگرسیون در علوم زیستی و پزشکی، رگرسیون مخاطره متناسب کاکس است. از آن‌جا که این رگرسیون با داده‌های خاصی به نام «داده‌های بقا» سر و کار دارد، در ادامه مفاهیم اساسی مربوط به این داده‌ها که شاخه‌ای از علم آمار به نام «تحلیل بقا» را می‌سازند، بررسی می‌گردد.

۴.۱ تحلیل بقا

برخی از انواع داده‌ها که مورد توجه محققین قرار گرفته است، فاصله‌ی زمانی تا وقوع بعضی حوادث مانند مرگ می‌باشند. یعنی پرداختن و توجه نمودن به گروهی از افراد به‌طوری که پس از مدتی برای هر کدام از آن‌ها یک نقطه‌ی شکست یا از کار افتادگی تعریف می‌گردد. از جمله مواردی که می‌تواند مصداق شکست یا واقعه‌ی مورد نظر باشد، طول عمر یک ماشین صنعتی و اولین زمان مراجعه یک اتومبیل نو به تعمیرگاه می‌باشند. از آن‌جایی که این روش‌ها در ابتدا برای مطالعات مرگ و میر طراحی شده‌اند، نام «تحلیل بقا» بر آن‌ها نهاده شده است.

مطالعه‌ای را در نظر بگیرید که هدف از آن بررسی اثر پرتو درمانی در کاهش خطر پیشروی سرطان باشد. فرض کنید این مطالعه بدین صورت انجام شود که از ابتدای سال ۱۳۸۵ تا انتهای سال ۱۳۸۶ تمام بیمارانی که در بیمارستان تحت پرتو درمانی قرار می‌گیرند، پیگیری شده‌اند و تمام موارد پیشروی سرطان ثبت شده است. مطالعه‌ی دیگری را در نظر بگیرید که با هدف مشابهی انجام شده، اما به جای آن که تا پایان سال ۱۳۸۶ پیگیری شده باشند تا پایان عمرشان در مطالعه حضور داشته باشند. در چنین حالتی تحلیل داده‌ها آسان است. شما در مورد هر بیماری دقیقا می‌دانید که آیا تا پایان عمرش دچار پیشروی سرطان شده است یا خیر و نیز، می‌توانید درصد پیشروی در هر گروه را گزارش و مقایسه کنید. حتی می‌توانید به دقت تعیین کنید که فاصله‌ی زمانی بین عدم و وقوع پیشروی، در هر یک از دو گروه مطالعه، به‌طور متوسط چه‌قدر بوده است. اما مشکلی که در مورد این مطالعه وجود دارد آن است که عملاً اجرای آن بسیار دشوار است. مثلاً از آن جایی‌که همه بیماران باید تا پایان عمرشان پیگیری شوند، ممکن است مطالعه ۳۰ سال یا بیشتر ادامه پیدا کند.

مطالعه اول از لحاظ اجرایی منطقی‌تر است، اما در پایان این مطالعه همه داده‌هایی که در مطالعه دوم در اختیار پژوهشگر بود، به‌دست نیامده‌اند. مثلاً چون بیماران فقط تا پایان سال ۱۳۸۶ پیگیری شده‌اند، اگر بیماری در فروردین ۱۳۸۵ عدم پیشروی داشته باشد و تا پایان مطالعه پیشروی نداشته باشد، می‌توان گفت به مدت دو سال پیشروی در وی رخ نداده، اما بیماری که در اسفند ۱۳۸۶ وارد مطالعه می‌شود، فقط یک ماه پیگیری می‌شود و نمی‌توان گفت که او نیز تا دو سال دچار پیشروی نخواهد شد؛ چه بسا او در فروردین ۱۳۸۷ دچار پیشروی شود.

در این پدیده‌ها، زمان رخداد از اهمیت بسزایی برخوردار است، و از آنجایی که مطالعات در پزشکی برای بررسی زنده ماندن یا مردن افراد بود از آن به عنوان زمان بقا یا در حالت کلی داده‌های بقا یاد می‌کنند. در حالت کلی، «تجزیه و تحلیل بقا» از نظر علم آمار مجموعه‌ای از روش‌های متفاوت آماری در تحلیل متغیرهای تصادفی نامنفی است که مقدار آن‌ها می‌تواند مدت زمان انتظار تا شکست (از کار افتادگی) یک مولفه‌ی فیزیکی یا زمان مرگ یک واحد بیولوژیک (انسان، حیوان، یا سلول زنده) و به‌طور کلی مدت زمان مورد انتظار تا رخداد پیشامد مورد نظر باشد. به عنوان مثال، این پیشامد در بیماری سرطان، مرگ یا رویداد مجدد بیماری است. در مطالعات پرستاری، مرخص کردن بیمار از بیمارستان و در پیوند اعضا، رد عضو پیوندی است.

اولین بار تحلیل بقا در سال ۱۶۶۲ میلادی توسط شخصی به‌نام جان گرونت[۳۰] به‌صورت کارهای ابتدایی بر روی جدول‌های مرگ و میر به‌کار برده شده است. وی اولین فهرست مرگ و میر در لندن را منتشر کرد. بعدها در یادداشت‌های ادموند هالی[۳۱]، ستاره‌شناس معروف (کاشف ستاره دنباله‌دار هالی) در سال ۱۶۹۳ نیز جدول‌هایی مشاهده شد که نشان می‌داد چند درصد از آدم‌ها در هر دوره سنی می‌میرند. سپس دنیل برنولی[۳۲] در سال ۱۷۶۰ زمانی‌که زمان مورد انتظار زندگی را محاسبه کرد، این جدول‌ها را گسترش داد (فریدمن[۳۳]، ۲۰۰۸). اخیرا با پیشرفت تکنولوژی، دوره نوین به کارگیری این روش در رشته‌های مختلف به‌ویژه مهندسی آغاز شده

[۳۰] John Graunt
[۳۱] Edmond Halley
[۳۲] Daniel Bernoulli
[۳۳] Freedman

است که بنا به اهمیت آن نام «قابلیت اطمینان» را برای آن استفاده می‌کنند. در سایر علوم نیز از این روش‌ها بهره می‌گیرند. مثلا در جامعه‌شناسی برای تحلیل رویدادهای تاریخی یا در صنایع برای پیش‌بینی زمان از کار افتادن یک دستگاه به‌کار می‌رود اما در سراسر این پایان‌نامه، از همان نام «تحلیل بقا» استفاده می‌شود.
تحلیل بقا به روش‌های تحلیلی زمان رخ دادن پیشامدها اشاره دارد.

تعریف ۱.۴.۱ (زمان بقا). در اغلب پژوهش‌های پزشکی و جمعیت‌شناختی، متغیر مورد توجه، نقطه‌ی زمانی است که در آن پیشامد مورد نظر رخ می‌دهد. این نقطه‌ی زمانی را زمان بقا، زمان از کارفتادگی یا زمان شکست گویند.

برای شروع تحلیل و جمع‌آوری داده ابتدا باید سه مولفه‌ی زیر به‌طور واضح و دقیق تعریف شود و هیچ‌گونه ابهامی برای آن‌ها وجود نداشته باشد.

(۱) مبدا زمان و نیز پایان مطالعه: یعنی زمان ورود به مطالعه‌ی واحد مورد پژوهش و نیز زمانی که تا آن موقع تحت مطالعه بوده است، مشخص باشد. لزومی ندارد زمان شروع برای تمام افراد یا مولفه‌ها، مقدار مشخص ثابتی باشد ولی باید زمان ورود به مطالعه و زمان رخ دادن پیشامد مورد انتظار برای آن‌ها دقیقا ثبت شود. در این صورت زمان بقا یا شکست، فاصله‌ی بین نقطه‌ی زمانی ورود به مطالعه و زمان رخ دادن پیشامد مورد انتظار است.

(۲) مقیاس و واحد اندازه‌گیری گذشت زمان: باید در مورد کلیه‌ی واحدهای تحت مطالعه برای اندازه‌گیری گذشت زمان مقیاس واحدی مورد استفاده قرار گیرد.

(۳) مفهوم شکست یا وقوع پیشامد: یعنی در ابتدای مطالعه دقیقا و بدون هیچ ابهامی، پیشامد مورد انتظار تعریف شود.

به عنوان مثال، پیشامد مورد نظر می‌تواند طول عمر از تولد (نقطه‌ی شروع یا مبدا زمان) تا مرگ باشد و بر حسب سال اندازه‌گیری شود. زمان‌های بقا بر حسب ماه از زمان تشخیص تا مرگ در بیماران هموفیلی زیر پنجاه سال که به ویروس ایدز آلوده شده‌اند یا زمان‌های بهبودی بر حسب

هفته از ارزیابی توانایی یک داروی جدید در حفظ بهبودی بیماران مبتلا به سرطان حاد خونی در مدت مطالعه یک سال، مثال‌های دیگری است که می‌توان به آن‌ها اشاره کرد.

وقتی بیماری تا پایان مطالعه پیگیری شود و پیشامد مورد نظر در وی رخ ندهد یا به هر دلیلی از مطالعه خارج شود، در مطالعات تحلیل بقا به این داده اصطلاحاً سانسورشده[۳۴] گفته می‌شود. وجود داده‌های سانسور شده سبب تمایز تحلیل بقا از سایر تحلیل‌های آماری شده است. هدف مطالعه تحلیل بقا آن است که با وجود داده‌های سانسورشده، بتواند تخمینی از شاخص‌های مرتبط با زمان را به دست بیاورد؛ مثلاً در مطالعه‌ی مطرح‌شده در ابتدای این بخش، هدف آن است که بتوان تخمینی از متوسط عمر عدم پیشروی در بیماران، یا ریسک (خطر) پیشروی بعد از دو سال به‌دست آورد.

۱.۴.۱ داده‌های سانسورشده

مسئله مهمی که همواره در مطالعات تحلیل بقا مطرح می‌شود، داده‌های سانسور و گم‌شده است، که به داده‌هایی اطلاق می‌شود که در طول مدت پیگیری از دست رفته یا گم‌شده اند. در طبیعت گاه نمی‌توان همه داده‌های موجود را (مثلا برای بررسی طول عمر موجودات و حتی لوازم و اشیا از قبیل وسایل برقی) مورد بررسی قرار داد.

برای درک بهتر، فرض کنید می‌خواهید طول عمر لامپ‌ها را بررسی کنید اما زمان لازم برای این کار را ندارید. مثلا تا ۴۰ روز منتظر می‌مانید اما تعداد لامپ‌هایی که می‌سوزند اندک هستند. به دلیل محدود بودن زمان مطالعه بقیه لامپ‌ها را مورد آزمایش قرار نمی‌دهید و استنباط‌های خود را بر اساس همین لامپ‌های موجود انجام می‌دهید. به این داده‌ها، داده‌های سانسورشده می‌گویند.

به‌طور کلی چندین دلیل برای رخ‌داد سانسور ممکن است وجود داشته باشد که به چهار مورد اشاره می‌شود:

(۱) فرد مورد پژوهش، رویداد را قبل از پایان مطالعه تجربه نکند (در مثال بالا، هیچ کدام از لامپ‌ها نسوزند).

[۳۴] Censored

(۲) فرد مورد پژوهش برای پیگیری در طول مطالعه مفقود شود.

(۳) فرد مورد پژوهش از مطالعه صرف‌نظر کند.

(۴) فرد مورد پژوهش، بعد از شروع مطالعه و گذشت مدت زمانی مورد توجه قرار گیرد (مثلا یک لامپ را بعد از گذشت ۱۰ روز به مجموعه خود اضافه کنید).

قبل از شروع آزمایش با توجه به پیچیدگی‌های داده‌ها در تحلیل بقا باید طرح مشخصی از ابتدا برای جمع‌آوری داده‌ها در نظر گرفت. مثلا فرض کنید بخواهید n داده را جمع آوری کنید. مهم‌ترین طرح‌هایی که وجود دارند، عبارتند از:

(۱) طرح سانسور نوع اول (سانسور زمان):

فرض کنید می‌خواهید آزمایش را در زمان از قبل مشخص‌شده t به‌پایان برسانید، به طوری‌که زمان رخداد رویداد، که قبل از t از بین رفته‌اند یادداشت می‌شود. داده‌هایی که به این ترتیب به‌دست می‌آیند یک نمونه سانسور شده نوع اول را می‌سازد.

(۲) سانسور نوع دوم (سانسور شکست):

اگر زمان از قبل تعیین‌شده‌ای برای انتهای آزمایش در نظر گرفته شود و پژوهش تا از کار افتادن rامین واحد ادامه پیدا کند ($1 \leq r \leq n$)، نمونه سانسورشده نوع دوم به دست می‌آید.

(۳) سانسور تصادفی:

فرض کنید به واحد iام شکست، متغیر تصادفی زمان شکست C_i نسبت داده شود. اگر $Y_i = \min(T_i, C_i)$ و به‌ازای $Y_i = T_i$، $\delta_i = 1$ زمان بقا را نشان می‌دهد و در غیر این صورت، $\delta_i = 0$ که $Y_i = C_i$ زمان سانسور است. توجه کنید که Y_iها زمان بقا (شکست) یا زمان سانسور هستند. این طرح به نام طرح سانسور تصادفی شناخته می‌شود و در آزمایش‌های کلینیکی خیلی متداول است که در این پایان‌نامه داده‌ها براساس سانسور تصادفی جمع‌آوری شده‌اند.

طرح‌های دیگری مانند سانسورهای فزاینده وجود دارند که به انواع دیگری مانند سانسور فزاینده نوع دوم از راست، سانسور فزاینده نوع دوم کلی، سانسور فزاینده نوع اول از راست، سانسور

فزاینده نوع اول کلی و سانسور فزاینده نوع دوم دورگه[۳۵] تقسیم می‌شوند.

متغیر تصادفی نامنفی T که نقطه‌ی شروع و مقیاس اندازه‌گیری آن مشخص است، زمان بقا را نشان می‌دهد. فرض کنید $F(t)$ تابع توزیع تجمعی متغیر تصادفی T باشد. این متغیر ممکن است گسسته، پیوسته یا ترکیبی از این دو باشد. تابع توزیع احتمال را می‌توان به یکی از روش‌های (۱) تابع چگالی احتمال[۳۶]، (۲) تابع بقا[۳۷] و (۳) تابع مخاطره[۳۸] در تحلیل بقا محاسبه کرد.

تعریف ۲.۴.۱ (تابع چگالی احتمال). تابع چگالی احتمال T با نماد $f(t)$ نمایش داده می‌شود و برابر است با:

$$f(t) = \lim_{ht \to ۰} \frac{P(t \leq T < h + t)}{h} = \frac{dF(t)}{dt}. \qquad (۱۹.۱)$$

تعریف ۳.۴.۱ (تابع بقا). تابع بقا با $S(t)$ نشان داده می‌شود و برابر است با احتمال اینکه مدت زمان بقای یک فرد یا مؤلفه بیشتر از t باشد، بدین معنی که فرد بیش از زمان از قبل مشخص شده، زنده باشد. تابع بقا یک مفهوم اساسی برای تحلیل بقا می‌باشد چرا که به‌دست آوردن احتمال بقا برای مقادیر مختلف زمان t اطلاعات خلاصه حیاتی درباره آن فرد از داده‌های بقا را فراهم می‌کند؛ تابع بقا در مورد توزیع‌های گسسته و پیوسته به‌صورت زیر تعریف می‌شود:

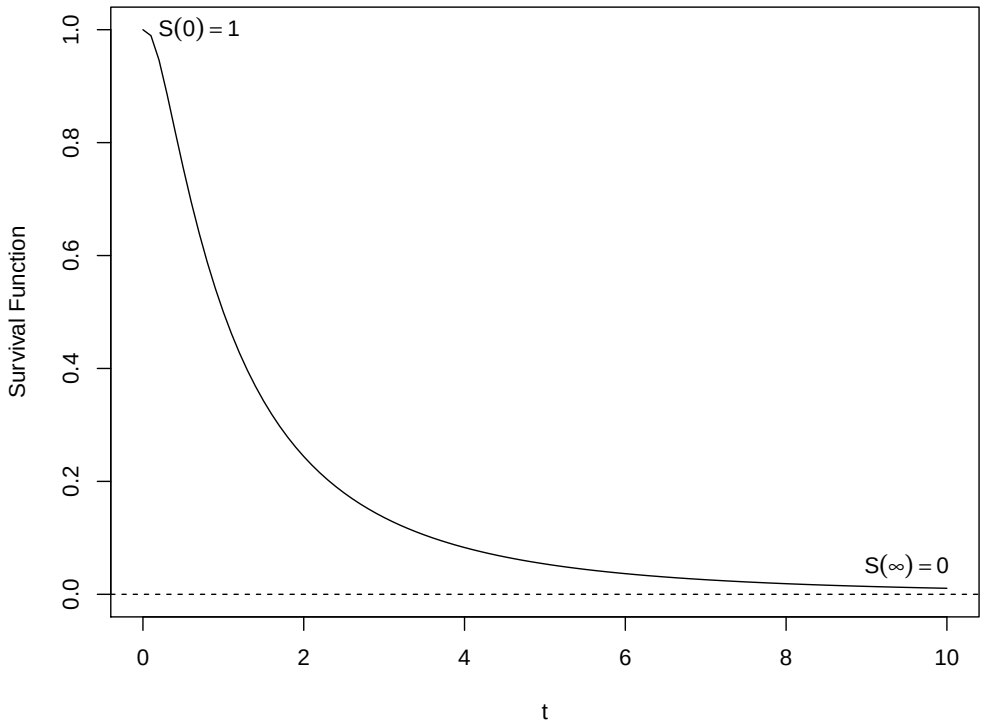
$$S(t) = Pr(T > t) = ۱ - Pr(T \leq t) = ۱ - F_T(t) \qquad (۲۰.۱)$$

با توجه به ویژگی‌های تابع توزیع تجمعی، ویژگی‌های تابع بقا به‌صورت زیر است.

توجه کنید که مدت زمان بقا، T، مقادیر نامنفی اختیار می‌کند ($F(-\infty) = ۰$) و در نتیجه، $S(۰) = ۱ - F(-\infty) = ۱$. هم‌چنین $S(\infty) = ۱ - F(\infty) = ۰$. بنابراین، در اولین لحظه، تمام واحدها زنده هستند و بالاخره به پایان می‌رسند (مرگ یا شکست اتفاق می‌افتد) و زمان شکست بی‌نهایت وجود ندارد. بدیهی است که $S(t)$ یک تابع غیرصعودی و از چپ پیوسته است. شکل ۳.۱، منحنی تابع بقا را با در نظر گرفتن توزیع لگ‌نرمال برای داده‌ها نشان می‌دهد.

---

[۳۵] Hybrid
[۳۶] Probability density function
[۳۷] Survival function
[۳۸] Hazard function



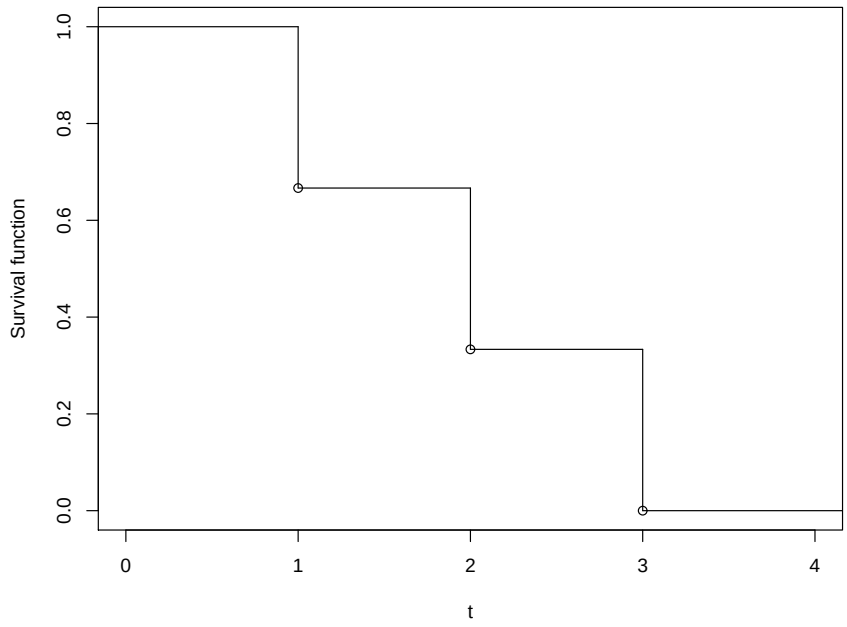
شکل ۳.۱: منحنی تابع بقا برای توزیع لگ‌نرمال

در عمل، در هنگام استفاده از داده‌های واقعی، معمولا نمودار به‌دست آمده، پله‌ای است، علاوه بر این، به دلیل اینکه طول دوره مطالعه، در واقعیت بی‌نهایت نیست و هم‌چنین در هر زمانی ممکن است مولفه‌هایی در معرض شکست (یا مرگ) باشند، این احتمال وجود دارد که هیچ فردی پیشامد مورد نظر را تجربه نکند. قضیه‌ی ۱.۴.۱ اهمیت منحنی بقا را نشان می‌دهد.

قضیه ۱.۴.۱. (احراری خلف، ۱۳۹۰) فرض کنید متغیر تصادفی T دارای تابع چگالی $f(\cdot)$ و تابع توزیع $F(\cdot)$ است. اگر μ میانگین این توزیع باشد، آن‌گاه می‌توان نشان داد سطح زیر منحنی بقا برابر μ است.

اگر T یک متغیر تصادفی گسسته باشد که مقادیر t_1، t_2، ... را اختیار کند، آن‌گاه تابع جرم احتمال آن به‌صورت زیر است:

$$P(t_i) = P(T = t_i), \quad i = ۱, ۲, \ldots. \qquad (۲۱.۱)$$



شکل ۴.۱: منحنی تابع بقا برای مثال ۱.۴.۱

تابع بقا برای این متغیر تصادفی گسسته با رابطه‌ی زیر تعریف می‌شود.

$$S(t) = P(T > t) = \sum_{i|t_i>t} P(t_i). \qquad (۲۲.۱)$$

مثال ۱.۴.۱. فرض کنید T دارای تابع جرم احتمال زیر باشد:

$$P(t_i) = P(T = i) = \frac{۱}{۳}, \quad i = ۱, ۲, ۳ \qquad (۲۳.۱)$$

آنگاه تابع بقا مربوط به آن به‌صورت زیر می‌باشد:

$$S(t) = P(T > t) = \sum_{i|t_i>t} P(t_i) = \begin{cases} ۱ & ۰ \leq t < ۱ \\ \frac{۲}{۳} & ۱ \leq t < ۲ \\ \frac{۱}{۳} & ۲ \leq t < ۳ \\ ۰ & t \geq ۳ \end{cases} \qquad (۲۴.۱)$$

شکل ۴.۱ تابع بقا مربوط به مثال ۱.۴.۱ را نشان می‌دهد.

تابع بقا را می‌توان با یکی از دو روش زیر محاسبه و برآورد نمود.

(۱) روش کاپلان_مه‌یر[۳۹] (KM): این روش برای نمونه‌های کوچک‌تر که زمان وقوع رخداد

[۳۹] Kaplan-Meier method

به‌دقت ثبت و اندازه‌گیری می‌شود، بسیار مفید است. در این روش، تعداد کل افراد مورد مطالعه که تحت پیگیری و مراقبت قرار گرفته‌اند، n نفر و زمان‌های مشاهده‌شده به‌صورت $t_۱$، $t_۲$، ... و t_n است. زمان بقا بعضی از این افراد ممکن است ثبت نشده باشد. همچنین ممکن است بیش از یک مشاهده برای هریک از زمان‌های بقا وجود داشته باشد.

