

برآوردگرهای توانیده در مدل رگرسیون کاکس

زهرا اسلامی^۱، مینا نوروزی راد^۲، محمد آرشی^۳

تاریخ دریافت: ۹۸/۱۲/۱۵

تاریخ پذیرش: ۹۹/۱۱/۱

چکیده:

در تحلیل داده‌های بقای سانسور شده، مدل‌های رگرسیونی کاکس از اهمیت ویژه‌ای برخوردار هستند. با افزایش متغیرها در یک مدل، به‌منظور دستیابی به مدل‌های کاراتر، می‌توان از روش‌های توانیده استفاده کرد. در این مقاله، به مروری بر مدل رگرسیون کاکس توانیده برای برخی از توابع توان معروف پرداخته شده است. هم‌چنین، مجموعه داده‌های پزشکی mgus2 بررسی شده و نشان داده شده که مدل‌های توانیده بهتر از رگرسیون کاکس به این داده‌ها برازش می‌شود که توان لاسو، بهترین عملکرد را برای این داده‌ها دارد.

واژه‌های کلیدی: تابع بقا، تابع خطر، رگرسیون کاکس، رگرسیون توانیده، لاسو.

۱ مقدمه

متناسب کاکس در شناسایی عوامل مؤثر بر بیماران مبتلا به متاستاز مغزی [۴] مدل رگرسیون نیمه پارامتری کاکس در تعیین عوامل مؤثر بر مدت اقامت بیماران بستری [۱۲] مقایسه مدل رگرسیون کاکس و شبکه عصبی مصنوعی در پیش‌بینی بقای بیماران لوسمی حاد [۱۶] بحث‌هایی در این مورد داشته‌اند. علی‌رغم پیشنهادها سازنده‌شان، هنوز هم ناشناخته‌های فراوانی وجود دارد که مقالات قبلی پاسخی برای آن‌ها نیافته است و چندان آموزنده و کاربردی نیستند و این ناشناخته‌ها، می‌توانند موضوع یک مطالعه در پژوهش‌های آینده باشند. در این مقاله مدل رگرسیون تابع خطر را معرفی نموده و با استفاده از یک مثال کاربردی پارامترهای مدل کاکس را تفسیر می‌کنیم. در بخش بعدی برآوردگرهای توانیده را توضیح داده و برخی از توابع توان را ذکر کرده‌ایم. سپس در بخش ۳ به معرفی اجمالی برخی از توابع توان در قالب مدل رگرسیون کاکس پرداخته و در بخش ۴، برآوردگرهای توانیده در مدل کاکس بیان شده است. در ادامه، در بخش ۵، یک مثال کاربردی را تحلیل کرده و خروجی را تفسیر می‌کنیم. بخش ۶ به نتیجه‌گیری اختصاص پیدا کرده است.

[۷]، مدلی را ارائه نمود که از آن زمان تاکنون به‌صورت گسترده‌ای استفاده شده است. این مدل برای فرمول‌بندی متغیر پاسخ زمان بقا یا زمان تا رویداد حادثه براساس یک سری متغیرهای پیش‌گو (توضیحی) و تبیینی به کار می‌رود. مدل کاکس تعمیمی از رگرسیون [۲۱] و مشابه رگرسیون نمایی [۱۱] در حالتی خاص است. در مدل‌بندی زمان بقا در حالت تک متغیره می‌توان از روش‌های جایگزین منحنی‌های کاپلان-میر و آزمون‌های لگاریتم رتبه‌ای استفاده کرد. برای آگاهی بیشتر می‌توان به [۲۰] مراجعه کرد.

آماردانان کاربردی از مدل رگرسیون کاکس، در مطالعاتشان استفاده کردند. به‌عنوان مثال می‌توان به تحلیل و بررسی داده‌ها در پزشکی، آزمایش‌های بالینی، بیماری‌های سرطانی، بیماری‌های ویروسی و مسری اشاره کرد. در نرم‌افزار R، بسته‌های زیادی وجود دارد که مدل رگرسیون کاکس را به داده‌های متنوعی برازش می‌دهد.