فرض کنید تعداد افراد شرکت‌کننده در زمان t_i، n_i نفر بوده که تعداد d_i نفر آنها از بین رفته‌اند. بنابراین احتمال اینکه یک نفر در فاصله زمانی فوق بمیرد، برابر با d_i/n_i و در نتیجه احتمال بقا در فاصله زمانی فوق برابر است با:

$$\hat{S}^{\mathrm{KM}}(t) = \prod_{i=۱}^{n}(۱ - \frac{d_i}{n_i}). \qquad (۲۵.۱)$$

(۲) روش جدول طول عمر[۴۰]: وقتی که تعداد مشاهدات و بیماران مورد بررسی زیاد باشند، احتمال وجود زمان‌های بقای یکسان قوت می‌گیرد. بدین معنا که ممکن است بیش از یک پیشامد در هر زمان رخ دهد. در این صورت، روش KM جدول‌های طولانی را ایجاد می‌کنند که ارائه و تفسیر آن مطلوب نبوده و وقت‌گیر است. بنابراین روش دیگری را به نام روش جدول طول عمر استفاده می‌کنند که در آن زمان وقوع پیشامدها را به‌صورت فاصله‌های زمانی تقسیم‌بندی می‌کنند. این فاصله‌ها به دلخواه می‌تواند کم و زیاد گردد و الزامی به تساوی فاصله‌ی بین گروه‌ها نیست. هر چند مساوی بودن آن‌ها متداول است.

فرض کنید تعداد افراد شرکت‌کننده در فاصله زمانی iام، n_i نفر بوده که تعداد d_i نفر آن‌ها از بین رفته‌اند. احتمال مرگ برابر $q_i = d_i/n_i$ است لذا احتمال بقا برابر $۱ - q_i$ است و بنابراین برآورد تابع بقا عبارت است از:

$$\hat{S}(t) = \prod_{i=۱}^{n}(۱ - q_i) = \prod_{i=۱}^{n}\frac{n_i - d_i}{n_i} \qquad (۲۶.۱)$$

تعریف ۴.۴.۱ (تابع مخاطره). تابع مخاطره، خطر آنی مرگ را در زمان t به شرط اینکه تا آن زمان زنده باشد را نشان می‌دهد و به‌صورت زیر تعریف می‌شود:

$$h(t) = \lim_{\Delta t \to ۰}\frac{P(t \le T < \Delta t + t | T \ge t)}{\Delta t}. \qquad (۲۷.۱)$$

[۴۰] Life time table method

برای یک مقدار مشخص از t، تابع مخاطره $h(t)$ دارای مشخصات زیر است:

(۱) $h(t) \geq ۰$

(۲) $h(t)$ کران بالا ندارد.

هدف از معرفی توابع مطرح شده، برآورد تابع توزیع احتمال T می‌باشد زیرا اطلاعات در مورد احتمال زنده بودن فرد و همچنین رفتار داده‌ها در طول مطالعه مورد نیاز است؛ همچنین می‌توان تابع بقا را از تابع مخاطره (و بالعکس) طبق روابط زیر به‌دست آورد

$$S(t) = \exp\left[-\int_{۰}^{t} h(u)du\right] \qquad (۲۸.۱)$$

$$h(t) = -\left[\frac{S'(t)}{S(t)}\right] = -d\ln S(t) \qquad (۲۹.۱)$$

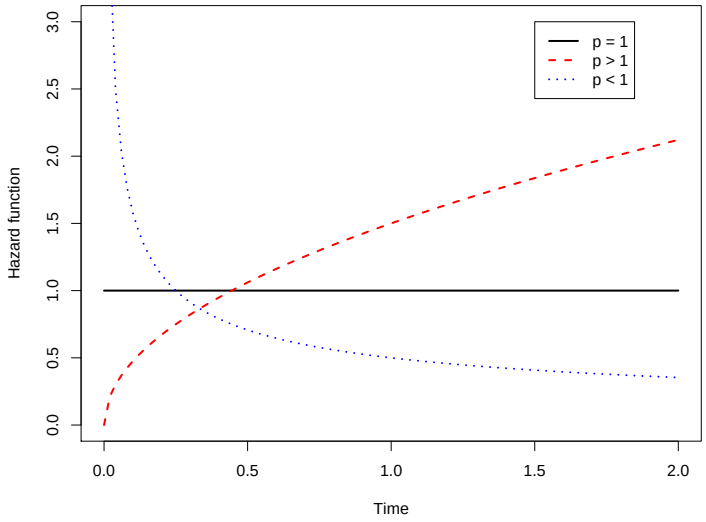
نکته جالب توجه این است که وقتی تابع مخاطره در دسترس باشد، می‌توان تابع بقا را به‌طور مستقیم از آن به دست آورد.

چند نمونه از تابع مخاطره به شکل زیر می‌باشند:

(۱) تابع خطر ثابت یا نمایی: در این حالت ، در یک مطالعه از سلامتی اشخاص، خطر مرگ ثابت است و بدون توجه به مقدار زمان t از قبل مشخص‌شده، $h(t)$ مقدار یکسانی را اختیار می‌کند. همچنین مخاطره آنی برای بیمار شدن فرد در طول دوره پیگیری ثابت است. هنگامی که تابع خطر ثابت باشد مدل را بقا نمایی گویند.

برای مثال، فرض کنید $h(t) = \lambda$ تابع خطر ثابت باشد آن‌گاه $S(t) = \exp(-\lambda t)$، تابع توزیع تجمعی $F(t) = ۱ - \exp(-\lambda t)$ و بنابراین T دارای توزیع نمایی است.

(۲) تابع خطر وایبل $h(t) = p\lambda^{p}t^{p-۱}$: برای این نوع از تابع خطر با توجه به پارامتر شکل (p) سه حالت ممکن است رخ دهد: (۱) $p = ۱$. در این حالت تابع خطر ثابت است. (۲) $p > ۱$. در این حالت تابع خطر در طول زمان افزایش می‌یابد که به آن، تابع خطر وایبل افزایشی می‌گویند و برای بیماران سرطانی که در آنها رویداد مورد بررسی مرگ است به‌کار برده می‌شود. همان‌طور که سن بیمار افزایش پیدا می‌کند به عنوان پیش‌آگاهی بد بر این اساس، خطر آنی برای مرگ بیمار نیز افزایش می‌یابد. (۳) $p < ۱$. در این حالت تابع



شکل ۵.۱: منحنی تابع خطر وایبل

خطر در طول زمان در حال کاهش است که به تابع خطر وایبل کاهشی معروف است. در این حالت انتظار می‌رود که احتمال پیشامد مرگ کاهش یابد. به‌عنوان مثال برای افرادی که عمل جراحی به عنوان درمان پیشنهاد شده است، انتظار می‌رود پس از انجام عمل، خطر مرگ کاهش یابد.

فرض کنید تابع خطر $h(t) = p\lambda^p t^{p-1}$ باشد بنابراین تابع بقا $S(t) = \exp\{-(\lambda t)^p\}$، تابع توزیع تجمعی $F(t) = 1 - \exp\{-(\lambda t)^p\}$ و در نتیجه T دارای توزیع وایبل می‌باشد.

(۳) تابع خطر لگ‌نرمال: این تابع خطر برابر است با

$$h(t) = \frac{\frac{1}{t\sigma}\phi\left(\frac{\ln(t)}{\sigma}\right)}{\Phi\left(\frac{-\ln(t)}{\sigma}\right)}, \quad \sigma > 0, \qquad (۳۰.۱)$$

که در آن، $\phi(0)$ و $\Phi(0)$ به ترتیب تابع چگالی و توزیع تجمعی نرمال استاندارد است. در این حالت، تابع خطر در طول زمان ابتدا افزایش و سپس کاهش می‌یابد. این تابع مخاطره بیشتر برای بیماران مبتلا به سل، به کار می‌رود. در این بیماران، انتظار می‌رود که خطر آنی برای افزایش مرگ در مراحل اولیه افزایش و سپس کاهش پیدا می‌کند.

وقتی تابع خطر، لگ‌نرمال باشد آنگاه تابع‌های بقا و توزیع تجمعی به‌صورت زیر به‌دست

می‌آیند

$$\begin{aligned} S(t) &= ۱ - \Phi\left(\frac{\ln(t)-\mu}{\sigma}\right) \\ F(t) &= \Phi\left(\frac{\ln(t)-\mu}{\sigma}\right) \end{aligned}$$

و در نتیجه T دارای توزیع لگ‌نرمال است.

فصل ۲

مدل رگرسیون نرخ خطر کاکس

۱.۲ مقدمه

هدف بسیاری از مطالعات پزشکی، بررسی توزیع بقای بیماران سرطانی بر اساس ویژگی‌های خاصی است. روش‌های آماری بررسی این توزیع و اختلاف‌های آن می‌تواند دربرگیرنده‌ی مدل‌های مختلفی باشد. مدلی که رگرسیون ارایه می‌دهد تا وقتی معتبر است که گذشت زمان تاثیری در متغیر پاسخ ایجاد نکند، مثلا رگرسیون روش مناسبی برای مدل وزن نوزاد نیست. علی‌رغم اینکه در بحث سری‌های زمانی، زمان به‌صورت یک پارامتر اصلی در نظر گرفته می‌شود در محاسبه و تخمین داده‌های بقا مدل مناسبی ارایه نخواهد داد زیرا

(۱) فرض مانایی سری زمانی در اینجا درنظر گرفته نمی‌شود، مثلا برای طول عمر، بحث استهلاک یا پیری مطرح است.

(۲) داده‌های سانسورشده در بحث سری زمانی وجود ندارند.

مدل اساسی که برای داده‌های بقا به‌کار برده می‌شود، «رگرسیون کاکس» می‌باشد که به مدل

خطرات متناسب کاکس[۱] (CPH) معروف است.

۲.۲ تاریخچه

در سال ۱۹۷۲ کاکس، آماردان معروف انگلیسی، مدلی را ارایه نمود که از آن زمان تا کنون به‌صورت گسترده‌ای استفاده شده است. اگر چه مقاله کاکس (۱۹۷۲)، موفقیتی در تحلیل داده‌های سانسور راست بود که از بحث‌های جالب منتشرشده به همراه مقاله، قابل رویت است اما ممکن است برخی افراد مدل کاکس را تعمیمی از رگرسیون ایده‌های موجود در نوشته منتل[۲] (۱۹۶۶) در نظر بگیرند. حدودا در همان زمان، فیگل و زلن[۳] (۱۹۶۵)، مدل‌های مختلف رگرسیون نمایی را مد نظر قرار دادند. یکی از مدل‌های آن‌ها، معادل مدل کاکس است.

مدل رگرسیون کاکس به سرعت توسط آماردانان کاربردی، به‌خصوص در آزمایشات بالینی، پذیرفته شد. کاربرد آن در زمانی که نرم‌افزار کاربرپسند به سرعت در دسترس قرار گرفت، گسترش یافت. امروزه، فرد به سختی می‌تواند یک مجله پیشرفته آماری یا پزشکی را بدون حداقل یک ارجاع به مقاله‌ی کاکس (۱۹۷۲) پیدا کند. با این وجود، چندین مسئله وجود داشت که جامعه آماری را به چالش می‌کشید. برخی از این مسائل، مانند چگونگی مواجهه با گره‌ها یا هم‌رتبه‌ها (دو یا چند نفر با زمان شکست مشابه)، و مبنای برآوردگر مطرح شده، در جلسه انجمن آمار سلطنتی انگلستان مدنظر قرار گرفتند. کاکس با معرفی مفهوم درستنمایی جزیی مطابقت برآوردگر خودش را ارائه کرد. همان زمان بود که برآوردگرها، کارآمد تشخیص داده شدند. اثبات ویژگی‌های سازگاری و نرمال بودن مجانبی، تقریبا یک دهه طول کشید.

موازی با پیشرفت نظری، تحقیقات روی ایجاد و بررسی مدل، صورت گرفت. هم اکنون انواعی از ابزارها در دسترس هستند که شامل آزمون‌های نیکویی برازش، به‌خوبی باقیمانده‌ها و دیگر مباحث تشخیصی است. اندرسون و همکاران[۴] (۱۹۹۱) در مورد کیفیت ارائه تحلیل‌های رگرسیون در نوشته‌های پژشکی بحث کرده‌اند. علی‌رغم پیشنهادات سازنده‌شان، بخش روش‌های بسیاری از مقالات هنوز هم نسبت به مدل کاکس، چندان آموزنده نیست.

---

[۱] Cox proportional hazards (cph)
[۲] Mantel
[۳] Feigl and Zelen
[۴] Anderson

۳.۲ تعریف مدل

فرض کنید مجموعه داده‌ی $\{(y_1,\delta_1),(y_2,\delta_2),\ldots,(y_n,\delta_n)\}$ در دسترس است که در آن $y_i = \min\{T_i, C_i\}$ و $\delta_i = I(T_i \le C_i)$ به‌طوری که $I(A)$ تابع نشانگر مجموعه‌ی A است. ایده‌ی اولیه‌ی مدل رگرسیونی کاکس، مدل‌بندی تابع خطر به جای استفاده از میانگین در مدل رگرسیونی است. اگر $\boldsymbol{X}_i = (X_{i1},\cdots,X_{ip})^\top$ بردار متغیرهای پیش‌بین باشد، آن‌گاه مدل کلی زیر را می‌توان برای آن نوشت:

$$h(t,\boldsymbol{X}) = h_0(t)h(\boldsymbol{X}\boldsymbol{\beta}) \qquad (۱.۲)$$

که در آن، t وابستگی به زمان را نشان می‌دهد و بر اساس مقدار δ، یکی از مقادیر مشاهده‌شده از متغیرهای T_i یا C_i که جنس هر دو زمان است، می‌پذیرد. $\boldsymbol{X} = (\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)^\top$ ماتریس متغیرهای پیش‌بین و $x_i \in \mathbb{R}^p$، $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^\top$ ضرایب نامعلوم رگرسیونی و $h(\cdot)$ تابعی مثبت و $h_0(t)$ تابع مخاطره پایه و نامعلوم می‌باشد. کاکس در سال ۱۹۷۲، تابع مثبت $h(\boldsymbol{X}^\top\boldsymbol{\beta})$ را به‌صورت نمایی $\exp\{\boldsymbol{X}^\top\boldsymbol{\beta}\}$ پیشنهاد کرد. در این صورت مدل رگرسیون کاکس به‌صورت زیر نوشته می‌شود:

$$h(t,\boldsymbol{X}) = h_0(t)\exp\{\boldsymbol{X}^\top\boldsymbol{\beta}\} \qquad (۲.۲)$$

از آن‌جا که تابع مخاطره مثبت است، می‌توان با لگاریتم گرفتن از رابطه‌ی (۲.۲)، رگرسیون کاکس را به‌صورت زیر نوشت:

$$\ln(h(t,\boldsymbol{X})) = \ln(h_0(t)) + \boldsymbol{X}^\top\boldsymbol{\beta} \qquad (۳.۲)$$

بنابراین، روش کاکس شبیه به رگرسیون چندگانه است با این تفاوت که متغیر وابسته در رگرسیون کاکس، تابع مخاطره است.

۴.۲ برآورد پارامترهای مدل

کاکس در مقاله سال ۱۹۷۵ خود فرض کرده است $h_۰(t)$ اختیاری[۵] باشد. با پذیرفتن این فرض، هیچ اطلاعی درباره‌ی β در فاصله‌های زمانی که در آن سانسور رخ داده است به دست نمی‌آید زیرا امکان دارد همواره $h_۰(t) = ۰$ باشد، بنابراین روی داده‌های سانسورشده شرطی می‌شود.

یک مشاهده، در زمان t در خطر گفته می‌شود اگر تا آن زمان شکست رخ نداده باشد یا سانسور نشده باشد. این مفهوم می‌تواند برای مشاهداتی که در ابتدا وارد مطالعه نشده‌اند تعمیم داده شود. این مشاهدات را سانسور چپ می‌نامند و اغلب زمانی رخ می‌دهند که زمان t دقیقا زمان تشخیص بیماری باشد، بنابراین بیماران در زمان $t_i' > ۰$ وارد مطالعه می‌شوند.

احتمال شرطی را در نظر بگیرید که فردی در زمان t_j شکست بخورد (بمیرد) به شرطی که t_j یکی از n زمان مرگ مشاهده شده‌ی $\{t_۱, \ldots, t_n\}$ باشد. بنابراین،

$$\frac{P\left(\text{فردی با مقدار } x_j \text{ در زمان } t_j \text{ بمیرد}\right)}{P\left(\text{فردی در زمان } t_j \text{ بمیرد}\right)} \qquad (۴.۲)$$

فرض شده است که مشاهدات مستقل از یکدیگر هستند. بنابراین، احتمال شرطی $L_i(\beta)$، احتمال این است که برای مشاهده‌ی iام در زمان t_i، $i = ۱, ۲, \ldots, n$ شکست رخ داده باشد. به‌عبارت دیگر،

$$L_i(\beta) = \frac{P\left(\text{فردی با مقدار } x_j \text{ در زمان } t_j \text{ بمیرد}\right)}{\sum_{k \in \Re_i} P\left(\text{فرد } k\text{ام در زمان } t_j \text{ بمیرد}\right)} \qquad (۵.۲)$$

که در آن، $\Re_i$ مجموعه‌ی خطر برای مشاهده‌ی iام است و به‌صورت

$$\Re_i = \{j : t_j' < t_i \leq t_j; i = ۱, ..., n\}$$

تعریف می‌شود. اگر احتمال شکست در زمان t_j با احتمال شکست در بازه‌ی $[t_k, t_j + \Delta)$ جای‌گذاری شود، آن‌گاه

$$\frac{\frac{P\left(\text{فردی با مقدار } x_j \text{ در } [t_j, t_j + \Delta) \text{ بمیرد}\right)}{\Delta}}{\sum_{k \in \Re_i} \frac{P\left(\text{فرد } k\text{ام در زمان } [t_j, t_j + \Delta) \text{ بمیرد}\right)}{\Delta}} \qquad (۶.۲)$$

[۵] Optional

وقتی $\Delta \to ۰$، رابطه‌ی بالا برابر است با

$$\frac{\text{خطر در زمان } t_j \text{ برای فردی با مقدار } x_j}{\text{خطر در } t_j \text{ برای فرد } k\text{ام} \sum_{k\in\Re_i}} = \frac{h(t_j, \boldsymbol{X}_j)}{\sum_{k\in\Re_i} h(t_j, \boldsymbol{X}_k)} \qquad (۷.۲)$$

بنا به رابطه‌ی (۲.۲)، احتمال اینکه فردی در زمان t_j شکست بخورد (بمیرد) به شرطی که t_j، زمان سانسور نباشد، برابر است با

$$\frac{h_۰(t_j)\exp\{\boldsymbol{X}_j^\top \boldsymbol{\beta}\}}{\sum_{k\in\Re_i} h_۰(t_j)\exp\{\boldsymbol{X}_k^\top \boldsymbol{\beta}\}} = \frac{\exp\{\boldsymbol{X}_j^\top \boldsymbol{\beta}\}}{\sum_{k\in\Re_i}\exp\{\boldsymbol{X}_k^\top \boldsymbol{\beta}\}} \qquad (۸.۲)$$

از آن‌جا که این درستنمایی روی داده‌های سانسور نشده، شرطی شده است، درستنمایی شرطی[۶] به‌دست می‌آید.

$$L_c(\boldsymbol{\beta}) = \prod_{i=۱}^{n} \frac{\exp\left\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\right\}}{\sum_{j\in\Re_i}\exp\left\{\boldsymbol{X}_j^\top \boldsymbol{\beta}\right\}} \qquad (۹.۲)$$

که n تعداد داده‌های سانسورنشده است. کاکس پیشنهاد می‌کند، که از درستنمایی شرطی به جای درستنمایی معمولی استفاده شود و آن را درستنمایی جزیی می‌نامند. فرض کنید $\delta = (\delta_۱, \ldots, \delta_n)$ که در آن اگر در مشاهده iام سانسور رخ‌داده باشد $\delta_i = ۰$ و اگر سانسور رخ‌نداده باشد $\delta_i = ۱$. بر این اساس، درستنمایی جزیی برابر است با

$$L(\boldsymbol{\beta}) = \prod_{i=۱}^{n} \left[\frac{\exp\left\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\right\}}{\sum_{j\in\Re_i}\exp\left\{\boldsymbol{X}_j^\top \boldsymbol{\beta}\right\}}\right]^{\delta_i} \qquad (۱۰.۲)$$

این تابع درستنمایی، مشابه درستنمایی جزیی، فقط احتمال‌ها را برای مشاهدات کامل (سانسورنشده) در نظر می‌گیرد. احتمال‌ها را برای داده‌های سانسورشده مستقیما حساب نمی‌کند اما داده‌های سانسور شده را در مجموعه‌ی خطر به‌حساب می‌آورد.

برای پیدا کردن برآورد حداکثر درستنمایی از بردار امتیاز و ماتریس اطلاعات نمونه، به صورت زیر استفاده می‌شود:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log L_c(\boldsymbol{\beta}) = \left(\frac{\partial}{\partial \boldsymbol{\beta}_۱} \log L_c(\boldsymbol{\beta}), \cdots, \frac{\partial}{\partial \boldsymbol{\beta}_p} \log L_c(\boldsymbol{\beta})\right)^\top$$

$$I(\boldsymbol{\beta}) = -\frac{\partial^۲}{\partial \boldsymbol{\beta}^۲} \log L_c(\boldsymbol{\beta}) = -\begin{bmatrix} \frac{\partial^۲}{\partial \beta_۱ \partial \beta_۱} \log L_c(\boldsymbol{\beta}) & \cdots & \frac{\partial^۲}{\partial \beta_۱ \partial \beta_p} \log L_c(\boldsymbol{\beta}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^۲}{\partial \beta_p \partial \beta_۱} \log L_c(\boldsymbol{\beta}) & \cdots & \frac{\partial^۲}{\partial \beta_p \partial \beta_p} \log L_c(\boldsymbol{\beta}) \end{bmatrix}$$

[۶] Conditional likelihood

باید معادلات زیر که معمولا به روش تکراری منجر می‌شود، حل شوند:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ln(L_c(\boldsymbol{\beta})) = ۰ \qquad (۱۱.۲)$$

بدین منظور،

$$\ln(L_c(\boldsymbol{\beta})) = \sum_{i=۱}^{n} \left[\boldsymbol{x}_i^\top \boldsymbol{\beta} - \ln \left(\sum_{j \in \Re_i} \exp\left\{ \boldsymbol{x}_j^\top \boldsymbol{\beta} \right\} \right) \right]$$

$$\frac{\partial}{\partial \boldsymbol{\beta}_k} \ln(L_c(\boldsymbol{\beta})) = \sum_{i=۱}^{n} \left(x_{ik} - \frac{\sum_{j \in \Re_i} x_{jk} \exp\left\{ \boldsymbol{x}_j^\top \boldsymbol{\beta} \right\}}{\sum_{j \in \Re_i} \exp\left\{ \boldsymbol{x}_j^\top \boldsymbol{\beta} \right\}} \right), \quad k = ۱, \ldots, p$$

$$\begin{aligned} I_{kl}(\boldsymbol{\beta}) &= -\frac{\partial^۲}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\beta}_l} \ln\left(L_c(\boldsymbol{\beta})\right) \\ &= \sum_{i=۱}^{n} \frac{\sum_{j \in \Re_i} x_{jk} x_{jl} \exp\left\{ \boldsymbol{x}_j^\top \boldsymbol{\beta} \right\}}{\sum_{j \in \Re_i} \exp\left\{ \boldsymbol{x}_j^\top \boldsymbol{\beta} \right\}} \\ &\quad - \sum_{i=۱}^{n} \frac{\sum_{j \in \Re_i} x_{jk} \exp\left\{ \boldsymbol{x}_j^\top \boldsymbol{\beta} \right\}}{\sum_{j \in \Re_i} \exp\left\{ \boldsymbol{x}_j^\top \boldsymbol{\beta} \right\}} \times \frac{\sum_{j \in \Re_i} x_{jl} \exp\left\{ \boldsymbol{x}_j^\top \boldsymbol{\beta} \right\}}{\sum_{j \in \Re_i} \exp\left\{ \boldsymbol{x}_j^\top \boldsymbol{\beta} \right\}}, \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k, l = ۱, \ldots, p \end{aligned}$$

برآوردگر $\hat{\boldsymbol{\beta}}$ از الگوریتم تکراری نیوتن– رافسون[۷] برای یافتن ریشه‌ی معادله‌ی (۱۱.۲) به‌صورت زیر به دست می‌آید.

$$\boldsymbol{\beta}_{n+۱} = \boldsymbol{\beta}_n - \left(\frac{\partial^۲}{\partial \boldsymbol{\beta}^۲} \log L_c(\boldsymbol{\beta}) \right)^{-۱} \left(\frac{\partial}{\partial \boldsymbol{\beta}} \log L_c(\boldsymbol{\beta}) \right). \qquad (۱۲.۲)$$

بنابراین می‌توان الگوریتم زیر را برای به‌دست آوردن برآوردگر $\hat{\boldsymbol{\beta}}$ در گرسیون کاکس پیشنهاد کرد:

(۱) برآوردگر $\boldsymbol{\beta}^{(۰)}$ را به عنوان یک برآوردگر اولیه در نظر بگیرید. ساسینی[۸] (۲۰۱۴) پیشنهاد $\boldsymbol{\beta}^{(۰)} = ۰$ را مطرح کرد.

(۲) به ازای $m = ۱, \ldots$ برآوردگر $\boldsymbol{\beta}^{(m)}$ از رابطه‌ی زیر به‌دست می‌آید:

$$\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m-۱)} + I^{-۱}(\hat{\boldsymbol{\beta}}^{(m-۱)}) \left[\frac{\partial}{\partial \boldsymbol{\beta}} \ln\left(L_c(\hat{\boldsymbol{\beta}}^{(m-۱)}) \right) \right]$$

(۳) مرحله‌ی (۲) را آن‌قدر ادامه می‌دهیم تا $\left| \hat{\boldsymbol{\beta}}^{(m+۱)} - \hat{\boldsymbol{\beta}}^{(m)} \right|$ مقدار کوچکی شود، به عبارت دیگر، الگوریتم همگرا شود.

[۷] Newton-Raphson algorithm
[۸] Sasieni

۵.۲ برآورد تابع بقا با استفاده از مدل cph

دو کمیت مطلوب اولیه دیدگاه تحلیل بقا، برآورد نسبت خطر و منحنی تابع بقا می‌باشند. اگر هیچ مدلی در برازش فراخوانی نشده باشد، منحنی بقا می‌تواند با یکی از روش‌های تجربی مانند KM و جدول طول عمر که در فصل ۱ به آن‌ها اشاره شد، برآورد شود. اما می‌توان تابع بقا را بر اساس تابع خطر نیز محاسبه کرد. بنابراین، پس از اینکه در مدل cph، پارامترها برآورد شدند، می‌توان تابع خطر و سپس تابع بقا را برآورد نمود. بدین منظور، طبق رابطه‌ی (۲۸.۱)، تابع بقای متناسب با مدل رگرسیون کاکس به‌صورت زیر است:

$$\begin{aligned} S(t,\boldsymbol{X}) &= \exp\left\{-\exp\{\boldsymbol{X}^\top\boldsymbol{\beta}\}\int_0^t h_0(u)du\right\} \\ &= S_0(t)^{\exp\{\boldsymbol{X}^\top\boldsymbol{\beta}\}} \end{aligned} \qquad (۱۳.۲)$$

که در آن $S_0(t) = \exp\left\{-\int_0^t h_0(u)du\right\}$ است. انتگرال $\int_0^t h_0(u)du$، تابع تجمعی خطر پایه نامیده می‌شود و آن را با نماد $H(t)$ نشان می‌دهند. روش‌های گوناگونی برای برآورد این تابع وجود دارد (کلین و موشبرگر[۹]، ۱۹۹۷). بنابراین، تابع بقا به‌صورت زیر برآورد می‌شود:

$$\hat{S}(t,\boldsymbol{X}) = \left(\hat{S}_0(t)\right)^{\exp\{\boldsymbol{X}^\top\hat{\boldsymbol{\beta}}\}} \qquad (۱۴.۲)$$

که در آن $\hat{\boldsymbol{\beta}}$، برآورد پارامتر مجهول رگرسیونی $\boldsymbol{\beta}$ در مدل cph است. از طرف دیگر، $\hat{S}_0(t)$، برآورد تابع $S_0(t)$ است که روش‌های متفاوتی برای برآورد آن وجود دارد. از معروف‌ترین این روش‌ها، می‌توان به روش‌های پیشنهادی برسلو[۱۰] (۱۹۷۲) و تسیاتیس[۱۱] (۱۹۷۸) اشاره نمود.

برسلو (۱۹۷۲) فرض می‌کند که $h_0(t)$ بین مشاهدات سانسورنشده ثابت است. بنابراین، $S_0(t)$ به‌صورت زیر برآورد می‌شود:

$$\hat{S}_0^{\mathrm{B}}(t,\boldsymbol{X}) = \prod_{y(i)\le t}\left(1 - \frac{\delta_i}{\sum_{j\in\Re_i}\exp\{\boldsymbol{X}_j\boldsymbol{\beta}\}}\right). \qquad (۱۵.۲)$$

[۹] Klein and Moeschberger
[۱۰] Breslow
[۱۱] Tsiatis

در این حالت، $\hat{S}_{0,\beta}(t)$ می‌تواند مقادیر منفی را اختیار کند. برای رفع این مشکل، تسیاتیس (۱۹۷۸) روش زیر را پیشنهاد می‌دهد:

$$\hat{S}_0^{\mathrm{T}}(t, \boldsymbol{X}) = \exp\left\{-\hat{\Lambda}_{0,T}(t)\right\},$$

که در آن

$$\hat{\Lambda}_{0,\beta}(t) = \sum_{y_i \le t} \frac{\delta_i}{\sum_{j \in \Re_i} \exp\left\{\boldsymbol{X}_j^\top \boldsymbol{\beta}\right\}}. \qquad (۱۶.۲)$$

۶.۲ ویژگی‌های مدل کاکس

بر اساس مدل (۳.۲)، $h_0(t)$ تابع خطر برای فردی است که مقدار متغیر توضیحی‌اش برابر صفر است. این تابع فقط به t وابسته است و به تغییرات در $\boldsymbol{X}$ بستگی ندارد. مولفه‌ی نمایی رگرسیون کاکس، فقط شامل $\boldsymbol{X}$ است و به زمان t وابستگی ندارد. ماتریس X از زمان مستقل است.

منظور مفهوم «مستقل از زمان» این است که مقدار یک متغیر با زمان تغییری نمی‌کند، می‌توان به متغیر جنسیت به عنوان مثالی از آن اشاره کرد. به‌طور کلی متغیرهایی که با گذشت زمان، احتمال کمی وجود دارد که مقدار آن‌ها تغییر کند را متغیرهای مستقل از زمان گویند مانند وضعیت سیگاری بودن. با این وجود، متغیرهایی وجود دارند که گاهی اوقات بسته به مسئله مورد نظر از آن‌ها به‌عنوان متغیرهای مستقل از زمان یاد می‌شود مانند متغیرهای سن و وزن.

یکی دیگر از خواص جذاب مدل کاکس این است که، حتی اگر $h_0(t)$ نامعلوم باشد، می توان β را در مدل رگرسیون کاکس برآورد نمود و از این رو، نسبت خطر محاسبه می‌گردد.

تابع خطر $h(t, \boldsymbol{x})$ و به‌طور مشابه منحنی بقا $S(t, \boldsymbol{x})$ می‌توانند برای مدل کاکس وقتی که تابع خطر پایه مشخص نباشد برآورد شوند. بنابراین، با مدل کاکس، با استفاده از حداقل فرضیات، می‌توان اطلاعات اولیه مطلوب از تحلیل بقا، شامل، نسبت خطر و منحنی بقا را به‌دست آورد.

در مدل کاکس (۲.۲) تابع خطر پایه، $h_0(t)$، نامعلوم است و هیچ فرضیه‌ی خاصی برای شکل آن در نظر گرفته نمی‌شود. اما از آن‌جا که برای تابع $h(\boldsymbol{X}^\top \boldsymbol{\beta})$ در رابطه‌ی (۱.۲) برای مدل‌بندی تابع مخاطره، کاکس صورت نمایی $\exp\{\boldsymbol{X}^\top \boldsymbol{\beta}\}$ را پیشنهاد داد، مدل کاکس یک مدل نیمه پارامتری است. می‌توان فرض کرد تابع خطر پایه، شکل تابعی کاملا مشخصی دارد که در

این حالت، مدل پارامتری به‌وجود می‌آید. برای مثال، فرض کنید تابع خطر را بتوان با توزیع وایبل مدل‌بندی کرد، در این صورت

$$h(t, \boldsymbol{X}) = pt^{p-1}\lambda \qquad (۱۷.۲)$$

که در آن $\lambda = \exp\{\boldsymbol{X}^\top \boldsymbol{\beta}\}$ و $h_0(t) = pt^{p-1}$.

مدل کاکس یک مدل قوی است به‌طوری که نتایج حاصل از مدل کاکس نتایج تقریبا نزدیکی به مدل‌های پارامتری دارد. برای مثال، اگر مدل پارامتری، وایبل باشد، استفاده از مدل کاکس نتایج قابل مقایسه‌ای با نتایج به‌دست آمده برای مدل وایبل خواهد داد. اگر از مدل پارامتری اطمینان وجود داشته باشید، بهتر است از مدل‌های پارامتری استفاده شود. اگرچه روش‌های متنوعی برای ارزیابی آزمون نیکویی برازش مدل‌های پارامتری وجود دارند، اما ممکن است نتوان به‌طور دقیق یک مدل پارامتری مناسب انتخاب کرد. بنابراین، در این موقعیت‌ها، مدل رگرسیون کاکس نتایج به اندازه کافی قابل اطمینان می‌دهد که انتخاب مدل صحیح است.

مدل رگرسیونی کاکس در رابطه‌ی (۲.۲)، تابع خطری را نتیجه می‌دهد که از حاصل‌ضرب تابع خطر پایه در t و حالت نمایی که شامل $\boldsymbol{X}$ها مستقل از t می‌باشد. تابع‌های $h_0(t)$ و $\exp\{\boldsymbol{X}^\top \boldsymbol{\beta}\}$ نامنفی هستند، بنابراین تابع خطر همواره نامنفی برآورد می‌شود. قسمت نمایی ضرب جذاب است زیرا برازش مدل را تضمین می‌کند و همیشه برآورد خطر نامنفی خواهد داد. اگر به جای قسمت نمایی مدل، از تابع خطی $\boldsymbol{X}$ استفاده شود، ممکن است ضرایب منفی به‌دست آیند، که اجازه چنین چیزی وجود ندارد.

کاربردی‌ترین روش برای یافتن ارتباط متغیرهای توضیحی با متغیر پاسخ تابع خطر، مدل رگرسیون کاکس است اما این مدل، محدودیت‌هایی نیز دارد. یکی از این محدودیت‌ها، فرض خطرهای متناسب است. به این معنی که لازم است نسبت خطرها روی زمان ثابت باشند که خطر برای هر شخص متناسب با خطر برای شخص دیگر است که ثابت تناسب، مستقل از زمان می‌باشد.

دو بردار متغیر تصادفی دلخواه $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ و $\boldsymbol{X}^* = (X_1^*, \ldots, X_p^*)^\top$ را در نظر بگیرید. نسبت خطر[۱۲] به‌صورت زیر تعریف می‌شود:

$$HR = \frac{\hat{h}(t, \boldsymbol{X}^*)}{\hat{h}(t, \boldsymbol{X})} \tag{۱۸.۲}$$

معمولا فرض می‌شود که $\boldsymbol{X}^*$، گروهی با خطر بیشتر باشد (مانند گروه دارونما) و $\boldsymbol{X}$، گروهی با خطر کمتر باشد (مانند گروه درمان). بنابراین، به دلیل تفسیرپذیری، معمولا هدف، آن است که $HR \geq 1$. به‌عبارت دیگر،

$$\hat{h}(t, \boldsymbol{X}^*) \geq \hat{h}(t, \boldsymbol{X}). \tag{۱۹.۲}$$

در نسبت خطر، توابع خطر پایه ساده می‌شوند. بنابراین،

$$\begin{aligned} HR &= \frac{\hat{h}_0(t) \exp\{\sum_i \hat{\boldsymbol{\beta}}_i \boldsymbol{X}_i^*\}}{\hat{h}_0(t) \exp\{\sum_i \hat{\boldsymbol{\beta}}_i \boldsymbol{X}_i\}} \\ &= \exp\left\{\sum_i \hat{\boldsymbol{\beta}}_i \left(\boldsymbol{X}_i^* - \boldsymbol{X}_i\right)\right\} \end{aligned} \tag{۲۰.۲}$$

بنابراین، در تفسیر مدل کاکس، ضریب رگرسیونی مثبت برای یک متغیر مستقل به معنی خطر بیشتر است و برعکس. برای درک بهتر، فرض کنید دو مشاهده از متغیر دلخواه X_i، فقط یک واحد اختلاف داشته باشند.

چنان‌چه ملاحظه می‌شود نسبت فوق فاقد $h_0(t)$ است، چون از صورت و مخرج کسر حذف گردیده است. بنابراین نسبت خطرات در همه زمان‌ها مقدار ثابتی است و اگر لگاریتم خطرات برای هر دو نفر دلخواه به‌صورت نموداری رسم شود تابع خطرات آن‌ها یا به طور معادل، تابع‌های بقای آنها کاملا موازی خواهند بود. برای درک بهتر، رابطه‌ی (۱۳.۲) را در نظر بگیرید. با دو بار لگاریتم گرفتن، نتیجه می‌شود

$$\ln\left(-\ln(S(t, \boldsymbol{X}))\right) = -\ln\left(-\ln(S_0(t))\right) - \boldsymbol{X}^\top \boldsymbol{\beta}. \tag{۲۱.۲}$$

بنابراین، برای دو مجموعه از مشاهدات $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ و $\boldsymbol{x}^* = (x_1^*, \ldots, x_p^*)^\top$،

$$\ln\left(-\ln(S(t, \boldsymbol{x}^*))\right) - \ln\left(-\ln(S(t, \boldsymbol{x}))\right) = (\boldsymbol{x}^* - \boldsymbol{x})^\top \boldsymbol{\beta}. \tag{۲۲.۲}$$

[۱۲] Hazard ratio

بنابراین، فضای بین دو منحنی بقا، نسبت به زمان ثابت است و منحنی‌ها، موازی هستند.

بر اساس نسبت خطر، می‌توان بیان کرد که در فرضیه‌ی خطرهای متناسب، باید نسبت خطر با گذشت زمان ثابت باشد. لزوما این فرضیه، همیشه برقرار نیست. فرض کنید در یک مطالعه بیماران سرطانی به دو گروه تقسیم شوند. در گروه اول، برای درمان، روش جراحی انجام می‌شود و در گروه دوم، از روش‌های درمانی بدون عمل جراحی استفاده می‌کنند. مداخله‌ی عمل جراحی، خطر زیادی دارد اما منجر به چشم‌انداز بسیار خوبی از زندگی طولانی‌مدت می‌شود. درمان بدون جراحی، خطر کمی دارد اما چشم‌انداز کمی از زندگی را داراست. بنابراین، در اینجا نسبت خطر ثابت نیست. در ابتدا، زیاد و سپس در امتداد زمان، کاهش می‌یابد.

تاکنون در مطالعات زیادی، رگرسیون کاکس به‌کار برده شده است اما بر اساس یک مطالعه سیستماتیک، تنها در ۵ درصد آن‌ها فرضیه‌ی متناسب بودن خطرها بررسی شده است (آکس[۱۳]، ۱۹۸۳). رهیافت‌های مختلفی برای بررسی نسبت خطر ثابت یا خطرهای متناسب در مدل کاکس وجود دارند که می‌توان به روش‌های نموداری و آزمون نیکویی برازش اشاره کرد. اگر این فرضیه برقرار نباشد، می‌توان از مدل رگرسیون کاکس لایه‌بندی‌شده[۱۴] استفاده کرد.

۱.۶.۲ آزمون معنی‌داری مدل رگرسیون کاکس

در بخش بعد یک مثال کاربردی مربوط به مجموعه داده‌های ژنومی و بالینی بیماران سرطان سلول‌های مغز استخوان با استفاده از مدل رگرسیون کاکس مورد بررسی قرار داده شده است. برای این منظور از بسته Coxph در نرم افزار R استفاده می‌شود. نتایج این تحلیل در جدول ۳.۲ ارائه شده است. در این جدول ستون اول نشان‌دهنده متغیرهای موجود در مدل می‌باشد. ستون دوم ضرایب رگرسیونی برآورد شده (β) در مدل را با استفاده از نتایج بخش ۴.۲ نمایش می‌دهد. ستون چهارم خطاهای استاندارد ضرایب رگرسیون برآورد شده ($s.e.(\beta)$) و ستون پنجم نشان دهنده آماره z می‌باشد. این آماره به آماره والد معروف بوده، که یکی از دو روش آماری استفاده شده در برآورد ماکزیمم درستنمایی است. آماره z از طریق فرمول زیر محاسبه می‌شود:

$$z = \frac{\beta}{s.e.(\beta)}. \qquad (۲۳.۲)$$

[۱۳] Oakes
[۱۴] Stratified Cox regression model

در جدول ۳.۲ ستون آخر نشان‌دهنده p ـ مقدارهای به‌دست آمده براساس آزمون والد است که معنی‌داری هر کدام از متغیرها را نشان می‌دهد. لازم به ذکر است که در مدل رگرسیون کاکس از آماره آزمون نسبت درستنمایی[۱۵] یا آماره LR هم می‌توان استفاده نمود. برای آگاهی بیشتر در خصوص آزمون‌های والد و LR به کلینبوم و کلین [۱۶] مراجعه کنید.

۷.۲ مثال

در این بخش، برآوردگر معرفی شده در مجموعه داده‌های ژنومی و بالینی بیماران سرطان سلول‌های مغز استخوان از کرال و همکاران[۱۷] (۱۹۷۵) به‌کار برده شده است. این داده‌ها مربوط به ۶۵ بیمار است. از بین بیماران در زمان مطالعه، ۴۸ بیمار از بین رفته‌اند و ۱۷ نفر زنده مانده‌اند. در این مجموعه داده که به نام MYELOMA یا سرطان مغز استخوان معروف است، متغیر Time، زمان بقا را بر حسب ماه از لحظه‌ی تشخیص نشان می‌دهد. متغیر VSTATUS، دو مقدار ۰ و ۱ را پذیرفته است که به ترتیب نشان می‌دهد یک بیمار زنده مانده یا فوت کرده است. اگر مقدار VSTATUS، صفر باشد، مقدار متناظر متغیر TIME، سانسورشده است.

متغیرهایی که محقق تصور می‌کرده است با بقا ارتباط دارند عبارتند از: LogBUN، (لگاریتم BUN در زمان تشخیص)، HGB (هموگلوبین در زمان تشخیص)، PLATELET، (تعداد پلاکت‌ها در خون در زمان تشخیص، ۰ برای وضعیت غیرنرمال و ۱ برای وضعیت نرمال)، AGE (سن بر حسب سال در زمان تشخیص)، LogWBC (لگاریتم تعداد گلبول‌های سفید خون در زمان تشخیص)، FRAC (شکستگی کامل یا ناقص در استخوان در زمان تشخیص، ۰، عدم وجود شکستگی و ۱، وجود شکستگی)، LogPBM (لگاریتم درصد پروتین‌های پلاسما در مغز استخوان)، PROTEIN (پروتینوری در زمان تشخیص)، SCALC (سرم کلسیم در زمان تشخیص).

جدول ۱.۲، آمار توصیفی متغیرهای کمی این مجموعه داده را نشان می‌دهد.

[۱۵] likelihood ratio
[۱۶] Kleinbaum and Klein
[۱۷] Krall et al

جدول ۱.۲: جدول آماره‌های توصیفی متغیرهای کمی مجموعه داده‌ی سرطان مغز استخوان

متغیرها	کمینه	چارک اول	میانه	میانگین	چارک سوم	بیشینه
Time	۱/۲۵	۷/۰۰	۱۵/۰۰	۲۴/۰۱	۳۵/۰۰	۹۲/۰۰
LogBUN	۰/۷۷	۱/۱۴	۱/۳۲	۱/۳۹	۱/۵۶	۲/۲۳
HGB	۴/۹۰	۸/۸۰	۱۰/۲۰	۱۰/۲۰	۱۲/۰۰	۱۴/۶۰
AGE	۳۸/۰۰	۵۱/۰۰	۶۰/۰۰	۶۰/۱۵	۶۷/۰۰	۸۲/۰۰
LogWBC	۳/۳۶	۳/۶۴	۳/۷۳	۳/۷۶	۳/۸۷	۴/۹۵
LogPBM	۰/۴۷	۱/۳۶	۱/۶۲	۱/۵۴	۱/۸۴	۲/۰۰
PROTEIN	۰/۰۰	۰/۰۰	۱/۰۰	۳/۶۱	۴/۰۰	۲۷/۰۰
SCALC	۷/۰۰	۹/۰۰	۱۰/۰۰	۱۰/۱۲	۱۰/۰۰	۱۸/۰۰

به منظور بررسی فرضیه‌ی متناسب بودن خطرات مدل کاکس، همزمان از دو روش مبتنی بر آزمون (روش پیشنهادی گرامبش و سرنیا[۱۸]، ۱۹۹۴) و نمودار باقیمانده‌های شونفلد[۱۹] استفاده شد که نتایج آزمون نیکویی برارش (جدول ۲.۲) نشان داد فرضیه‌ی متناسب بودن خطرات برای تمامی متغیرها برقرار است. شکل ۱.۲، نمودار باقیمانده‌ها را برای تمامی متغیرها در این مجموعه داده نشان می‌دهد که نتیجه‌ی آزمون را تایید می‌کند. در این نمودار، باقیمانده‌ها نباید وابستگی به زمان را نشان دهند و بیشتر داده‌ها در راستای زمان بین خطوط عبور داده شده قرار گیرند.

[۱۸] Grambsch and Therneau
[۱۹] Schoenfeld residuals

جدول ۲.۲: نتایج آزمون نیکویی برازش برای بررسی فرضیه‌ی متناسب بودن متغیرهای مستقل در مجموعه‌ی داده‌ی سرطان مغز استخوان

متغیرها	ρ	χ^2	P ـ مقدار
LogBUN	−۰/۲۰	۳/۰۰	۰/۰۸
HGN	۰/۰۱	۰/۰۱	۰/۹۲
PLATELET	۰/۰۰	۰/۰۰	۰/۹۸
AGE	−۰/۱۳	۰/۹۷	۰/۳۲
LogWBC	−۰/۱۶	۱/۶۵	۰/۱۹
FRAC	−۰/۰۴	۰/۰۵	۰/۸۱
LogPBM	−۰/۱۹	۲/۱۱	۰/۱۴
PROTEIN	−۰/۰۶	۰/۱۹	۰/۶۶
SCALC	−۰/۱۷	۲/۴۴	۰/۱۱

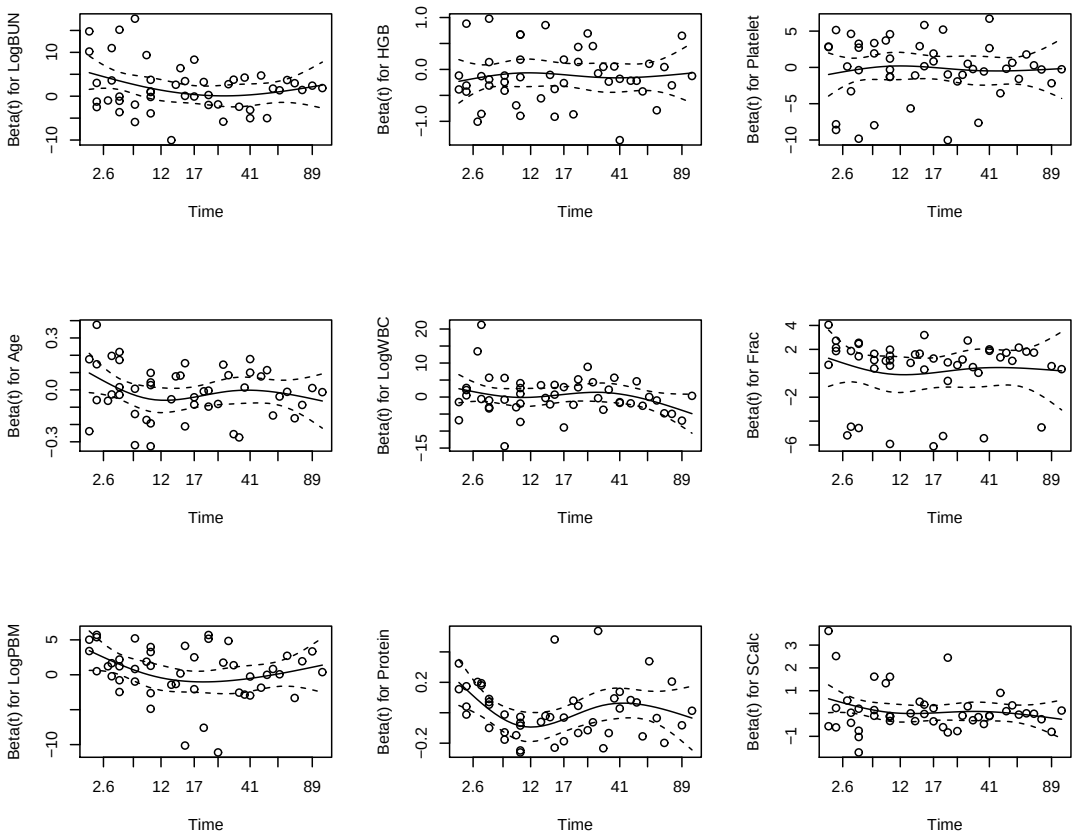
هدف این مدل، پیدا کردن یک مدل آگاهی‌بخش برای زمان بقا بیماران براساس داده‌های ژنومی است. علاوه بر این، بررسی اثرهای متغیرهای مستقل بر بقای بیماران، هدف رگرسیون cph است. برای تعیین برآورد ضرایب مدل خطرات متناسب کاکس از تابع درستنمایی جزیی استفاده می‌کنیم که در بین آن‌ها، هرکدام که مقدار $\exp\{\beta\}$ از یک بزرگتر باشد باعث افزایش خطر مرگ در بیماری است. $\exp\{\beta\}$، نسبت خطر را برای هر متغیر نشان می‌دهد. جدول ۳.۲ نتایج رگرسیون کاکس را ارایه می‌دهد که بیان می‌کند فقط متغیرهای LogBUN و HGB از نظر آماری در سطح خطای ۵ درصد معنی‌دار هستند.

مدل رگرسیون کاکس در رابطه‌ی (۳.۲) به‌صورت زیر برآورد می‌شود:

$$\begin{aligned}\ln\left(h(t,\boldsymbol{X})\right) &= \ln\left(h_{\circ}(t)\right) + ۱/۸۵\text{LogBUN} - ۰/۱۲\text{HGB} - ۰/۲۵\text{PLATELET} \\ &\quad -۰/۰۱\text{AGE} + ۰/۳۵\text{LogWBC} + ۰/۳۴\text{FRAC} + ۰/۳۸\text{LogPBM} \\ &\quad +۰/۰۱\text{PROTEIN} + ۰/۱۲\text{SCALC} \qquad (۲۴.۲)\end{aligned}$$

در نتیجه می‌توان مدل رگرسیونی cph را در رابطه‌ی (۲.۲) به صورت زیر نوشت:

$$\begin{aligned}h(t,\boldsymbol{X}) &= h_{\circ}(t)\exp\{۱/۸۵\text{LogBUN} - ۰/۱۲\text{HGB} - ۰/۲۵\text{PLATELET} \\ &\quad -۰/۰۱\text{AGE} + ۰/۳۵\text{LogWBC} + ۰/۳۴\text{FRAC} + ۰/۳۸\text{LogPBM} \\ &\quad +۰/۰۱\text{PROTEIN} + ۰/۱۲\text{SCALC}\} \qquad (۲۵.۲)\end{aligned}$$

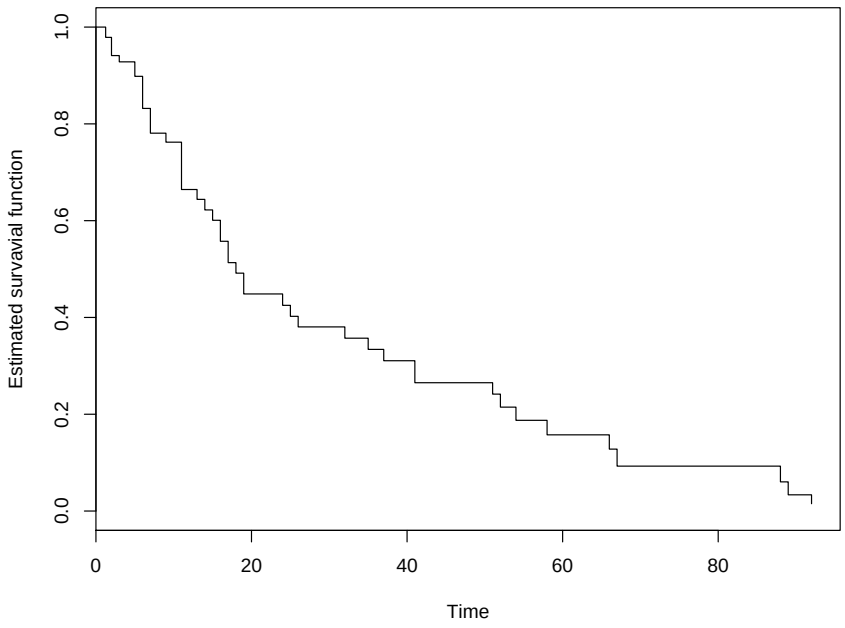


شکل ۱.۲: نمودار باقیمانده‌های شوینفلد برای متغیرهای مجموعه داده‌ی سرطان مغز استخوان

با توجه به نتایج به‌دست آمده از جدول ۳.۲، میزان خطر مرگ افراد با یک واحد افزایش در LogBUN، LogWBC، FRAC، LogPBM، PROTEIN و SCALC به‌ترتیب به اندازه‌ی ۶/۳۹، ۱/۴۲، ۱/۴۰، ۱/۴۶، ۱/۰۱ و ۱/۱۳ برابر افزایش می‌یابد که مقدار بزرگتری از یک است و نشان می‌دهد که با افزایش عوامل گفته شده خطر مرگ در بیماران سرطان مغز استخوان افزایش می‌یابد. شکل ۲.۲ نمودار بقا را برای این مجموعه نشان می‌دهد.

جدول ۳.۲: نتایج مدل رگرسیونی cph بر داده‌های سرطان مغز استخوان

متغیرها	β	$\exp\{\beta\}$	$s.e.(\beta)$	z	P ـ مقدار
LogBUN	۱/۸۵	۶/۳۹	۰/۶۵	۲/۸۳	۰/۰۰
HGB	−۰/۱۲	۰/۸۸	۰/۰۷	−۱/۷۵	۰/۰۷
PLATELET	−۰/۲۵	۰/۷۷	۰/۵۱	−۲/۵۰	۰/۶۱
AGE	−۰/۰۱	۰/۹۸	۰/۰۱	−۰/۶۷	۰/۵۰
LogWBC	۰/۳۵	۱/۴۲	۰/۷۱	۰/۴۹	۰/۶۲
FRAC	۰/۳۴	۱/۴۰	۰/۴۰	۰/۸۴	۰/۴۰
LogPBM	۰/۳۸	۱/۴۶	۰/۴۸	۰/۷۸	۰/۴۳
PROTEIN	۰/۰۱	۱/۰۱	۰/۰۲	۰/۵۰	۰/۶۱
SCALC	۰/۱۲	۱/۱۳	۰/۱۰	۱/۲۴	۰/۲۱



شکل ۲.۲: نمودار بقای برآورد شده بر اساس مدل cph برای مجموعه داده‌ی بیماران سرطان مغز استخوان

فصل ۳

مدل رگرسیون خطرات متناسب کاکس جریمه‌شده

۱.۳ مقدمه

امروزه با پیشرفت علم، تکنولوژی میکروآرایه‌های[۱] DNA امکان شبیه‌سازی از هزاران ژن را فراهم ساخته است که رویکردی قدرتمند در مطالعه ژنتیکی انواع مختلف تومور است. تصاویر ژنومی به‌طور گسترده‌ای می‌تواند برای طبقه‌بندی مولکولی سرطان‌ها، مطالعه سطوح مختلف پاسخ‌های یک دارو در حوزه داروسازی و نیز پیش‌بینی پیامدهای بالینی بیماران مورد استفاده قرار گیرد. با این حال، توسعه کمتری در رابطه با پروفایل‌های ژن به فنوتیپ‌های[۲] دیگر، مانند فنوتیپ‌های مداوم کمی یا فنوتیپ‌های بقا سانسورشده مانند زمان بازگشت به سرطان و زمان مرگ، صورت گرفته است. با توجه به تنوع زیادی در زمان به برخی رویدادهای بالینی مانند بازگشت سرطان در بیماران، مطالعه فنوتیپ‌های بقای احتمالی سانسورشده[۳] می‌تواند نسبت به درمان فنوتیپ‌ها

[۱] Microarrays
[۲] Phenotypes
[۳] Possibly censored survival phenotypes

به عنوان متغیرهای دودویی یا رسته‌ای[۴] مفید باشد.

مدل رگرسیون cph، محبوب‌ترین روش در تحلیل رگرسیونی برای داده‌های سانسورشده بقا است. با این حال، با توجه به فضای بسیار زیادی از پیشگوها، یعنی ژن‌هایی با سطوح مختلف که با آزمایش‌های ریزآرایه اندازه‌گیری شده‌اند، روش‌های کلاسیک مانند درستنمایی ماکزیمم جزیی نمی‌توانند به‌طور مستقیم برای برآورد پارامترها استفاده شوند.

در برآورد ضرایب مدل cph، هیچ یک از ضرایب رگرسیون برآوردشده دقیقا برابر صفر نیست و تمام متغیرها در مدل نهایی حضور دارند. در نتیجه، قادر به انتخاب متغیرهای مهم و مدیریت حالتی که تعداد متغیرها بیشتر از تعداد مشاهدات باشد، نخواهیم بود. در موارد معمول که اندازه n بیمار از تعداد p متغیر پیشگو بیشتر باشد، می‌توان پارامتر نامعلوم β را با ماکزیمم کردن لگاریتم درستنمایی جزیی برآورد کرد. اما برخی تحقیقات نشان داده‌اند که مدل cph به‌دلیل فضای بالای بعد پیش‌گوها و همخطی بالای برخی از ژن‌ها نمی‌تواند به‌طور مستقیم برای پیشگویی زمان بقا در حالت $p > n$ در ریزآرایه‌ها کاربردی باشد. هنگامی که از داده‌های ریزآرایه برای تحلیل بقا استفاده می‌شود، مجموعه داده همیشه از داده‌ی هزاران حالت ژن درست می‌شود. بنابراین به دلیل اینکه ممکن است (۱) مدل بسیار پیچیده باشد (۲) مدل کامل و درست، ناشناخته باشد و (۳) وجود متغیرهای اضافی در مدل باعث بیش‌برآوردی شود، نمی‌توان از همه متغیرهای پیشگو در مدل استفاده کرد. و از آنجایی که مدل‌های با تعداد متغیرهای کمتر همواره دلخواه تحلیل‌گران آماری بوده است بنابراین باید مدلی انتخاب شود که هم آسان و قابل فهم و هم قابل توضیح برای دیگران باشد.

روش‌های مختلفی برای انتخاب مدل وجود دارند. در عمل، انتخاب مدل با ترکیبی از (۱) آگاهی از علم، (۲) آزمایش و خطا، (۳) روش‌های انتخاب متغیر خودکار مانند: انتخاب پیشرو[۵]، پسرو[۶] و گام به گام[۷]، (۴) اندازه‌گیری تغییرات توضیح داده‌شده، (۵) معیارهای اطلاع مانند AIC و BIC و (۶) کاهش بعد در ابعاد بالا، انجام می‌شود.

از دیدگاه بیولوژیکی، باید انتظار داشت که فقط یک زیرمجموعه کوچک از ژن‌ها برای

---

[۴] Categorical
[۵] Forward selection
[۶] Backward
[۷] Stepwise

پیش‌گویی فنوتیپ‌ها مناسب باشد. استفاده از تمام ژن‌ها در مدل پیش‌گو موجب نویز می‌شود و انتظار می‌رود که عملکرد پیش‌گو ضعیف باشد. با توجه به ابعاد بالا، روش‌های کلاسیک انتخاب متغیر استاندارد مانند انتخاب بهترین زیرمجموعه[۸]، و انتخاب گام به گام نمی‌توانند اعمال شوند. تیبشیرانی[۹](۱۹۹۷) روش جریمه‌ی لاسو را که خودش در سال ۱۹۹۶ مطرح کرده بود، برای مسئله‌ی انتخاب متغیر برای مدل‌های cph گسترش داد و پیشنهاد کرد از روش برنامه‌ریزی درجه دوم برای ماکزیمم کردن درستنمایی جزئی جریمه‌شده لاسو به منظور به‌دست آوردن برآورد پارامترها استفاده کند.

علاوه بر اینکه بعد ماتریس طرح در مسایل ژنومی بالا است، سطوح بیان برخی از ژن‌ها هم اغلب بسیار همبسته هستند، که باعث ایجاد همخطی چندگانه بالا می‌شود. برای مقابله با مشکل همخطی، رایج‌ترین رویکرد استفاده از رگرسیون ریج است که در فصل ۱ به آن اشاره شد. اضافه کردن جریمه‌ی ریج به تابع درستنمایی در مدل رگرسیون cph، باعث می‌شود که برآوردگر کاکس معمولی را به سمت صفر منقبض کند. تفاوت رگرسیون خطرات متناسب کاکس ریج با رگرسیون ریج معمولی در این است که اینجا تابع درستنمایی که با تابع ریج جریمه شده است، ماکزیمم می‌شود.

همان‌طور که اشاره شد در مقایسه با روش جریمه ریج با محدودیت در مجموع توان دوم ضرایب، روش لاسو روشی را برای انتخاب متغیر ارائه می‌دهد. این روش‌های جریمه‌شده بیشتر در تنظیمات جایی‌که در آن حجم نمونه بیشتر از تعداد پیشگوها است مورد بررسی قرار گرفته‌اند. اخیرا روش جدیدی توسط ژو و هیستی (۲۰۰۵) پیشنهاد شده است که ترکیب خطی ریج و لاسو می‌باشد و به الاستیک‌نت معروف است. ترکیب روش الاستیک‌نت و مدل رگرسیون CPH، مدل رگرسیون خطرات متناسب الاستیک‌نت را می‌سازد که تمام ویژگی‌های رگرسیون‌های cph و الاستیک‌نت را همزمان دارد.

در مدل رگرسیون کاکس، ناپایداری ضرایب رگرسیونی برآوردشده می‌تواند با ماکزیمم کردن لگاریتم درستنمایی جزیی جریمه‌شده کاهش یابد که در آن تابع جریمه ضرایب رگرسیون از لگاریتم درستنمایی جزیی ($l(\beta) = \ln L(\beta)$) که در رابطه‌ی (۱۰.۲) تعریف شده، کسر شده

[۸] Best subset selction
[۹] Tibshirani

است. به عبارت دیگر،

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_j \left[\boldsymbol{X}_i^\top \boldsymbol{\beta} - \ln \left(\sum_{j \in \Re_i} \exp \left\{ \boldsymbol{X}_j^\top \boldsymbol{\beta} \right\} \right) \right] \tag{۱.۳}$$

که در آن اگر $\delta_j = 1$ باشد داده سانسور نشده و اگر $\delta_j = 0$ داده سانسورشده است. آن‌گاه، برآوردگر رگرسیونی کاکس جریمه‌شده در حالت کلی به‌صورت زیر به دست می‌آید:

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} \left\{ l(\boldsymbol{\beta}) - P_\lambda(\boldsymbol{\beta}) \right\}. \tag{۲.۳}$$

تابع جریمه عمومی با $P_\lambda(\cdot)$ نشان داده می‌شود که $\lambda > 0$ یک پارامتر تنظیم‌کننده است. می‌توان مسیله‌ی فوق را مشابه روش کمترین توان‌های دوم در رگرسیون چندگانه، به یک مسئله‌ی کمینه‌سازی تبدیل کرد. رابطه‌ی (۲.۳) را می‌توان به‌طور معادل به‌صورت زیر نوشت:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ -l(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}) \right\} \tag{۳.۳}$$

یعنی از ابتدا، تابع جریمه‌ی $P_\lambda(\cdot)$ را به مسئله‌ی کمینه‌سازی تابع $-l(\boldsymbol{\beta})$ اضافه کرد. معمولا می‌توان از تابع جریمه‌ی زیر استفاده می‌شود.

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|^\gamma \tag{۴.۳}$$

انواع جریمه‌ها به ازای مقادیر مختلف γ معرفی و به صورت زیر ارائه می‌شوند:

(۱) اگر $\gamma = 0$ باشد، تابع جریمه به $p\lambda$ کاهش می‌یابد. در نتیجه برآوردگر به‌دست آمده معادل با برآوردگری است که از روش کلاسیک انتخاب متغیر بهترین زیر مجموعه به‌دست می‌آید. معیاری که در انتخاب بهترین زیرمجموعه از متغیرها می‌تواند مورد استفاده قرار گیرد، معیار AIC است. از آنجایی که در مدل کاکس، $h_0(t)$ نامشخص است، انتخاب مدل بر روی $\boldsymbol{\beta}$ تمرکز می‌کند. چون درستنمایی جزیی دارای خواص درستنمایی کلاسیک است و طبق مرتبه دوم سری تیلور در اثبات AIC کلاسیک استفاده شده است بنابراین AIC برای مدل کاکس به‌صورت زیر است:

$$\text{AIC} = -2l(\hat{\boldsymbol{\beta}}) + 2p \tag{۵.۳}$$

که $l(\hat{\boldsymbol{\beta}})$ لگاریتم درستنمایی جزیی، و p بعد $\hat{\boldsymbol{\beta}}$ می‌باشد که $\hat{\boldsymbol{\beta}}$ برآوردگری است که روش انتخاب متغیر پیشنهاد می‌کند.

(۲) اگر $\gamma = ۱$ باشد، تابع جریمه $P_\lambda(\boldsymbol{\beta})$ به جریمه‌ی لاسو تبدیل می‌شود.

(۳) اگر $\gamma = ۲$ باشد، تابع جریمه $P_\lambda(\boldsymbol{\beta})$ به تابع جریمه‌ی ریج تبدیل می‌شود.

در ادامه، به معرفی توابع جریمه ریج، لاسو و ترکیبی از آن‌ها پرداخته می‌شود.

۲.۳ رگرسیون cph با جریمه ریج

وقتی در مدل رگرسیونی cph، متغیرهای پیشگو همبستگی بالایی داشته باشند، روش‌های برآورد استاندارد مانند ماکزیمم درستنمایی می‌تواند برآوردهای ناپایدار و واریانس بزرگ را به همراه داشته باشد. این مشکل به‌طور واضح در رگرسیون خطی معمولی در فصل ۱ توضیح داده شده است. هورل و کنارد (۱۹۷۰) روش ریج را پیشنهاد دادند که بعضی مواقع به‌عنوان یک راه‌حل برای این مشکل به‌کار می‌رود. روش‌های رگرسیون ریج در مدل‌های خطی یک مقدار اریبی ناچیز در برآوردگرها می‌پذیرد که باعث پایداری بیشتر پارامتر و تقویت برآورد واریانس می‌شود. رویکرد درستنمایی جریمه‌شده در رگرسیون کاکس توسط ورویج و ون‌هاولینجن[۱۰] (۱۹۹۴) استفاده شده است که در آن لگاریتم درستنمایی جزیی جریمه‌شده به صورت زیر تعریف می‌شود:

$$l_\lambda^{\text{ridge}}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda \sum_{j=۱}^{p} \beta_j^۲ \qquad (۶.۳)$$

که λ در اینجا پارامتر ریج نامیده شده است و مقدار انقباض برآورد $\boldsymbol{\beta}$ را کنترل می‌کند.

فرض کنید $U(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta})$، و $\Omega(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{X}$ ماتریس اطلاع[۱۱] باشد. بنابراین برای تابع لگاریتم درستنمایی جزیی با جریمه‌ی ریج، فرض کنید

$$U_\lambda(\boldsymbol{\beta}) = U(\boldsymbol{\beta}) - \lambda \boldsymbol{\beta} \qquad \Omega_\lambda(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p$$

که $\boldsymbol{I}_p$ یک ماتریس همانی $p \times p$ است. با روش نیوتن‌ـرافسون، ماکزیمم تابع $l_\lambda(\boldsymbol{\beta})$ که در

[۱۰] Verwij and van Houwelingen
[۱۱] Information matrix

رابطه‌ی (؟؟) تعریف شده است، طبق الگوریتم تکراری زیر به دست می‌آید.

$$\begin{aligned}\boldsymbol{\beta}_{i+1} &= \boldsymbol{\beta}_i + (\Omega_\lambda(\boldsymbol{\beta}_i))^{-1} U_\lambda(\boldsymbol{\beta}_i) \\ &= \boldsymbol{\beta}_i + (\Omega(\boldsymbol{\beta}_i) + \lambda \boldsymbol{I}_p)^{-1} (U(\boldsymbol{\beta}_i) - \lambda \boldsymbol{\beta}_i) \\ &= (\Omega(\boldsymbol{\beta}_i) + \lambda \boldsymbol{I}_p)^{-1} (U(\boldsymbol{\beta}_i) + \Omega(\boldsymbol{\beta}_i)) .\end{aligned}$$

این الگوریتم تا همگرا شدن ادامه پیدا می‌کند و در نهایت $\hat{\boldsymbol{\beta}}_\lambda$ برآوردگر ریج تکراری به‌دست می‌آید. این برآوردگر، در اولین مرحله، از برآوردگر کاکس که با روش درستنمایی ماکزیمم به‌دست آمده است، استفاده می‌کند. بنابراین، برآورد مرحله‌ی اول ریج، برابر است با

$$\hat{\boldsymbol{\beta}}_1 = \left(\Omega(\hat{\boldsymbol{\beta}}) + \lambda I\right)^{-1} \Omega(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}$$

وقتی درجه همخطی شدید باشد، ویژگی‌های برآوردگر ریج به‌صورت زیر می‌باشد:

(۱) وقتی $\lambda = 0$ باشد، $\hat{\boldsymbol{\beta}}_\lambda = \hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}$، که همان برآوردگری است که از رگرسیون cph به‌دست می‌آید.

(۲) وقتی $\lambda > 0$ باشد، هر دو برآوردگر $\hat{\boldsymbol{\beta}}_\lambda$ و $\hat{\boldsymbol{\beta}}_1$ اریب هستند.

(۳) در مدل‌های خطی، برای یک مقدار λ ثابت، $\hat{\boldsymbol{\beta}}_\lambda = \hat{\boldsymbol{\beta}}_1$ است.

(۴) مدل ریج یک مدل انقباضی[۱۲] در برآورد ضرایبی است که متغیرها نسبت به یکدیگر خیلی همبسته باشند.

(۵) مواقعی که تعداد پارامترها بیشتر از تعداد مشاهدات باشد کاربرد دارد.

(۶) برآوردگر ریج، انتخاب متغیر انجام نمی‌دهد.

۳.۳ رگرسیون cph با جریمه لاسو

اگر چه رگرسیون ریج عملکرد پیشگوها را بهبود می‌دهد، اما نمی‌تواند ضرایب غیرضروری را حذف کند. تیبشیرانی (۱۹۹۶)، روش کمترین قدر مطلق انقباض و انتخاب عملگر که به لاسو معروف شده است را معرفی کرد. رگرسیون لاسو هم انقباض و انتخاب متغیر را انجام می‌دهد.

[۱۲] shrinkage model

تیبشیرانی (۱۹۹۷) جریمه لاسو را برای انتخاب متغیر و انقباض در مدل رگرسیون خطرات متناسب کاکس ارائه داد. طرح پیشنهادی او، لگاریتم درستنمایی جزیی آزمودنی را با تابع جریمه مجموع مقادیر قدر مطلق پارامترها ماکزیمم می‌کند. به‌عبارت دیگر،

$$\hat{\boldsymbol{\beta}}^{\text{LASSO}} = \text{argmax}_{\boldsymbol{\beta}} \left\{ l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{۷.۳}$$

ماهیت لاسو، کاهش مقدار ضرایب و ارائه بعضی از ضرایب دقیقا صفر است. تیبشیرانی این موضوع را در مطالعات شبیه‌سازی تایید کرد. رویکرد رگرسیون ریج ضرایب را انقباض می‌دهد اما دقیقا برابر صفر قرار نمی‌دهد اما لاسو یک ابزار برای دستیابی به یک مدل صرفه‌جو[۱۳] است؛ در واقع در رهیافت ریج، ضرایب دقیقا صفر بعید است رخ دهند. اما جریمه‌ی لاسو، برخی از برآوردهای ضرایب وقتی که λ به اندازه کافی بزرگ باشد را دقیقا برابر صفر قرار می‌دهد. برای محاسبه برآورد لاسو باید از روش‌های بهینه‌سازی برنامه‌ریزی[۱۴] درجه دو استفاده کرد که برعکس ریج راه‌حل غیرخطی است.

تیبشیرانی (۱۹۹۷) دو الگوریتم برای روش لاسو بر اساس تکنیک‌های بهینه‌سازی ارائه داد. الگوریتم‌های تکراری و شامل راه‌حل مکرر، مشکلات توان‌های دوم مطلق می‌باشد. استراتژی برای حل (۷.۳) این است که از روش نیوتون رافسون به‌روز شده[۱۵] استفاده شود.

فرض کنید $\boldsymbol{X}$ ماتریس طرح متغیرهای پیشگو را مشخص کند و $\boldsymbol{\eta} = \boldsymbol{X}^\top \boldsymbol{\beta}$

$$\begin{aligned} \boldsymbol{U} &= \frac{\partial}{\partial \boldsymbol{\eta}} l(\boldsymbol{\beta}) \\ \boldsymbol{A} &= -\frac{\partial^2 l}{\partial \boldsymbol{\eta} \boldsymbol{\eta}^\top} \end{aligned}$$

و $\boldsymbol{Z} = \boldsymbol{\eta} + \boldsymbol{A}^{-1}\boldsymbol{U}$ باشد. بنابراین جمله اول سری بسط تیلور برای $l(\boldsymbol{\beta})$ به‌صورت زیر است

$$(\boldsymbol{Z} - \boldsymbol{\eta})^\top \boldsymbol{A} (\boldsymbol{Z} - \boldsymbol{\eta}) \tag{۸.۳}$$

قبل از بیان الگوریتم، بهتر است یافتن برآوردگر لاسو را به‌صورت یک مسئله‌ی بهینه‌سازی نوشت. به عبارت دیگر،

$$\hat{\boldsymbol{\beta}}_n^{\text{LASSO}} = \max \{ l(\boldsymbol{\beta}) \} \quad \text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \leq s \tag{۹.۳}$$

[۱۳] Parsimonous model
[۱۴] Programming
[۱۵] Updated Newton-Raphson method

از این رو برای یافتن برآوردگر لاسو از الگوریتم زیر استفاده شده است:

(۱) s مقدار ثابت دلخواهی در نظر گرفته شود و قرار دهید $\hat{\boldsymbol{\beta}} = ۰$.

(۲) $\boldsymbol{\eta}$، $\boldsymbol{U}$، $\boldsymbol{A}$ و $\boldsymbol{Z}$ را براساس مقدار کنونی $\hat{\boldsymbol{\beta}}$ محاسبه کنید.

(۳) $(\boldsymbol{Z}-\boldsymbol{\eta})^{\top}\boldsymbol{A}(\boldsymbol{Z}-\boldsymbol{\eta})$ را با شرط $\sum|\boldsymbol{\beta}_i| \leq s$ کمینه کنید.

(۴) مراحل ۲ و ۳ را آن قدر تکرار کنید که $\hat{\beta}$ تغییر نکند.

کمینه‌سازی در مرحله (۳) با برنامه‌ریزی درجه دو انجام می‌شود.

خواص رگرسیون لاسو را می‌توان به صورت زیر بیان کرد:

(۱) رگرسیون لاسو هم انقباض و هم انتخاب متغیر انجام می‌دهد.

(۲) برای $p > n$، روش رگرسیون لاسو، حداکثر n متغیر انتخاب می‌کند.

(۳) جریمه لاسو برتری بیشتری به جریمه ریج در کاهش برخی ضرایب به صفر دارد.

(۴) لاسو برای انتخاب متغیر در مدل کاکس یک رقیب شایسته نسبت به انتخاب گام به گام به نظر می‌رسد.

از روش اعتبارسنجی متقابل برای به‌دست آوردن پارامتر تنظیم‌کننده پس از پایداری خطای اعتبارسنجی استفاده می‌شود.

۴.۳ رگرسیون cph با جریمه الاستیک‌نت

تابع جریمه الاستیک‌نت یکی از توابع جریمه جدید در این مدل‌هاست که اولین بار ژو و هیستی (۲۰۰۵) آن را معرفی کردند. این تابع جریمه، ترکیب جریمه‌های لاسو و ریج می‌باشد و شبیه لاسو انتخاب متغیر اتوماتیک با صفر قرار دادن برخی از ضرایب انجام می‌دهد با این تفاوت که الاستیک‌نت به انتخاب متغیرهای بیشتری در مقایسه با لاسو تمایل دارد. استفاده از برآوردگر الاستیک‌نت به‌ویژه در شرایطی که متغیرها با یکدیگر همبستگی بالایی دارند، کاربردی‌تر است زیرا لاسو فقط یکی از متغیرها را از بین مجموعه متغیرهای همبسته می‌تواند انتخاب کند.

تابع جریمه الاستیک نت به‌صورت زیر تعریف می‌شود:

$$p_{\alpha,\lambda}(|\beta_j|) = \lambda \times \left(\alpha|\beta_j| + (۱ - \alpha)\beta_j^۲\right).$$

در اینجا، $\alpha \in [۰, ۱]$ تعیین تاثیر جریمه‌ی لاسو نسبت به جریمه‌ی ریج است. بنابراین، برآوردگر الاستیک‌نت در مدل رگرسیون خطرات متناسب کاکس با تابع جریمه‌ی الاستیک‌نت که آن را با نماد $\hat{\boldsymbol{\beta}}^{\text{EN}}$ نشان می‌دهند، از رابطه‌ی زیر به‌دست می‌آید.

$$\hat{\boldsymbol{\beta}}^{\text{EN}} = \text{argmax}_{\boldsymbol{\beta}} \left\{ l(\boldsymbol{\beta}) + \lambda\alpha \sum_{j=۱}^{p} |\beta_j| + \lambda(۱ - \alpha) \sum_{j=۱}^{p} \beta_j^۲ \right\} \qquad (۱۰.۳)$$

تابع جریمه الاستیک نت شامل دو پارامتر تنظیم‌کننده است که به داده‌ها وابسته هستند. چالش برای پیدا کردن یک مجموعه از پارامترهای تنظیم‌کننده (α, λ) است، برای اینکه تابع درستنمایی جزیی جریمه شده را ماکزیمم کند. معمولا از یک روش جستجوی شبکه‌ای استفاده می‌شود که به محاسبات سیستمی در انحراف جریمه لگاریتم درستنمایی در هر نقطه از شبکه نیاز دارد که به این معنی است که تراکم شبکه در دقت و پیچیدگی زمانی الگوریتم کارآمد است (آبرامسکی و جانگ[۱۶]، ۱۹۹۴) .

در قسمت بعد دو الگوریتم برای برازش مدل کاکس با جریمه الاستیک‌نت ارایه شده‌اند.

۵.۳ الگوریتم

در سال ۲۰۰۷، پارک و هیستی[۱۷] الگوریتم تصحیح‌کننده‌ی متغیرهای پیش‌گو را برای محاسبه‌ی مدل رگرسیون cph با جریمه‌ی الاستیک‌نت پیشنهاد کردند، جیومن [۱۸] (۲۰۱۰) الگوریتم گرادیان نزولی[۱۹] را برای یافتن پاسخ مدل رگرسیون cph با جریمه‌ی لاسو معرفی کرد. ابتدا بر اساس یانگ و ژو [۲۰] (۲۰۱۳) الگوریتم مختصات نزولی [۲۱] (CMD) و سپس الگوریتم «مسیر[۲۲]» که سیمون و همکاران[۲۳] برای محاسبه‌ی برآوردگر الاستیک‌نت در مدل کاکس را پیشنهاد کردند،

[۱۶] Abramsky and Jung
[۱۷] Park and Hastie
[۱۸] Geoman
[۱۹] Gradient descent algorithm
[۲۰] Yang and Zou
[۲۱] Coordinate-majorixation-descent (CMD)
[۲۲] Pathwise
[۲۳] Simon et al

معرفی می‌شوند.

ابتدا فرض کنید

$$\dot{l}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta})$$

و

$$\ddot{l}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^\top} l(\boldsymbol{\beta})$$

۱.۵.۳ الگوریتم CMD برای رگرسیون کاکس با جریمه‌ی الاستیک‌نت

(۱) برآوردگر $\tilde{\boldsymbol{\beta}}$ را به‌صورت زیر به‌دست آورید

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{n} \left\{ \sum_{i=1}^{n} -l_i(\boldsymbol{\beta}) + \lambda P_\alpha(\boldsymbol{\beta}) \right\}$$

که در آن،

$$P_\alpha(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left(\alpha|\beta_j| + (1-\alpha)\beta_j^2 \right)$$

(۲) برای $j = 1, \ldots, p$،

$$D_j = \sum_{i=1}^{n} \frac{1}{4n} \left(\max_{j \in R_i}(\boldsymbol{x}_{ij}) - \min_{j \in R_i}(\boldsymbol{x}_{ij}) \right)^2$$

$$\dot{l}_j(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left[-\boldsymbol{x}_{ij} + \frac{\sum_{j \in R_i} \boldsymbol{x}_{ij} \exp\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\}}{\sum_{j \in R_i} \exp\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\}} \right]$$

و

$$\tilde{\boldsymbol{\beta}}_j^{\text{new}} = \frac{S\left(D_j \tilde{\boldsymbol{\beta}}_j - \dot{l}_j(\tilde{\boldsymbol{\beta}}).\lambda\alpha \right)}{D_j + \lambda(1-\alpha)}$$

که $S(z.t) = (|z| - t)^+ \text{sgn}(z)$ که در آن تابع sgn(.)، تابع علامت است و به‌صورت زیر تعریف می‌شود

$$\text{sgn}(a) = \begin{cases} 1 & a > 0 \\ 0 & a = 0 \\ -1 & a < 0 \end{cases}$$

و $(\boldsymbol{x})^+ = max(0, \boldsymbol{x})$ است.

(۳) برآوردگر $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}^{new}$ را به‌روز رسانی کنید

(۴) مرحله‌های ۲ و ۳ را آن‌قدر تکرار کنید تا برآوردگر $\tilde{\boldsymbol{\beta}}$ همگرا شود.

۲.۵.۳ الگوریتم مسیر برای رگرسیون کاکس با جریمه‌ی الاستیک‌نت

(۱) برآوردگر $\tilde{\boldsymbol{\beta}}$ را به‌صورت

$$\tilde{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} \left\{ \frac{۲}{n} \left(\boldsymbol{X}^\top \boldsymbol{\beta} - \ln \left(\sum_{i \in R_s} \exp\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\} \right) - \lambda P_\alpha(\boldsymbol{\beta}) \right) \right\}$$

و مجموعه $\tilde{\boldsymbol{\eta}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}}$ بیابید که در آن،

$$P_\alpha(\boldsymbol{\beta}) = \sum_{j=۱}^{p} \left(\alpha|\beta_j| + (۱ - \alpha)\beta_j^۲ \right)$$

(۲) $\ddot{l}(\tilde{\boldsymbol{\eta}})$ و $\boldsymbol{Z}(\tilde{\boldsymbol{\eta}})$ را محاسبه کنید. با توجه به‌بسط سری تیلور،

$$\begin{aligned} l(\boldsymbol{\beta}) &\approx l(\tilde{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \dot{l}(\tilde{\boldsymbol{\beta}}) + \frac{(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \ddot{l}(\tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})}{۲} \\ &= l(\tilde{\boldsymbol{\beta}}) + (\boldsymbol{X}\boldsymbol{\beta} - \tilde{\boldsymbol{\eta}})^\top \dot{l}(\tilde{\boldsymbol{\eta}}) + \frac{(\boldsymbol{X}\boldsymbol{\beta} - \tilde{\boldsymbol{\eta}})^\top \ddot{l}(\tilde{\boldsymbol{\eta}})(\boldsymbol{X}\boldsymbol{\beta} - \tilde{\boldsymbol{\eta}})}{۲} \end{aligned}$$

که $\tilde{\boldsymbol{\eta}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}}$ است. یا به‌عبارت جبری ساده‌تر،

$$l(\boldsymbol{\beta}) \approx \frac{۱}{۲} \left(z(\tilde{\eta}) - \boldsymbol{X}\boldsymbol{\beta} \right)^\top l''(\tilde{\eta}) \left(\boldsymbol{Z}(\tilde{\eta}) - \boldsymbol{X}\boldsymbol{\beta} \right) + C(\tilde{\eta}, \tilde{\boldsymbol{\beta}})$$

که

$$\boldsymbol{Z}(\tilde{\eta}) = \tilde{\boldsymbol{\eta}} - \ddot{l}(\tilde{\boldsymbol{\eta}})^{-۱} \dot{l}(\tilde{\boldsymbol{\eta}})$$

و $C(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\beta}})$ به $\boldsymbol{\beta}$ وابسته نیست.

(۳) $\hat{\boldsymbol{\beta}}$ را از کمینه کردن

$$\frac{۱}{n} \sum_{i=۱}^{n} w(\tilde{\boldsymbol{\eta}})_i \left(\boldsymbol{Z}(\tilde{\boldsymbol{\eta}})_i - \boldsymbol{x}_i^\top \boldsymbol{\beta} \right)^۲ + \lambda P_\alpha(\boldsymbol{\beta})$$

پیدا کنید. که $w(\tilde{\eta})_i$، iامین ورودی متعامد $\ddot{l}(\tilde{\eta})$ است.

(۴) مجموعه $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$، و $\tilde{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ است.

(۵) مرحله‌های ۲ و ۴ را آن‌قدر تکرار کنید تا برآوردگر $\hat{\boldsymbol{\beta}}$ همگرا شود.

۶.۳ محاسبات با R

در نرم افزار R دو پکیج برای محاسبه جریمه الاستیک‌نت در مدل کاکس وجود دارند.

۱.۶.۳ بسته glmnet

بسته glmnet برای مدل‌های خطی تعمیم یافته، با قید جریمه

$$(۱-\alpha)\sum_{j=۱}^{p}|\boldsymbol{\beta}_j|^۲ + \alpha\sum_{j=۱}^{p}|\boldsymbol{\beta}_j| \quad \alpha \in (۰,۱] \tag{۱۱.۳}$$

به‌کار می‌رود (فریدمن، هیستی و تیبشیرانی، ۲۰۰۸). در این بسته، از الگوریتم CMD استفاده شده است. برای نصب و دانلود این بسته در R می‌توان از دستور زیر استفاده کرد.

```
> install.packages("glmnet")
```

برای فراخوانی این بسته باید به صورت زیر عمل کرد

```
> library("glmnet")
```

و برای به‌روزرسانی آخرین نسخه این بسته از دستور زیر استفاده می‌شود

```
> glmnet.upgrade(testing=TRUE)
```

همچنین این بسته را می‌توان به صورت زیر اجرا کرد

```
> glmnet(x, y, family="cox", alpha  )
```

که در آن x ماتریس ورودی که هر سطر آن بردار مشاهدات است، y متغیر پاسخ نامیده می‌شود که در اینجا متغیر بقا است و با دو متغیر time و status با استفاده از تابع Surv از بسته‌ی survival تعیین می‌شود، family نوع رگرسیون را مشخص می‌کند که در اینجا باید رگرسیون کاکس انتخاب شود. متغیر status وضعیت مشاهده از نظر کامل یا سانسور بودن نشان می‌دهد. همچنین alpha پارامتر تنظیم‌کننده در رگرسیون جریمه شده است و مقداری بین $[۰,۱]$ را می‌پذیرد. در این بسته‌ی نرم‌افزاری، λ در رابطه‌ی (۱۰.۳)، برابر ۱ فرض شده است و بنا به تعریف تابع جریمه

در این بسته، مقدار $\alpha = ۰$، برآوردگر ریج و مقدار $\alpha = ۱$، برآوردگر لاسو را نتیجه می‌دهد. اگر α، مقداری را از بازه‌ی $(۰, ۱)$ بپذیرد، در این صورت، برآوردگر الاستیک‌نت در رگرسیون cph به‌دست می‌آید.

با استفاده از دستور زیر می‌توان مقدار پارامتر تنظیم‌کننده را به روش اعتبارسنجی متقابل، تعیین کرد.

```
> cv.glmnet(x, y, family="cox", alpha )
```

پس از تعیین پارامتر α، می‌توان متغیرهای انتخاب‌شده را به‌ازای آن مشخص کرد. مثالی ساده از برنامه‌ی R با یک مجموعه‌ی داده‌ی فرضی که متغیرهای x، $time$ و $status$ را دارد، به‌صورت زیر است:

```
> library("glmnet")
> library("survival")
> cv.fit <- cv.glmnet(x,Surv(time,status),family="cox")
> fit <-glmnet(x,Surv(time,status),family = "cox")
> Coefficients <- coef(fit, s= cv.fit$lambda.min)
> Coefficients
```

۲.۶.۳ بسته coxnet

بسته coxnet برای مدل cph بر اساس الگوریتم مسیر (سیمون و همکاران، ۲۰۱۱) آماده شده است. در این بسته، تابع جریمه به‌صورت زیر تعریف شده است:

$$\lambda\alpha\sum_{j=۱}^{p}|\beta_j| + \frac{(۱-\alpha)}{۲}\sum_{j=۱}^{p}\beta_j^۲ \quad \alpha \in (۰, ۱] \qquad (۱۲.۳)$$

برای نصب و دانلود این بسته در R می‌توان از دستور زیر استفاده کرد.

```
> install.packages("Coxnet")
```

برای فراخوانی این بسته باید به صورت زیر عمل کرد

```
> library("Coxnet")
```

و برای به‌روزرسانی آخرین نسخه این بسته از دستور زیر استفاده می‌شود

```
> Coxnet.upgrade(testing=TRUE)
```

همچنین این بسته را می‌توان به صورت زیر اجرا کرد

```
> Coxnet(x, y, penalty =  "Enet", alpha = 1, lambda = NULL)
```

که در آن x ماتریس ورودی که هر سطر آن بردار مشاهدات است، y متغیر پاسخ که است که در اینجا متغیر بقا است که از ستون‌هایی به نام time و status تشکیل شده است. penalty نوع جریمه را انتخاب می‌کند که می‌تواند جریمه‌ی لاسو را نیز بپذیرد. اگر مقدار $\alpha = ۱$ باشد، تابع جریمه الاستیک‌نت در این بسته به تابع جریمه‌ی لاسو تغییر می‌کند و انتخاب تابع جریمه برای آرگومان penalty تاثیری ندارد. برای انجام رگرسیون الاستیک‌نت در مدل cph، باید این آرگومان را برابر با "Enet" قرار داد. آرگومان lambda، پارامتر تنظیم‌کننده‌ی دیگر رگرسیون الاستیک‌نت در این بسته است که بر اساس روش اعتبار سنجی متقابل تعیین می‌شود.

دوباره مجموعه‌ی داده‌ی فرضی با متغیرهای x، $time$ و $status$ را در نظر بگیرید. برنامه‌ی ساده‌ای از این بسته به‌صورت زیر است:

```
 y=cbind(time,status)
fit<-Coxnet(x,y,penalty="Enet",alpha = 1,lambda=....)
fit$Beta
```

مقدار پیش‌فرض alpha، برابر ۱ است که صرف نظر از تابع جریمه، همواره جریمه‌ی لاسو را در نظر می‌گیرد. مشکلی که در این پکیج وجود دارد این است که باید همواره مقداری را برای alpha در نظر گرفت. این تابع، λ، دیگر پارامتر تنظیم‌کننده را با روش اعتبارسنجی متقابل به‌دست می‌آورد. پیشنهاد می‌شود که پارامتر α نیز به روش اعتبارسنجی متقابل با نوشتن کد دیگری بدین منظور تعیین شود.

از آنجایی که در جریمه الاستیک‌نت دو پارامتر تنظیم‌کننده‌ی alpha و lambda باید محاسبه شوند و چون در بسته‌ی glmnet فقط پارامتر lambda قابل محاسبه می‌باشد؛ در ادامه به معرفی پکیج c06 پرداخته شده است که هر دو پارامتر تنظیم پکیج glmnet را محاسبه می‌کند.

۳.۶.۳ بسته c06

بسته c06 برای مدل‌های کاکس تنظیم شده و خطی تعمیم یافته، با قید جریمه

$$\lambda \left((1-\alpha) \sum_{j=1}^{p} |\beta_j|^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right) \quad \alpha \in (0, 1] \qquad (۱۳.۳)$$

به‌کار می‌رود (سیل و همکاران، ۲۰۱۵).

برای نصب و دانلود این بسته در R می‌توان از دستور زیر استفاده کرد.

```
> install.packages("c06")
```

برای فراخوانی این بسته باید به صورت زیر عمل کرد

```
> library("c06")
```

و برای به‌روزرسانی آخرین نسخه این بسته از دستور زیر استفاده می‌شود

```
> c06.upgrade(testing=TRUE)
```

همچنین برای محاسبه پارامترهای تنظیم پکیج glmnet از طریق پکیج c060 باید به صورت زیر عمل کرد

```
> bounds <- t(data.frame(alpha = c(0, 1)))
> colnames(bounds) <- c("lower", "upper")
> nfolds <- 10
> set.seed(1234)
> foldid <- balancedFolds(class.column.factor = y[, 2], cross.outer = nfolds)
```

معمولا برای پیدا کردن تنظیمات از مقادیر پارامترهای تنظیم (α, λ)، برای این‌که اعتبار سنجی متقابل ۱۰ـ تایی مشتق لگاریتم درستنمایی جزیی جریمه شده مدل کمینه باشد. برای هر α داده شده λ بهینه با محاسبه مسیر تنظیم با تابع hc glmnet دستور زیر پیدا می‌شود.

```
> type.min = "lambda.1se"
```

تابع tune.glmnet.interval مشتق لگاریتم درستنمایی جزیی مدل با پارامترهای تنظیم داده شده (α, λ) محاسبه می‌کند و با استفاده از دستور زیر می‌توان هر دو پارامتر را به‌طور همزمان مشخص کرد.

```
> fit <- epsgo(Q.func = "tune.glmnet.interval", bounds = bounds,
+ parms.coding = "none", seed = 1234, fminlower = -100, x = x, y = y,
+ family = "cox", foldid = foldid, type.min = "lambda.1se",
+ type.measure = "deviance")


sumint <- summary(fit, verbose = TRUE)
```

فصل ۴

کاربردها

۱.۴ مقدمه

در این فصل، از تکنیک جریمه شده در رگرسیون خطرات متناسب کاکس که در فصل ۳ معرفی شد، بر روی داده‌های ژنومی استفاده کرده و کاربردی از تکنیک جریمه شده برای داده‌های پزشکی بررسی شده است.

از آن‌جا که تعداد متغیرها در داده‌های ژنومی بسیار زیاد است، بررسی روش‌های جریمه شده در آن‌ها به سادگی امکان‌پذیر نمی‌باشد. بنابراین، به منظور درک بهتری از رهیافت‌های مطرح شده در فصل ۳، ابتدا یک مجموعه داده با تعداد متغیر کمتری در نظر گرفته شده است و ضرایب رگرسیونی در آن مورد مقایسه قرار گرفته است.

۲.۴ مجموعه داده‌های بیماران سرطان سلول‌های مغز استخوان

این محموعه داده در فصل ۲ مورد بررسی قرار گرفته است و فرضیات لازم برای مدل رگرسیون cph در آن برقرار است.

یادآوری می‌شود متغیرهای این مجموعه داده عبارتند از:

LogBUN لگاریتم BUN در زمان تشخیص

HGB هموگلوبین در زمان تشخیص

PLATELET تعداد پلاکت‌ها در خون در زمان تشخیص، ۰ برای وضعیت غیرنرمال و ۱ برای وضعیت نرمال

AGE سن بر حسب سال در زمان تشخیص

LogWBC لگاریتم تعداد گلبول‌های سفید خون در زمان تشخیص

FRAC شکستگی کامل یا ناقص در استخوان در زمان تشخیص، ۰، عدم وجود شکستگی و ۱، وجود شکستگی

LogPBM لگاریتم درصد پروتین‌های پلاسما در مغز استخوان

PROTEIN پروتینوری در زمان تشخیص

SCALC سرم کلسیم در زمان تشخیص

هدف از این مطالعه، پیدا کردن یک مدل آگاهی‌بخش برای زمان بقای بیماران بر اساس داده‌های بیماران سرطان سلول مغز استخوان است. علاوه بر این، پیش‌بینی بقای بیماران و تعیین ضرایب (اندازه‌ی تاثیر متغیرها) از اهداف دیگر این پژوهش است.

برای انجام رگرسیون cph، باید داده‌ها استاندارد شوند. بنابراین، جدول ۱.۴، برآورد ضرایب مدل رگرسیون خطرات متناسب کاکس را برای داده‌های استاندارد شده نشان می‌دهد.

جدول ۱.۴: برآورد ضرایب مدل cph مجموعه داده‌های استاندارد شده بیماران سرطان مغز استخوان

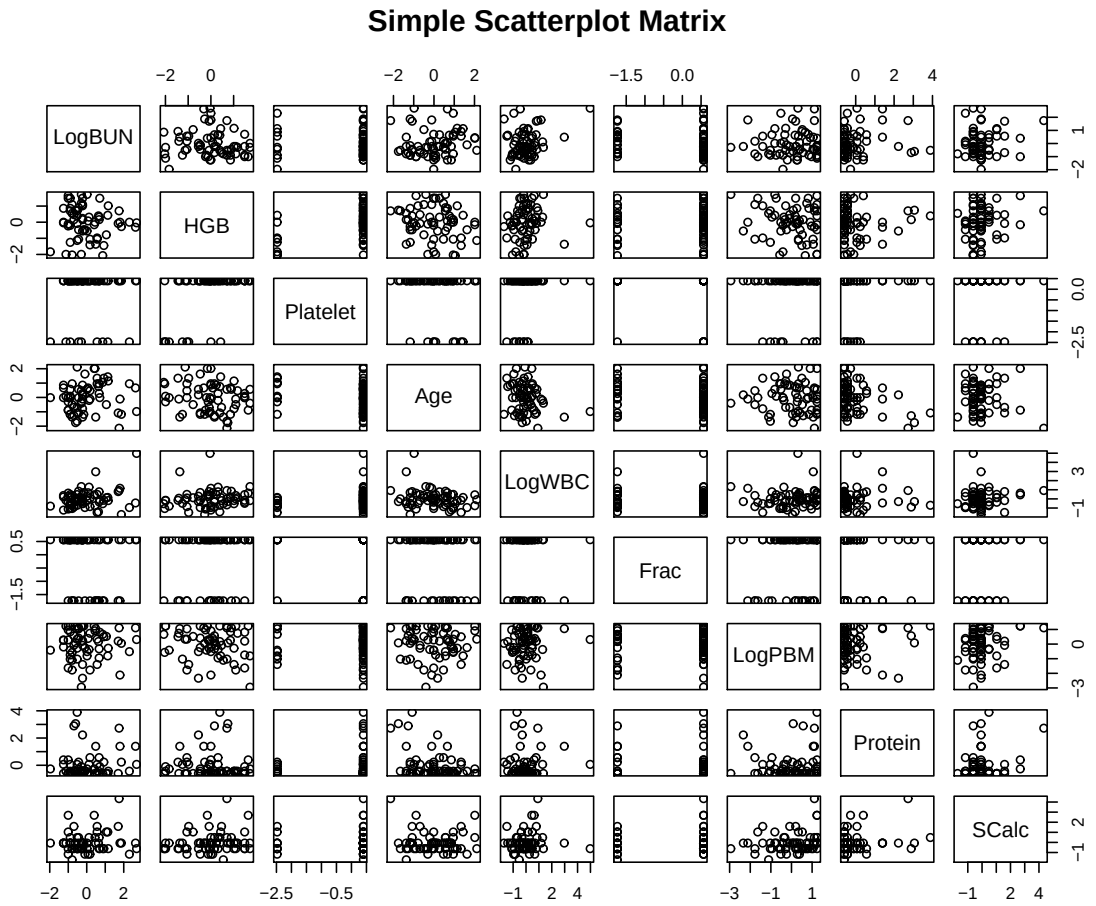
متغیرها	β	$\exp\{\beta\}$	$s.e.(\hat{\beta})$
LogBUN	۰/۵۸	۱/۷۸	۰/۲۰۵۲
HGB	−۰/۳۲	۰/۷۲	۰/۱۸۴۵
PLATELET	−۰/۰۸	۰/۹۱	۰/۱۷۸۲
AGE	−۰/۱۳	۰/۸۷	۰/۲۰۲۲
LogWBC	۰/۰۸	۱/۰۸	۰/۱۷۰۴
FRAC	۰/۱۴	۱/۱۶	۰/۱۷۶۸
LogPBM	۰/۱۳	۱/۱۴	۰/۱۷۷۵
PROTEIN	۰/۰۷	۱/۰۸	۰/۱۵۷۱
SCALC	۰/۲۳	۰/۱۹	۰/۱۹۰۷

نمودار پراکنش متغیرهای مستقل مجموعه داده‌ی بیماران سرطان مغز استخوان در شکل ۱.۴ نشان داده شده است. بر اساس این نمودار، بین متغیرهای HGB و LogWBC و همین‌طور بین متغیرهای HGB و AGE رابطه‌ی خطی دیده می‌شود.

رگرسیون cph با جریمه ریج روی این مجموعه داده پیاده‌سازی شد و نتایج آن در جدول ۲.۴ بیان شده است. انحراف استاندارد ضرایب در این جدول، از روش بوت‌استرپ به‌دست آمده است. هدف این مدل، پیدا کردن یک مدل آگاهی‌بخش برای زمان بقا بیماران براساس داده‌های سرطان مغز استخوان است. علاوه بر این، بررسی اثرهای متغیرهای مستقل بر بقای بیماران، هدف رگرسیون cph جریمه شده است. از آنجایی که جریمه ریج ضرایب را به سمت صفر هدایت می‌کند برای تعیین برآورد ضرایب مدل خطرات متناسب کاکس با جریمه ریج از تابع درستنمایی جزیی جریمه شده استفاده شده است که در بین آن‌ها، هرکدام که مقدار $\exp\{\beta\}$ از یک بزرگتر باشد باعث افزایش خطر مرگ در بیماری است. $\exp\{\beta\}$، نسبت خطر را برای هر متغیر نشان می‌دهد.

مدل رگرسیون کاکس با جریمه ریج در رابطه‌ی (۳.۲) به‌صورت زیر برآورد می‌شود:

$$\begin{aligned}\ln\left(h(t,\boldsymbol{X})\right) &= \ln\left(h_{\circ}(t)\right) + ۰/۱۳\text{LogBUN} - ۰/۱۵\text{HGB} - ۰/۰۸\text{PLATELET} \\ &\quad - ۰/۰۶\text{AGE} + ۰/۰۵\text{LogWBC} + ۰/۲۱\text{FRAC} + ۰/۰۳\text{LogPBM}\end{aligned}$$



شکل ۱.۴: نمودار پراکنش داده‌های بیماران سرطان مغز استخوان

$$+۰/۱۲\text{PROTEIN} + ۰/۱۲\text{SCALC} \qquad (۱.۴)$$

در نتیجه می‌توان مدل رگرسیونی cph را در رابطه‌ی (۲.۲) به صورت زیر نوشت:

$$\begin{aligned} h(t, \boldsymbol{X}) \quad = \quad & h_۰(t) \exp\{۰/۱۳\text{LogBUN} - ۰/۱۵\text{HGB} - ۰/۰۸\text{PLATELET} \\ & -۰/۰۶\text{AGE} + ۰/۰۵\text{LogWBC} + ۰/۲۱\text{FRAC} + ۰/۰۳\text{LogPBM} \\ & +۰/۱۲\text{PROTEIN} + ۰/۱۲\text{SCALC}\} \end{aligned} \qquad (۲.۴)$$

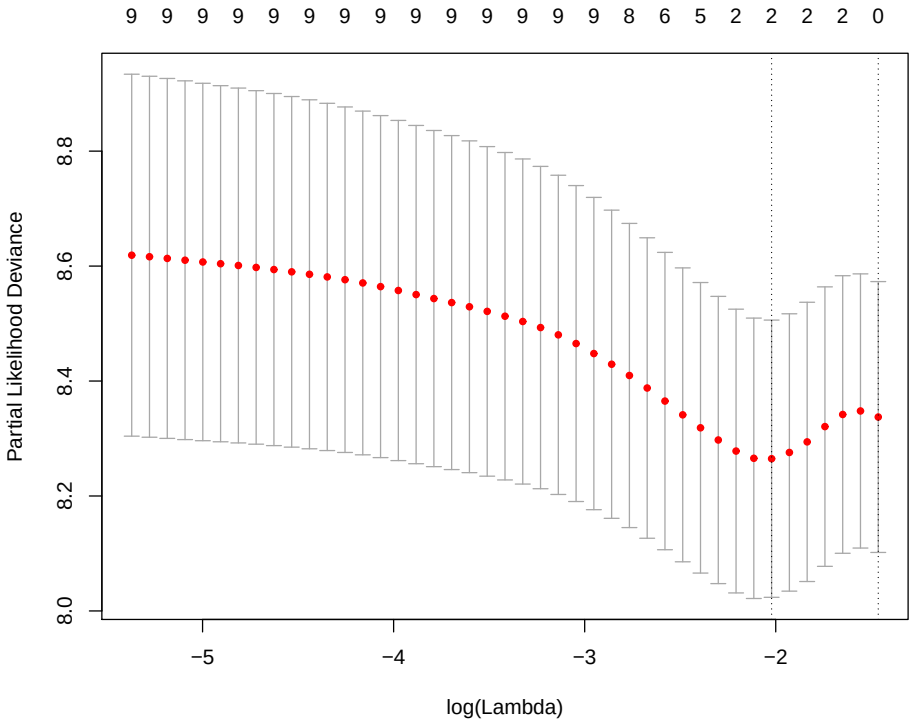
با توجه به نتایج به‌دست آمده از جدول ۲.۴، میزان خطر مرگ افراد با یک واحد افزایش در LogBUN، LogWBC، FRAC، LogPBM، PROTEIN و SCALC به‌ترتیب به اندازه‌ی ۱/۱۴، ۱/۰۵، ۱/۲۳، ۱/۰۳، ۱/۱۲ و ۱/۱۳ برابر افزایش می‌یابد که مقدار بزرگتری از یک است و نشان می‌دهد که با افزایش عوامل گفته شده خطر مرگ در بیماران سرطان مغز استخوان افزایش

جدول ۲.۴: برآورد ضرایب مدل cph با جریمه ریج مجموعه داده‌های استاندارد شده بیماران سرطان مغز استخوان

متغیرها	β	$\exp\{\beta\}$	$s.e.(\hat{\beta})$
LogBUN	۰/۱۳	۱/۱۴	۰/۰۰۶۷
HGB	−۰/۱۵	۰/۸۵	۰/۰۰۷۳
PLATELET	−۰/۰۸	۰/۹۱	۰/۰۰۶۸
AGE	−۰/۰۶	۰/۹۴	۰/۰۷۱
LogWBC	۰/۰۵	۱/۰۵	۰/۰۰۷۲
FRAC	۰/۲۱	۱/۲۳	۰۰۷۰
LogPBM	۰/۰۳	۱/۰۳	۰/۰۰۶۵
PROTEIN	۰/۱۲	۱/۱۲	۰/۰۰۶۹
SCALC	۰/۱۲	۱/۱۳	۰/۰۰۶۶

می‌یابد.

در نمودار ۲.۴، بر اساس معیار انحراف درستنمایی جزیی مدل اشباع‌شده از لگاریتم درستنمایی مدل کامل است. منظور از مدل اشباع شده، مدلی است که هر مشاهده یک پارامتر آزاد دارد. از این معیار برای تعیین λی بهینه در رگرسیون کاکس‌ـ لاسو استفاده می‌گردد.



شکل ۲.۴: نمودار اعتبارسنجی متقابل انحراف لگاریتم درستنمایی جزیی برای مجموعه داده‌های سرطان مغز استخوان. انحراف‌های استاندارد برای این داده‌ها نیز رسم شده است. خط عمودی نقطه‌چین مقادیر λ با کمترین انحراف (خط چپ) و بزرگترین λ با یک انحراف معیار از مینیمم انحراف‌ها.

جدول ۳.۴، ضرایب لاسو را نشان می‌دهد. انحراف استاندارد ضرایب با استفاده از روش بوت‌استرپ به‌دست آمده است.

برای تعیین برآورد ضرایب مدل خطرات متناسب کاکس با جریمه لاسو از تابع درستنمایی جزیی جریمه شده استفاده شده است که در بین آن‌ها، هرکدام که مقدار $\exp\{\boldsymbol{\beta}\}$ از یک بزرگتر باشد باعث افزایش خطر مرگ در بیماری است. $\exp\{\beta\}$، نسبت خطر را برای هر متغیر نشان می‌دهد.

مدل رگرسیون کاکس با جریمه لاسو در رابطه‌ی (۳.۲) به‌صورت زیر برآورد می‌شود:

$$\begin{aligned}\ln\left(h(t,\boldsymbol{X})\right) &= \ln\left(h_{\cdot}(t)\right) + \cdot/۲۱\,\text{LogBUN} - \cdot/۱۳\,\text{HGB} + \cdot\,\text{PLATELET} \\ &\quad - \cdot\,\text{AGE} + \cdot\,\text{LogWBC} + \cdot\,\text{FRAC} + \cdot\,\text{LogPBM} \\ &\quad + \cdot\,\text{PROTEIN} + \cdot\,\text{SCALC} \qquad (۳.۴)\end{aligned}$$

جدول ۳.۴: برآورد ضرایب مدل cph با جریمه لاسو مجموعه داده‌های استاندارد شده بیماران سرطان مغز استخوان

متغیرها	β	$\exp\{\beta\}$	$s.e.(\hat{\beta})$
LogBUN	۰/۲۱	۱/۲۴	۰/۰۱۹۶
HGB	−۰/۱۳	۰/۸۷	۰/۰۰۱۵
PLATELET	۰/۰۰	۱/۰۰	۰/۰۰۰۰
AGE	۰/۰۰	۱/۰۰	۰/۰۰۰۰
LogWBC	۰/۰۰	۱/۰۰	۰/۰۰۰۰
FRAC	۰/۰۰	۱/۰۰	۰۰۰۰
LogPBM	۰/۰۰	۱/۰۰	۰/۰۰۰۰
PROTEIN	۰/۰۰	۱/۰۰	۰/۰۰۰۰
SCALC	۰/۰۰	۱/۰۰	۰/۰۰۰۰

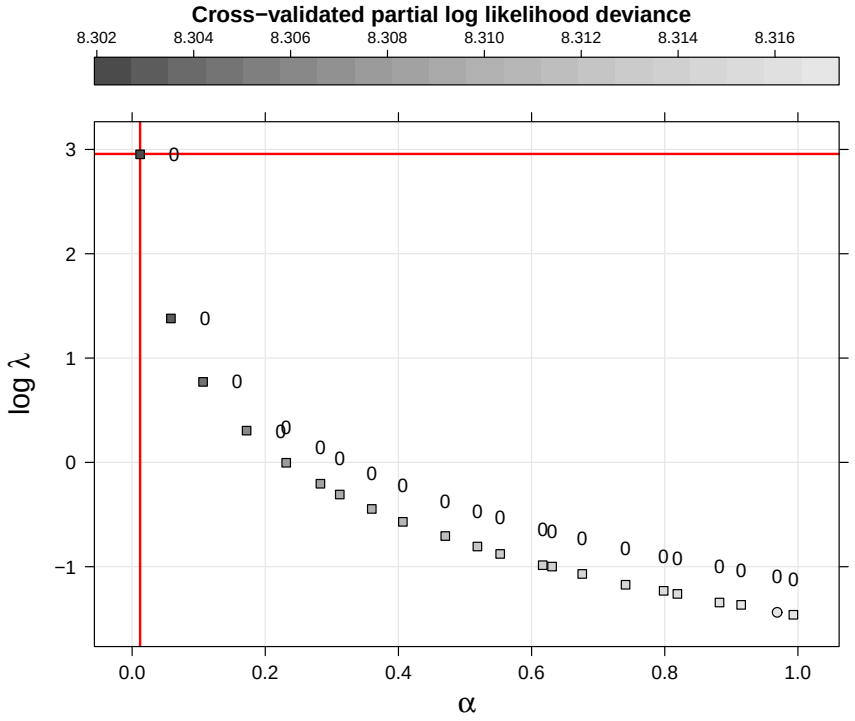
در نتیجه می‌توان مدل رگرسیونی cph جریمه شده را در رابطه‌ی (۲.۲) به صورت زیر نوشت:

$$\begin{aligned} h(t,\boldsymbol{X}) \;=\; & h_{\cdot}(t)\exp\{\cdot/۲۱\,\text{LogBUN} - \cdot/۱۳\,\text{HGB} + \cdot\,\text{PLATELET} \\ & + \cdot\,\text{AGE} + \cdot\,\text{LogWBC} + \cdot\,\text{FRAC} + \cdot\,\text{LogPBM} \\ & + \cdot\,\text{PROTEIN} + \cdot\,\text{SCALC}\} \end{aligned} \qquad (۴.۴)$$

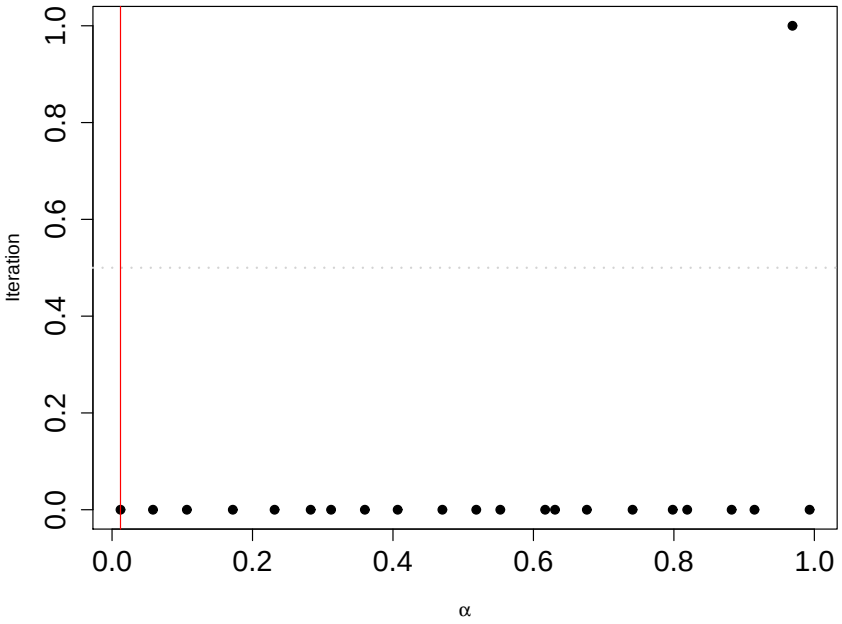
با توجه به نتایج به‌دست آمده از جدول ۳.۴، میزان خطر مرگ افراد با یک واحد افزایش در LogBUN به اندازه‌ی ۱/۲۴ برابر افزایش می‌یابد و با یک واحد افزایش در HGB به اندازه ۰/۸۷ کاهش می‌یابد. که برای مقدار بزرگتر از یک نشان می‌دهد که با افزایش عوامل گفته شده خطر مرگ در بیماران سرطان مغز استخوان افزایش می‌یابد و برای مقدار کوچک‌تر از یک نشان می‌دهد که با افزایش عوامل گفته شده خطر مرگ در بیماران سرطان مغز استخوان کاهش می‌یابد.

نمودارهای ۳.۴ و ۴.۴، نشان می‌دهد مقادیر بهینه‌ی بردار پارامترهای تنظیم‌کننده‌ی (α,λ) را در جریمه‌ی کاکس به‌صورت $(۰/۰۱۲, ۱۹/۲۳۸۶)$ به‌دست می‌آید.

با این پارامترها، رگرسیون الاستیک‌نت هیچ یک از متغیرها را برای حضور در مدل انتخاب نمی کند. جدول ۴.۴، مدل‌های مختلف رگرسیون کاکس جریمه شده را بر اساس معیار AIC مقایسه می‌کند که نشان می‌دهد رگرسیون لاسو عملکرد بهتری دارد.



شکل ۳.۴: نمودار انحراف درستنمایی جزیی به عنوان تابعی از هر دو پارامتر تنظیم‌کننده در جریمه‌ی الاستیک نت برای مجموعه داده‌ی بیماران سرطان مغز استخوان



شکل ۴.۴: نمودار توزیع نقاط پارامتر تنظیم‌کننده در جریمه‌ی الاستیک‌نت برای مجموعه داده‌ی بیماران سرطان مغز استخوان

جدول ۴.۴: مقایسه‌ی برآوردگرها برای مجموعه داده‌های بیماران سرطان مغز استخوان

	کاکس معمولی	کاکس ریج	کاکس لاسو
AIC	−۳۲۶/۳۸۹۵	−۳۶۰/۰۳۸۱	−۴۹۳/۱۰۰۴

۳.۴ داده‌های ژنومی

در این بخش، برآوردگرهای معرفی شده در فصل ۳ در مجموعه داده‌های ژنومی و داده‌های بالینی بیماران سیتوژنتیکی طبیعی لوسمی میلوئید حاد (AML) از متزلر و همکاران (۲۰۰۸) به‌کار برده شده است. این داده‌ها از پایگاه داده GEO[۱] با کد GSE12417 قابل فراخوانی است.

این داده‌ها مربوط به ۷۹ بیمار است که با استفاده از تکنولوژی ریزآرایه‌ای Affymetrix HG-U133 Plus 2.0 اندازه‌گیری شده است.

به دلیل راحتی، تعداد ۵۴۶۷۵ ویژگی ژن‌ها که در اینجا متغیرها را می‌سازند به ۱۰۰۰۰ ویژگی‌ای که بیشترین تغییرات را داشتند، کاهش داده شد.

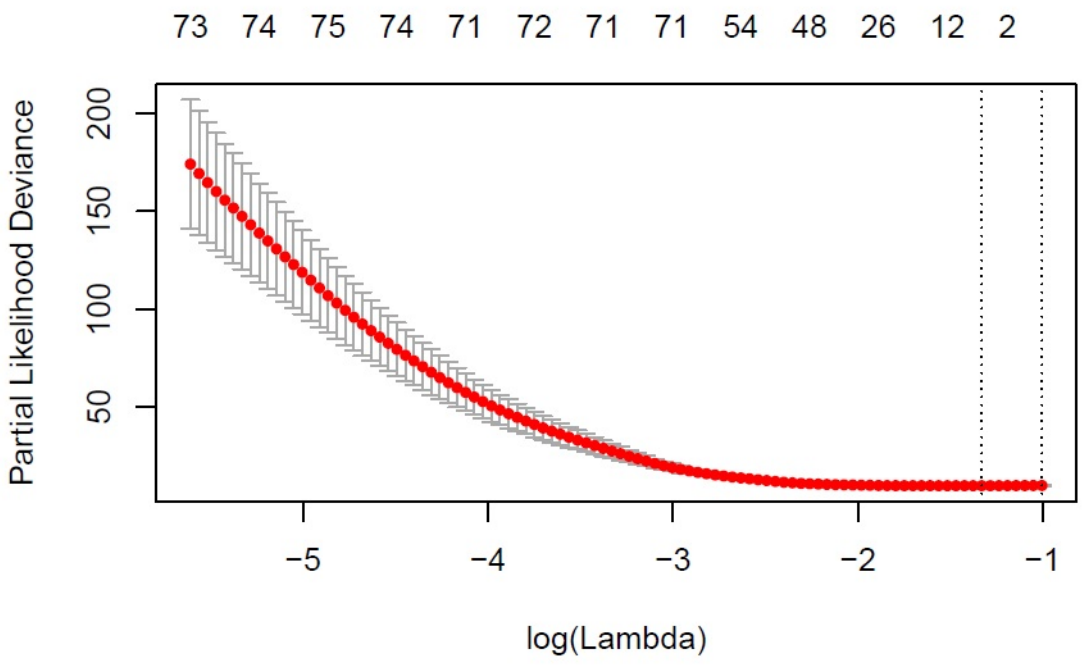
برای همه محاسبات مجموعه داده‌ها با عنوان ExpressionSet به نام eset ذخیره می‌شوند. ماتریس داده‌های ژن‌ها را می‌توان از طریق نامیدن exprs و داده‌های بقا کلی و سایر داده‌های مخصوص بیمار (به عنوان مثال، سن بیمار) در phenoData شخص ذخیره شده‌اند که توسط pData(eset) به‌دست آورد. زمان بقا کلی در متغیر os ذخیره شده‌است، متغیر وضعیت بقا مربوط os-status نامیده شده است و متغیر سن بیمار، age است.

فراخوانی این داده‌ها متفاوت از سایر داده‌ها در نرم‌افزار R است که در ضمیمه آمده است.

هدف این مطالعه، پیدا کردن یک مدل آگاهی‌بخش برای زمان بقای بیماران بر اساس داده‌های ژنومی است. علاوه بر پیش‌بینی زمان بقای کلی، به تعیین ژن‌های موثر در این‌باره هدف دیگر تحقیق است. ابتدا، مسئله‌ی انتخاب متغیر با برازش مدل رگرسیونی کاکس جریمه شده با جریمه‌ی لاسو انجام شده است. از تابع glmnet برای برازش مدل cph با جریمه‌ی لاسو به مجموعه داده AML استفاده شده است. برای تعیین پارامتر هموارسازی، از روش اعتبارسنجی متقابل ۱۰ـ تایی استفاده شده است. ۱۰۰ مقدار مختلف برای λ بر اساس دامنه‌ی داده‌ها درنظرگرفته شده است و در بین آنها، بر اساس نمودار ۵.۴ مقدار ۰/۲۶۵ برای λ انتخاب شد.

[۱] http://www.ncbi.nlm.nih.gov/geo/

($\log(\lambda) = -۱/۳۲۹$) بر این اساس، مدل رگرسیون لاسوی نهایی فقط ۵ متغیر (ژن) را انتخاب کرده است و برآوردهای ضرایب لاسو در این مدل در جدول ۵.۴ آمده است.

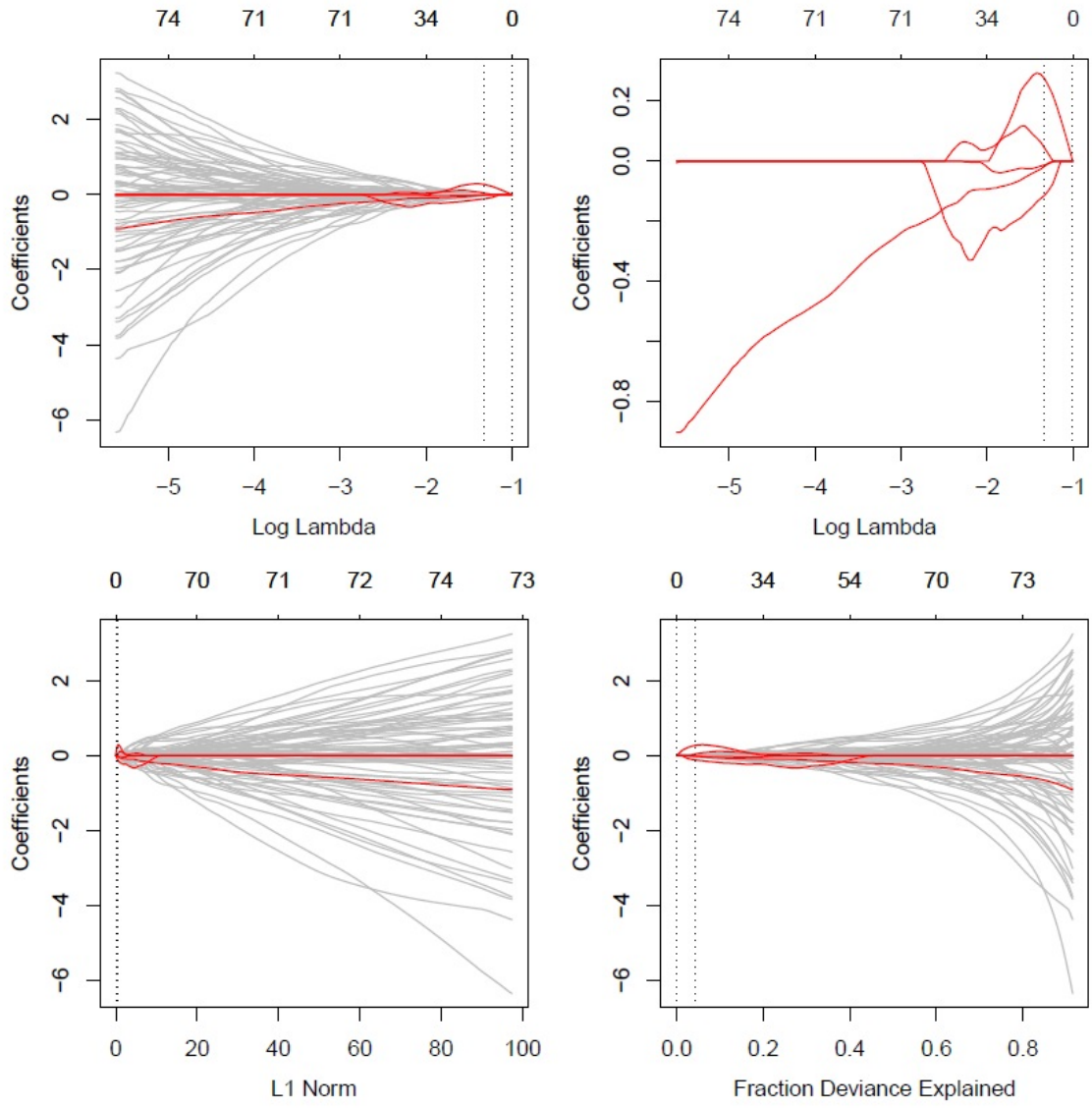


شکل ۵.۴: نمودار اعتبارسنجی متقابل انحراف لگاریتم درستنمایی جزیی برای مجموعه داده‌های AML. انحراف‌های استاندارد برای این داده‌ها نیز رسم شده است. اولین خط عمودی نقطه‌چین مقدار λ با کمترین انحراف است (خط چپ) و دومین خط (خط راست) بزرگترین λ با یک انحراف معیار از مینیمم انحراف‌ها را نشان می‌دهد.

جدول ۵.۴: مقدار ضرایب برآوردگر لاسو در مدل رگرسیون cph با جریمه‌ی لاسو

متغیرها	233371-at	226169-at	222462-s-at	204419-x-at	203640-at
برآورد لاسو	−۰/۰۱۲۲	۰/۰۴۳۰	۰/۲۷۴۲	−۰/۰۱۶۶	−۰/۱۱۳۴

در شکل ۶.۴، مسیر متغیرهای انتخاب شده با رنگ قرمز دیده می‌شود که بهبودی برآورد ضرایب رگرسیونی را با افزایش پارامتر تنظیم‌سازی λ نشان می‌دهد. در ابتدا بیشتر برآوردها صفر هستند و هر چه مقدار λ افزایش پیدا می‌کند، این برآوردگرها از صفر فاصله می‌گیرند. با وجودی که ۵ متغیر انتخاب شده، فقط متغیرهایی هستند که با مقدار بهینه‌ی λ انتخاب شده‌اند، وقتی جریمه کاهش می‌یابد، آن‌ها در بین متغیرها با بیشترین اثر باقی نمی‌مانند و متغیرهای بیشتری شروع می‌کنند تا وارد مدل شوند. در حقیقت، از بین ۵ متغیر، ۴تای آن‌ها به ازای مقادیر کوچک $log(\lambda)$ به صفر نزدیک می‌شود و این بدین معنی است که در مدل‌های خیلی بزرگ این چهار متغیر می‌تواند با متغیر دیگر جایگزین شود، به‌عبارت دیگر، متغیرها پایدار نیستند و می‌توانند با دیگر متغیرها در مدل‌های خیلی بزرگ جایگزین شوند.



شکل ۶.۴: مسیر ضرایب برای مدل‌های رگرسیون کاکس با جریمه‌ی لاسو برای داده‌های AML. متغیرها با مسیرهای قرمز رنگ، ضرایب غیرصفر در مدل با λی بهینه که از اعتبارسنجی متقابل ۱۰تایی به‌دست آمده است، را نشان می‌دهد. نمودارهای بالا مسیر ضرایب را بر اساس $log\lambda$ نشان می‌دهد. (بالا_چپ: مسیر کامل، بالا_راست: فقط مسیر متغیرهای انتخاب شده را نشان می‌دهد). نمودارهای پایین مسیر ضرایب را نسب به نرم L_1 (تابع جریمه لاسو) از بردار ضرایب برآورد شده (چپ) و نسبت انحراف لگاریتم درستنمایی جزیی اولیه توضیح داده‌شده (راست) را نشان می‌دهد. خط‌های عمودی نقطه‌چین، λهایی با کمترین انحراف‌ها و با یک انحراف معیار از مینیمم انحراف‌ها را نشان می‌دهد.

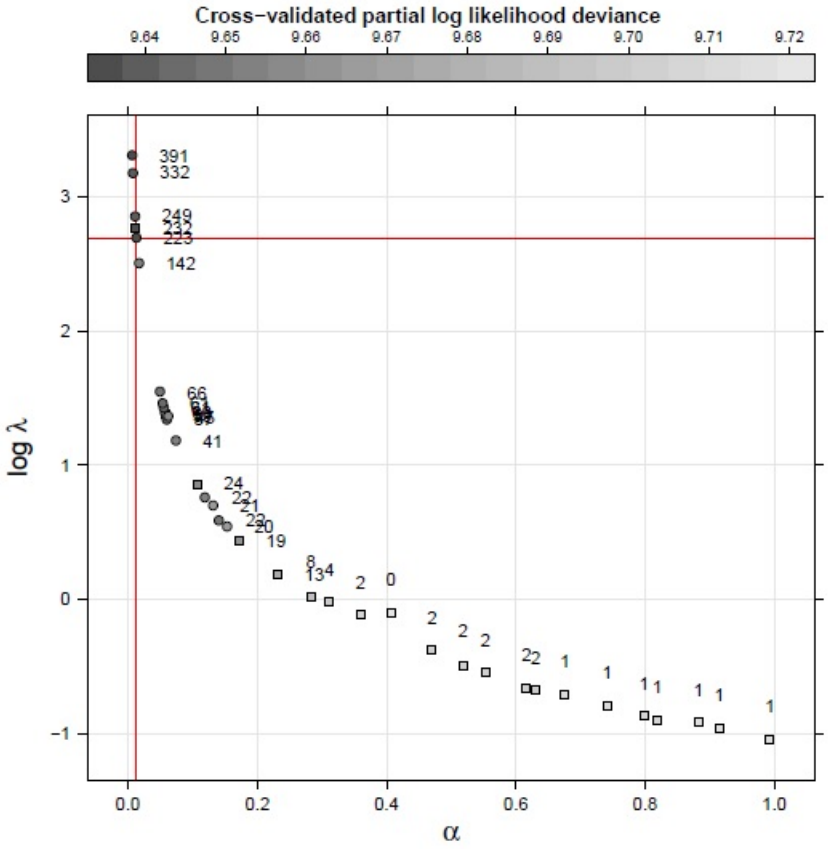
مدل رگرسیون جریمه‌شده‌ی کاکس برای پیش‌بینی داده‌های بقا خیلی خوب رفتار نمی‌کند. مدل لاسو که بر اساس λی بهینه از اعتبارسنجی متقابل ۱۰تایی شناخته شده است بسیار تنک است و فقط ۵ متغیر را در بر می‌گیرد. همان‌طور که اشاره شد، پایدار نیز نیست.

در ادامه، یک مدل رگرسیونی کاکس با جریمه الاستیک‌نت به جای جریمه‌ی لاسو را به داده‌های AML برازش داده شده است. همان‌طور که در تابع جریمه الاستیک‌نت مشاهده می‌شود، برای این تابع جریمه، باید دو پارامتر تنظیم‌کننده‌ی α و λ به‌طور همزمان مشخص شوند. با استفاده از الگوریتم جستجوی بازه‌ای با استفاده از تابع epsgo در بسته‌ی penalizedSVM، این دو پارامتر برابر با $\alpha = ۰/۰۱۳$ و $\log(\lambda) = ۲/۶۸۹$ یا $\lambda = ۱۴/۷۲$تعیین می‌شوند.

مدل لاسوی جریمه شده با جریمه لاسو از نظر پیش‌بینی بقا کلی برای مجموعه داده‌های AML به خوبی عمل نمی‌کند. مدل لاسو به‌عنوان مدل بهینه با روش اعتبارسنجی متقابل ۱۰تایی که خیلی تنک و دارای ویژگی است معرفی شده است. علاوه بر این ما مشاهده کردیم که این ویژگی‌ها پایدار نیستند و ۴تا از آنها در مجموعه‌ای از ویژگی های انتخابی هنگامی که مقدار تنظیمات کاهش می‌یابد باقی نمی‌مانند و ویژگی‌های بیشتری شروع به وارد شدن به مدل می کنند.

در ادامه، مدل الاستیک‌نت به جای مدل لاسو به داده‌های AML برازش داده شده است. همانگونه که گفته شد الاستیک‌نت نیاز به دو پارامتر همزمان تنظیم α و λ دارد. هدف پیدا کردن مقدار پارامترهای تنظیم‌کننده‌ی (α, λ) است، برای اینکه انحراف اعتبارسنجی متقابل ۱۰تایی لگاریتم درستنمایی جریمه شده مدل کمینه گردد.

شکل ۷.۴، نمودار اعتبارسنجی متقابل انحراف لگاریتم تابع درستنمایی را به‌ازای هر دو پارامتر α و λ نشان می‌دهد. مدل نهایی رگرسیون کاکس با جریمه‌ی الاستیک‌نت، ۲۲۰ متغیر دارد که مشاهده می‌شود تنکی کمتری در مقایسه با مدل لاسوی نهایی دارد. متغیرهای انتخاب شده با جریمه‌ی لاسو در بین متغیرهای انتخابی با الاستیک‌نت نیز حضور دارند. همچنین ضرایب متغیرهای انتخاب شده با جریمه‌ی الاستیک‌نت، پایدار می‌باشند. بنابراین به جای اندازه‌گیری ۱۰۰۰۰ ژن، می‌توان فقط ۲۲۰ ژن را بررسی کرد.



شکل ۷.۴: نمودار اعتبارسنجی متقابل انحراف لگاریتم درستنمایی جزیی برای مجموعه داده‌های AML برای (α, λ). برای هر نقطه‌ی محاسبه شده در فضای پارامتر، تعداد متغیرهای انتخاب شده در کنار آن نوشته شده است. خط قرمز، پاسخ نهایی را نشان می‌دهد که تابع خطا به اندازه‌ی یک انحراف معیار از مینیمم فاصله دارد.

۴.۴ پیشنهادات و آینده تحقیق

در این پایان‌نامه فرض شده است که زمان‌های تکراری (گره‌ها) در داده‌ها وجود ندارد. بررسی رگرسیون cph جریمه شده برای داده‌های بقا با وجود گره‌ها در آن می‌تواند موضوع تحقیق دیگری باشد.

متغیرهای پیش‌گو در رگرسیون خطرات متناسب کاکس جریمه‌شده در این پایان‌نامه فرض شده است که مستقل از زمان باشند اما یکی از قدرتمندی‌های رگرسیون cph توانایی آن برای تحلیل متغیرهایی است که با زمان در طی دوره‌ای که پژوهش انجام می‌شود، تغییر می‌کنند.

بررسی تکنیک جریمه‌شده بر روی این داده‌ها می‌تواند در آینده مورد توجه قرار بگیرد.

فرضیه‌ی خطرات متناسب نیز می‌تواند از رگرسیون cph نادیده گرفته شود و از رگرسیون کاکس لایه‌ای در این حالت استفاده شود. انجام تکنیک جریمه‌شده بر روی این رگرسیون، موضوع تحقیق دیگری است.

می‌توان از تکنیک جریمه‌شده بر روی مدل‌های پارامتری مانند وایبل، نمایی و ... نیز استفاده کرد و آن‌ها را با مدل cph جریمه شده مقایسه کرد.

در این پایان‌نامه فقط از توابع جریمه‌ی ریج، لاسو و الاستیک‌نت در رگرسیون خطرات متناسب کاکس استفاده شد. اما توابع جریمه‌ی دیگری نیز مانند تابع جریمه‌ی اسکد (فن و لی، ۲۰۰۱) را نیز استفاده کرد.

یکی از ویژگی‌های مهم برآوردگرهای جریمه‌شده، خاصیت پیش‌گویی[۲] آن‌ها است. این خاصیت بیان می‌کند که برآوردگر می‌تواند ضرایب رگرسیونی که واقعا صفر هستند را به‌طور خودکار دقیقا صفر برآورد کند و سایر ضرایب رگرسیونی برآورد شوند، اگر زیرمدل اصلی را از قبل بدانیم. فن و لی (۲۰۰۱) نشان دادند که برآوردگر الاستیک‌نت خاصیت پیش‌گویی ندارد. بنابراین، ژو و ژانگ[۳] (۲۰۰۹) روش الاستیک‌نت سازوار را پیشنهاد دادند. لی[۴] (۲۰۱۵) نشان دادند که برآوردگر الاستیک‌نت سازوار در مدل رگرسیونی کاکس این ویژگی را دارد.

اضافه کردن تابع جریمه‌ی تغییر یافته‌ی الاستیک‌نت منجر به برآوردگر الاستیک‌نت سازوار[۵] برای مدل رگرسیونی خطرات متناسب کاکس می‌شود

$$\hat{\boldsymbol{\beta}}^{\text{AEN}} = \text{argmax}_{\boldsymbol{\beta}} \left\{ l(\boldsymbol{\beta}) + \lambda_1^* \sum_{j=1}^{p} \hat{w}_j |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}$$

که $\lambda_1^* \geq 0$ و $\lambda_2 \geq 0$ پارامترهای تنظیم‌کننده هستند و $\hat{w}_j = \left(\hat{\beta}_i^{\text{EN}}\right)^{-1}$ را برآوردگر الاستیک‌نت سازوار قرار دهید، برآوردگرهای جریمه‌ی شده‌ی مطرح شده را در بر می‌گیرد. بدین منظور، (۱) اگر $\lambda_1^* = 0$، آنگاه رابطه‌ی بالا منجر به برآوردگر کاکس ریج می‌شود. (۲) انتخاب $\lambda_2 = 0$ و $\hat{w}_j = 1$ (به ازای هر j)، برآوردگر لاسو در مدل رگرسیونی کاکس را نتیجه می‌دهد. (۳) اگر $\lambda_1^*, \lambda_2 \neq 0$ و $\hat{w}_j = 1$ (به ازای هر j) باشد، در این صورت برآوردگر الاستیک‌نت سازوار به

[۲] Oracle
[۳] Zou and Zhang
[۴] Li
[۵] Adaptive Elastic Net Estimator

برآوردگر الاستیک‌نت معمولی ($\hat{\beta}^{EN}$) تبدیل می‌شود. بنابراین، این برآوردگر، حالت کلی‌تری از برآوردگرهای مطرح شده در این پایان‌نامه است و از عملکرد بهتری در مدل‌های بعد بالا و تنک برخوردار است که موضوع جالبی برای ادامه‌ی کار این پایان‌نامه است.

پیوست آ

برنامه‌های نوشته‌شده با نرم‌افزار R

کد مربوط به شکل۱.۱ :

```
Data <- read.table("https://onlinecourses.science.psu.edu/stat501/sites/
onlinecourses.science.psu.edu.stat501/files/data/skincancer.txt",header=T)
attach(Data)
Y <- Mort
X <- Lat
plot(X,Y, pch = 19 , xlab = "Latitude (at center of State)",
            ylab = "Mortality (Death per 10 million)",
            main = "Skin Cancer Mortality versus State Latitude")
abline(lm(Y~X),col="red",lwd= 2)
text(45,220,expression(paste(widehat(Y), "= 389.19 - 5.98", X)))
```

کد مربوط به شکل ۲.۱:

```
library(scatterplot3d)
Data <- read.table("https://onlinecourses.science.psu.edu/stat501/sites/
onlinecourses.science.psu.edu.stat501/files/data/skincancer.txt",header=T)
attach(Data)
Y <- Mort
X1 <- Lat
X2 <- Long
s3d <-scatterplot3d(X1,X2,Y, pch=19,type = "p",
              xlab = "Latitude (X1)", ylab = "Longitude (X2)", grid = FALSE,
               zlab = "Mortality (Y)",
           main="Skin Cancer Mortality versus State Latitude and Longitude")
fit<-lm(Y~X1+X2)
s3d$plane3d(fit, draw_polygon=TRUE)
mtext(expression(paste(hat(Y), " = 400.67 -5.93 X1 -0.15 X2")))
```

کد مربوط به مثال ۱.۳.۱:

```
set.seed(1)
X1 <- rnorm(20)
X2 <- rnorm(20,mean = X1, sd = 0.01)
cor(X1,X2)
E <- rnorm(20)
Y <- 3 + X1 + X2 + E
OLS.model <- lm(Y~X1+X2)
OLS<-coef(OLS.model); OLS
sum((Y- OLS[1] - OLS[2]*X1 - OLS[3]*X2)^2)
```

```
sum((Y- 3 - 1*X1 - 1*X2)^2)

OLS[3] + OLS[2]

(OLS[2])^2 + (OLS[3])^2

ridge.model<-lm.ridge(Y~X1+X2,lambda =1)

ridge<-coef(ridge.model); ridge
```

کد مربوط به شکل ۳.۱:

```
curve(1-plnorm(x),0,10,xlim=c(0,10),ylim=c(0,1),xlab = "t",

          ylab = "Survival Function")

text(0.7,1, expression(S(0) == 1))

text(9.5,0.05, expression(S(infinity) == 0))

abline(h = 0, lty= 2)
```

کد مربوط به شکل ۴.۱:

```
x <- c(1,2,3)

sfun  <- stepfun(x,c(1,2/3,1/3,0))

plot(sfun,xlab = "t",ylab = "Survival function",main = " ")
```

کد مربوط به شکل ۵.۱:

```
h<-function(t,p,lambda) p*(lambda^p)*(t^(p-1))

curve(h(x,1,1),0,2,lwd = 2,ylab = "Hazard function",

      xlab = "Time",ylim = c(0,3))

curve(h(x,1.5,1),0,2,col = "red", lwd = 2,lty =2, add=TRUE)

curve(h(x,0.5,1),0,2,col = "blue", lwd = 2, lty =3, add=TRUE)

legend(1.5,3,c("p = 1", "p > 1", "p < 1"),

       col = c("black","red","blue"),lwd = 2,lty = c(1,2,3))
```

کد مربوط به بخش ۷.۲:

```
library(survival)
myeloma=read.table("myeloma-65data.txt",header=T)
head(myeloma)
summary(myeloma)
attach(myeloma)
cox.myeloma<-coxph(Surv(Time,VStatus)~LogBUN + HGB + Platelet + Age
            + LogWBC + Frac + LogPBM + Protein + SCalc, data=myeloma)
cox.zph(cox.myeloma)
par(mfrow=c(3,3))
plot(cox.zph(cox.myeloma))
cox.myeloma
plot(survfit(cox.myeloma),conf.int = F,xlab = "Time",
       ylab = "Estimated survavial function")
```

کد مربوط به جدول ۱.۴:

```
set.seed(1234)
library(CoxRidge)
library(glmnet)
library(survival)
Myeloma = read.table("myeloma-65data.txt",header=T)
attach(Myeloma)
Myeloma[,3:11] <- apply(X,2,function(x){(x-mean(x))/sqrt(var(x))})
cox.Myeloma<-coxph(Surv(Time,VStatus)~LogBUN + HGB + Platelet + Age
                + LogWBC + Frac + LogPBM + Protein + SCalc, data=Myeloma)
cox.Myeloma
```

کد مربوط به شکل ۱.۴:

```
Myeloma =read.table("myeloma-65data.txt",header=T)
attach(Myeloma)
X <- cbind(LogBUN , HGB , Platelet , Age ,
           LogWBC , Frac , LogPBM , Protein , SCalc)
pairs(X, main="Simple Scatterplot Matrix")
```

کد مربوط به جدول ۲.۴:

```
set.seed(1234)
library(CoxRidge)
library(glmnet)
library(survival)
Myeloma = read.table("myeloma-65data.txt",header=T)
attach(Myeloma)
X <- cbind(LogBUN , HGB , Platelet , Age , LogWBC , Frac , LogPBM , Protein , SCalc)
X <- apply(X,2,function(x){(x-mean(x))/sqrt(var(x))})
fit.ridge=cox.ridge(Surv(Time,VStatus)~X,lambda=1);fit.ridge
N.boot = 1000
B.ridge <- matrix(0,nr= 9, nc = N.boot)
i = 1
while(i < N.boot + 1){
  index<-sample(1:nrow(Myeloma),nrow(Myeloma),replace = T)
  D.boot <- Myeloma[index,]
  fit=cox.ridge(Surv(Time,VStatus)~.,lambdaFixed = T,data = D.boot)
  B.ridge[,i]<-fit$coef
  i = i+1 }
```

```
apply(B.ridge,1,sd)/sqrt(N.boot)
```

کد مربوط به شکل ۲.۴:

```
library(glmnet)
Myeloma =read.table("myeloma-65data.txt",header=T)
attach(Myeloma)
X <- cbind(LogBUN , HGB , Platelet , Age ,
           LogWBC , Frac , LogPBM , Protein , SCalc)
X <- as.matrix(X)
X <- apply(X,2,function(x){(x-mean(x))/sqrt(var(x))})
cv.fit <- cv.glmnet(X, Surv(Time, VStatus),
                    family = "cox", maxit = 1000)
plot(cv.fit)
```

کد مربوط به جدول ۳.۴:

```
library(glmnet)
Myeloma =read.table("myeloma-65data.txt",header=T)
attach(Myeloma)
X <- cbind(LogBUN , HGB , Platelet , Age ,
           LogWBC , Frac , LogPBM , Protein , SCalc)
X <- as.matrix(X)
X <- apply(X,2,function(x){(x-mean(x))/sqrt(var(x))})
cv.fit <- cv.glmnet(X, Surv(Time, VStatus),
                    family = "cox", maxit = 1000)
fit.lasso <- glmnet(X, Surv(Time, VStatus), family = "cox", maxit = 1000)
Coefficients <- coef(fit.lasso, s = cv.fit$lambda.min)
```

```
Coefficients

exp(Coefficients)

N.boot = 1000

B.lasso <- matrix(0,nr= 9, nc = N.boot)

i = 1

while(i < N.boot + 1){

  index<-sample(1:nrow(Myeloma),0.7*nrow(Myeloma),replace = F)

  D.boot <- Myeloma[index,]

  X <- as.matrix(D.boot[,3:11])

  Time <- D.boot[,1]

  VStatus <- D.boot[,2]

 cv.fit.boot <- cv.glmnet(X, Surv(Time, VStatus), family = "cox", maxit = 1000)

 fit.lasso.boot <- glmnet(X, Surv(Time, VStatus), family = "cox", maxit = 1000)

  B.lasso[,i]<-as.matrix(coef(fit.lasso.boot, s = cv.fit.boot$lambda.min))

  i = i+1 }

apply(B.lasso,1,sd)/sqrt(N.boot)
```

کد مربوط به شکل‌های ۳.۴ و ۴.۴:

```
set.seed(1234)

library(CoxRidge)

library(glmnet)

library(c060)

Myeloma = myeloma=read.table("myeloma-65data.txt",header=T)

attach(myeloma)

X <- cbind(LogBUN , HGB , Platelet , Age , LogWBC , Frac , LogPBM , Protein , SCalc)

X <- as.matrix(X)
```

```
X <- apply(X,2,function(x){(x-mean(x))/sqrt(var(x))})
bounds <- t(data.frame(alpha = c(0, 1)))
colnames(bounds) <- c("lower", "upper")
nfolds <- 10
set.seed(1234)
foldid <- balancedFolds(class.column.factor = Y[, 2],cross.outer = nfolds)
fit <- epsgo(Q.func = "tune.glmnet.interval", bounds = bounds,
           parms.coding = "none", seed = 1234, fminlower = -100, x = X, y = Y,
               family = "cox", foldid = foldid, type.min = "lambda.1se",
               type.measure = "deviance")
s.fit <- summary(fit, verbose = TRUE)
plot(s.fit)
plot(s.fit,type = "points")
```

کد مربوط به جدول ۴.۴:

```
p = 9
AIC.cox = 2*cox.Myeloma$loglik -2*p; AIC.cox
AIC.ridge = 2*fit.ridge$loglik-2*p; AIC.ridge
index <- which(fit.lasso$lambda == cv.fit$lambda.min)
AIC.lasso = -2*deviance.glmnet(fit.lasso)[index]-2*2; AIC.lasso
```

کد مربوط به فراخوانی داده‌های ژنومی AML در فصل ۴:

```
source("https://bioconductor.org/biocLite.R")
biocLite("GEOquery")
library(GEOquery)
gds <- getGEO("GSE12417") #download from the site directly
```

```
dataDirectory <- system.file("extdata", package="Biobase")
gsm <- getGEO(filename=system.file("extdata/GSE12417.txt.gz",
                package="GEOquery"))
exprsFile <- file.path(dataDirectory, "exprsData.txt")
eset <- as.matrix(read.table(exprsFile, header=TRUE, sep="\t", row.names=1,
             as.is=TRUE))
pDataFile <- file.path(dataDirectory, "pData.txt")
pData <- read.table(pDataFile,row.names=1, header=TRUE, sep="\t")
```

کد مربوط به شکل ۵.۴:

```
library(c060)
library(glmnet)
library(survival)
cvres <- cv.glmnet(y = Surv(pData(eset)$os, pData(eset)$os_status),
x = t(exprs(eset)), family = "cox", nfolds = 10)
plot(cvres)
```

کد مربوط به جدول ۵.۴:

```
library(c060)
library(glmnet)
library(survival)
fit.lasso <- glmnet(y = Surv(pData(eset)$os, pData(eset)$os_status),
x = t(exprs(eset)), family = "cox")
coef(fit.lasso, s = cvres$lambda.min)
```

کد مربوط به شکل ۶.۴:

```
cof <- coef(res, s = cvres$lambda.min)
 Plot.coef.glmnet(cvfit = cvres, betas = rownames(cof)[which(cof != 0)])
```

کد مربوط به شکل ۶.۴:

```
x <- t(exprs(eset))
y <- cbind(time = pData(eset)$os, status = pData(eset)$os_status)
bounds <- t(data.frame(alpha = c(0, 1)))
colnames(bounds) <- c("lower", "upper")
nfolds <- 10
set.seed(1234)
foldid <- balancedFolds(class.column.factor = y[, 2],cross.outer = nfolds)
fit <- epsgo(Q.func = "tune.glmnet.interval", bounds = bounds,
parms.coding = "none", seed = 1234, fminlower = -100, x = x, y = y,
family = "cox", foldid = foldid, type.min = "lambda.1se",
type.measure = "deviance")
```

کد مربوط به برآوردگر الاستیک‌نت در مثال داده‌های ژنومی

```
fit.enet <- glmnet(y = Surv(pData(eset)$os, pData(eset)$os_status),
x = t(exprs(eset)), family = "cox",alpha = 0.013)
coef(fit.enet, lambda=14.72)
```

مراجع

[۱] احراری خلف، و. (۱۳۹۰). برآورد ناپارامتری تابع بقا تحت داده‌های سانسور شده، پایان نامه کارشناسی ارشد، دانشگاه بیرجند.

[۲] رستا، ر.، چینی‌پرداز، ر. و راسخی، ع. ا. (۱۳۸۹). معرفی روش کمترین توان‌های دوم تاوان داده در انتخاب متغیرهای توضیحی، ششمین همایش ملی آمار دانشگاه پیام نور اهواز، ۲۰۷-۲۱۰.

[3] Abramsky, S. and Jung, A. (1994). Domain Theory, in: S. Abramsky, D.M. Gabbay, T.S.E. Maibaum (Eds.), *Handbook of Logic in Computer Science*, **3**, Clarendon Press, Oxford, 1 - 68.

[4] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Proceeding 2nd Inter Symposium on Information Theory, 267-281, Budapest

[5] Andersen, P. K. (1991). Survival analysis 1982–1991: the second decade of the proportional hazards regression model, *Statistics in Medicine*, **10**, 1931 - 1941.

[6] Burnham, K.P. and Anderson, D.R. (2002). *Model Melection and Multimodel Inference: a Practical Information-Theoretic Approach*, Springer, New York.

[7] Cox, D. R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B 34*, 187 - 220.

[8] Cox, D. R. (1975). Partial likelihood, *Biometrika*, **62**, 269 - 276.

[9] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerische Mathematika*, **31**, 377-403.

[10] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, *Annals of Statistics*, **32**, 407-451.

[11] Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, **96**(4), 1348-1360.

[12] Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information, *Biometrics*, **21**, 826 - 838.

[13] Freedman, D. A. (2008). Survival analysis: a primer, *The American Statistician*, **62**(2), 110-119.

[14] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, **33**(1), 1-22.

[15] Goeman, J. J, (2010). L1 penalized estimation in the cox proportional hazards model, *Biometrical Journal*, **52**, 70–84.

[16] Grambsc, P. and Therneau, T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515-526.

[17] Hastie,T. and Tibshirani, R. (1990). Generalized Additive Models, Chapman and Hall .

[18] Hoerl, A.E. and Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55-67.

[19] Klein J. and Moeschberger M. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*, Springer Verlag, New York.

[20] Krall, J.M., Uthoff, V.A., and Harley, J. B. (1975). a step-up procedure for selecting variables associated with survival, *Biometrics*, **31**, 49 -57.

[21] Li, C., Wei, X., and Dai, H. (2015). Adaptive elastic net method for cox model, arXiv: 1507.06371v1

[22] Liua,G. and Piantadosib, S. (2016). Ridge estimation in generalized linear models and proportional hazards regressions, in:Merck Research Laboratories,351 N Sumneytown Pike., *Statistics in Medicine*, **13**.

[23] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer and Chemotherapy Reports*, **50**, 163 - 170.

[24] Metzeler KH, Hummel M, Bloomfield CD, Spiekermann K, Braess J, Sauerland MC, Heinecke A, Radmacher M, Marcucci G, Whitman SP, Maharry K, Paschka P, Larson RA, Berdel WE, Buchner T, Wormann B, Mansmann U, Hiddemann W, Bohlander SK, Buske C (2008). An 86 probe set gene expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood*, **112**(10), 4193–4201.

[25] Oakes D. (1983). Comparison of models for survival data, *Statistics in Medicine*, **2**, 305-311.

[26] Park, M. Y. and Hastie, T, (2007). L1-regularization path algorithm for generalized linear models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 659 -677.

[27] Saseini, P. (2014), Cox Regression Model, *Encyclopedia of Biostatistics*. DOI: 10.10020470011815.b2a11013.

[28] Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics*, **6**, 461-464.

[29] Sill, M., Hielscher, T., Becker, N., Zucknick, M. (2015) Extended inference with lasso and elastic-net regularized Cox and generalized linear models, *Journal of Statistical Software*, **62**(5) 1-23.

[30] Simon, N., Friedman, J., Hastie, T., Tibshirani, R. (2011) Regularization paths for Cox's proportional hazards model via coordinate descent, *Journal of Statistical Software*, **39**(5) 1-13

[31] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, **58**(1), 267 – 288.

[32] Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model, in:Department of Preventive Medicine and Biostatistics and Department of Statistics, University of Toronto , *Statistics in Medicine*, **16**, 385 - 395.

[33] Verweij, P J. M. and Van Houwelingen, H C. (1994). Penalized likelihood in Cox regression, in: Department of‘ Medical Statistics., *Statistics in Medicine*, **13**, 2427 - 2436.

[34] Yang, Yi and Zou, H, (2013). A cocktail algorithm for sloving the elastic net penalized Cox's regression in high dimensions, *Statistics and its Inference*, **6**, 167-173.

[35] Zou, H. (2006). The Adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**(476), 1418-1429.

[36] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B*, **67**(2), 301-320.

[37] Zou, H., Zhang, H.H. (2009). On the Adaptive elastic net with a diverging number of parameters. *Annals of Statistics*, **37**, 1733-1751.

واژه‌نامه فارسی به انگلیسی

رگرسیون ..... regression
رگرسیون خطی ساده ..... Simple linear regression
رگرسیون خطی چندگانه ..... Multiple linear regression
رگرسیون خطی جریمه ..... Penalized linear regression
ضریب همبستگی پیرسن ..... Pearson’s coefficient
مجموع توان‌های دوم مانده‌ها ..... Sum of square errors
کمترین توان‌های دوم معمولی ..... Ordinary Least Squares
معیار اطلاع آکاییک ..... Akaike Information Creterion
معیار اطلاع بیزی ..... Bayesian Information Creterion
تنکی ..... Spares
لاسو ..... Least absolute shrinkage and selection operato
رگرسیون خطی ساده ..... Least Angle Regression
اسکد ..... SCAD
الاستیک‌نت ..... ElasticNet
الاستیک‌نت اولیه ..... Naive ElasticNet
اعتبار سنجی ..... Cross Validation
اعتبار سنجی تعمیم یافته ..... Generalized Cross Validation(GCV)
ماتریس تصویری ..... Projection Matrix
سانسور شده ..... Censored
دورگه ..... Hybrid
تابع چگالی احتمال ..... Probability density function
تابع بقا ..... Survival function
تابع خطر ..... Hazard function
روش کاپلان_مه‌یر ..... Kaplan-Meier method
جدول طول عمر ..... Life time table method

خطرات متناسب کاکس ..... Cox proportional hazards (cph)
اختیاری ..... Optional
درستنمایی شرطی ..... Conditional likelihood
الگوریتم نیوتون ـ رافسون ..... Newton-Raphson algorithm
نسبت خطر ..... Hazard ratio
مدل رگرسیون کاکس گروه‌بندی شده ..... Stratified Cox regression model
باقیمانده‌های شویفلد ..... Schoenfeld residuals
میکروآرایه‌ها ..... Microarrays
فنوتیپ ..... Phenotype
فنوتیپ‌های بقای احتمالی سانسور شده ..... Possibly censored survival phenotyp
رسته‌ای ..... Categorical
انتخاب پیشرو ..... Forward selection
پسرو ..... Backward
گام به گام ..... Stepwise
انتخاب بهترین زیرمجموعه ..... Best subset selection
ماتریس اطلاع ..... Information matrix
مدل انقباضی ..... shrinkage model
مدل صرفه‌جو ..... Pasimonous model
برنامه‌ریزی ..... programming
روش نیوتون رافسون به‌روز شده ..... Updated Newton-Raphson methode
الگوریتم گرادیان نزولی ..... Gradient descent algorithm
الگوریتم مختصات نزولی ..... Coordinate-majorixation descent(CMD)
مسیر ..... Pathwise

واژه‌نامه انگلیسی به فارسی

معیار اطلاع آکاییک ..... Akaike Information Creterion
معیار اطلاع بیزی ..... Bayesian Information Creterion
روش کاپلان_مه‌یر ..... Kaplan-Meier method
لاسو ..... Least absolute shrinkage and selection operato
رگرسیون کمترین زاویه ..... Least Angle Regression
الگوریتم نیوتون ـ رافسون ..... Newton-Raphson algorithm
کمترین توان‌های دوم معمولی ..... Ordinary Least Squares
ضریب همبستگی پیرسن ..... Pearson's coefficient
رگرسیون خطی جریمه ..... Penalized linear regression
رگرسیون ..... regression
رگرسیون خطی ساده ..... Simple linear regression
مجموع توان‌های دوم مانده‌ها ..... Sum of square errors
پسرو ..... Backward
انتخاب بهترین زیرمجموعه ..... Best subset selection
رسته‌ای ..... Categorical
سانسور شده ..... Censored
درستنمایی شرطی ..... Conditional likelihood
الگوریتم مختصات نزولی ..... Coordinate-majorixation descent(CMD)
خطرات متناسب کاکس ..... Cox proportional hazards (cph)
اعتبار سنجی ..... Cross Validation
الاستیک‌نت ..... ElasticNet
انتخاب پیشرو ..... Forward selection
اعتبار سنجی تعمیم یافته ..... Generalized Cross Validation(GCV)
الگوریتم گرادیان نزولی ..... Gradient descent algorithm
تابع خطر ..... Hazard function

نسبت خطر . . . . . Hazard ratio
دورگه . . . . . Hybrid
ماتریس اطلاع . . . . . Information matrix
جدول طول عمر . . . . . Life time table method
میکروآرایه‌ها . . . . . Microarrays
رگرسیون خطی چندگانه . . . . . Multiple linear regression
الاستیک‌نت اولیه . . . . . Naive ElasticNet
اختیاری . . . . . Optional
مدل صرفه‌جو . . . . . Pasimonous model
مسیر . . . . . Pathwise
فنوتیپ . . . . . Phenotype
فنوتیپ‌های بقای احتمالی سانسور شده . . . . . Possibly censored survival phenotyp
تابع چگالی احتمال . . . . . Probability density function
برنامه‌ریزی . . . . . programming
ماتریس تصویری . . . . . Projection Matrix
اسکد . . . . . SCAD
باقیمانده‌های شویفلد . . . . . Schoenfeld residuals
مدل انقباضی . . . . . shrinkage model
تنکی . . . . . Spares
گام به گام . . . . . Stepwise
مدل رگرسیون کاکس گروه‌بندی شده . . . . . Stratified Cox regression model
تابع بقا . . . . . Survival function
روش نیوتون رافسون به‌روز شده . . . . . Updated Newton-Raphson methode

Aabstract

With the advancement of technology in data storage, recently the big data are appeared in clinical trials. Determining the effective variables in predicting the time of death in a disease leads to save cost and also time, and is therefore of great importance. Nowadays in many medical studies, microscopic technology is used to find the relationship between the behavior of the genes and the various clinical features of the patients.

In this dissertation, the three famous penalty functions, ridge, LASSO and a combination of these two functions, Elastic net are used in a Cox proportional hazards regression model and compared.

Keywords:
Cox proportional hazards regression model, Elastic net estimator; Penalized Regression; Survival analysis.

Shahrood University of Technology

Faculty Of Mathematical Sciences

MSc Thesis in: Mathematical Statistics

Penalized Coxs proportional hazards regression

By: Mohammad Golpaygani

Supervisor

Mohammad Arashi

August 2017