امروزه، فرد به‌سختی می‌تواند یک مجله پیشرفته آماری یا پزشکی را بدون حداقل یک ارجاع به مقاله [۷] پیدا کند. در ایران نیز، مقایسه رگرسیون کاکس و مدل‌های پارامتریک در تحلیل بقای بیماران مبتلا به سرطان معده، توسط [۲۲]، تحلیل بقای بیماران سرطان کولورکتال و عوامل پیش‌آگهی دهنده با استفاده از مدل رگرسیون کاکس [۲۳] به کارگیری مدل مخاطرات

^۱ کارشناس ارشد آمار ریاضی، دانشگاه صنعتی شاهرود eslamiz0107@yahoo.com

^۲ گروه آمار، دانشکده علوم ریاضی، دانشگاه صنعتی شاهرود

^۳ گروه آمار، دانشکده علوم ریاضی، دانشگاه فردوسی مشهد

۲ مدل رگرسیون تابع خطر

در تحلیل بقا، زمان تا رویداد حادثه را مدل‌بندی می‌کنیم و هدف اصلی در این مطالعات، بررسی وابستگی زمان بقا T به متغیرهای پیشین p بعدی $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$ است. یک روش معمول برای رسیدن به این هدف، تابع خطر است که تابع شرطی T به متغیر $\mathbf{X} = x$ بستگی دارد که به صورت زیر تعریف می‌شود:

$$h(t|x) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P\{t \leq T < t + \Delta t \mid T \geq t, \mathbf{X} = x\}$$

که $h(t|x)$ تابع خطر، T زمان بقا و Δt تغییرات لحظه‌ای است. اگر t مقدار مشخصی باشد، $h(t|x) \geq 0$ است و $h(t)$ کران بالا ندارد. تابع بقا در توزیع‌های پیوسته و گسسته به صورت زیر تعریف می‌شود:

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F_T(t)$$

اگر تابع خطر مشخص باشد، می‌توانیم تابع بقا را به طور مستقیم به دست آوریم. چند نمونه از توابع خطر در ادامه بیان شده است:

(۱) اگر $h(t) = \lambda$ تابع خطر ثابت باشد، تابع توزیع تجمعی به صورت $F(t) = 1 - \exp(-\lambda t)$ بوده و T دارای توزیع نمایی است.

(۲) اگر تابع خطر $h(t) = p\lambda t^{p-1}$ باشد، تابع توزیع تجمعی دارای صورت $F(t) = 1 - \exp\{-(\lambda t)^p\}$ بوده و T دارای توزیع وایبول است.

(۳) اگر تابع خطر

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

باشد، تابع توزیعی تجمعی $F(t) = \Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)$ است و T از توزیع لگ نرمال پیروی می‌کند.

فرض کنید مجموعه داده $\{(y_1, \delta_1), (y_2, \delta_2), \dots, (y_n, \delta_n)\}$ در دسترس است، همچنین $\delta_i = I(T_i \leq C_i)$ و $y_i = \min\{T_i, C_i\}$ متغیر تصادفی زمان شکست در واحد نام شکست است. T_i متغیر تصادفی نامنفی است که زمان بقا را در واحد نام مشخص می‌کند و $I(A)$ تابع نشانگر مجموعه A است. به جای استفاده از میانگین که در مدل رگرسیونی استفاده می‌شود، در رگرسیون کاکس، تابع خطر را مدل‌بندی می‌کنیم. مدل رگرسیون کاکس دارای صورت کلی زیر است:

$$h(t, \mathbf{X}) = h_0(t)h(\mathbf{X}\beta), \quad (1.2)$$

که در آن، وابستگی به زمان را نشان می‌دهد و براساس مقدار δ یکی از مقادیر مشاهده‌شده از متغیرهای T_i یا C_i که جنس هر دو زمان است، می‌پذیرد. [۷] تابع مثبت $h(\mathbf{X}^T\beta)$ را به صورت نمایی $\exp\{\mathbf{X}^T\beta\}$ پیشنهاد کرد. در این صورت مدل رگرسیون کاکس به صورت زیر نوشته می‌شود:

$$h(t, \mathbf{X}) = h_0(t) \exp\{\mathbf{X}^T\beta\}. \quad (2.2)$$

از آنجا که تابع مخاطره مثبت است، می‌توان با لگاریتم گرفتن از رابطه (۲.۲)، آن را به صورت زیر نوشت:

$$\ln(h(t, \mathbf{X})) = \ln(h_0(t)) + \{\mathbf{X}^T\beta\}. \quad (3.2)$$

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

مدل رگرسیون کاکس، کاربردی‌ترین روش برای یافتن ارتباط متغیرهای پیش‌گو با متغیر پاسخ تابع خطر است. مدل رگرسیون کاکس را بدین دلیل که نسبت خطر در همه زمان‌ها ثابت است و به زمان بستگی ندارد، مدل مخاطرات متناسب نیز می‌نامند. مثلاً نسبت خطر برای دو بردار تصادفی دلخواه $\mathbf{X}_0 = (X_1, X_2, \dots, X_p)^T$ و $\mathbf{X}_1 = (X_1, X_2, \dots, X_p)^T$ را در نظر بگیرید که به صورت (۴.۲) است. یعنی تابعی ثابت نسبت به زمان، و این یکی از مهم‌ترین مفروضات مدل کاکس است. [۸].

$$\frac{h(t|\beta, \mathbf{X}_1)}{h(t|\beta, \mathbf{X}_0)} = \frac{h_0(t) \exp(\mathbf{X}_1^T\beta)}{h_0(t) \exp(\mathbf{X}_0^T\beta)} = \exp\{(\mathbf{X}_1 - \mathbf{X}_0)\beta\} \quad (4.2)$$

در مدل رگرسیون کاکس، ناپایداری ضرایب رگرسیونی برآورد شده می‌تواند با بیشینه کردن لگاریتم درستنمایی جزئی کاهش یابد، که در آن تابع لگاریتم درستنمایی جزئی به صورت زیر است:

$$\ell_p(\beta) = \sum_{i=1}^n \delta_j \left[\mathbf{X}_i^T\beta - \ln \left(\sum_{j \in R_i} \exp\{\mathbf{X}_j^T\beta\} \right) \right], \quad (5.2)$$

که در آن اگر $\delta_j = 1$ داده سانسور نشده و اگر $\delta_j = 0$ داده سانسور شده است. برای آگاهی بیشتر در مورد جزئیات دستیابی به رابطه (۵.۲)، [۱] را ببینید.

۳ برآورد گره‌های تاوانیده

در این بخش ابتدا به معرفی مدل رگرسیون کاکس تاوانیده می‌پردازیم و در ادامه رابطه ریاضی برخی از توابع تاوان معروف بیان شده است.

²Multiple linear regression

۱.۳ رگرسیون کاکس تاوانیده

رابطه p متغیر پیش گو مدل رگرسیون خطی و یک متغیر پاسخ در رگرسیون خطی چند گانه^۲ به صورت زیر فرمول بندی می شود:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \quad (۱.۳)$$

مدل خطی (۱.۳) را در نظر گرفته و فرض کنید ماتریس واریانس-کواریانس مؤلفه خطا برابر $\sigma^2 \mathbf{I}$ است، به عبارتی $\text{cov}(\varepsilon) = \sigma^2 \mathbf{I}$ که در آن $\sigma^2 > 0$ یک مقدار عددی نامعلوم و $\mathbf{I}$ ماتریس همانی است. برآوردگر کمترین توان های دوم^۳ (LS)، برآوردگری است که مجموع توان های دوم خطاها یعنی

$$S(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) \quad (۲.۳)$$

را کمینه می کند. در صورت نامنفرد بودن ماتریس $\mathbf{X}^\top \mathbf{X}$ (معکوس پذیر بودن)، برآوردگر LS بردار مقادیر برازش شده β به صورت زیر به دست می آید

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (۳.۳)$$

گاهی اوقات تعداد متغیرهای پیش گو در مسئله مورد بررسی زیاد است. بدیهی است مدل حاصل پیچیده بوده و تفسیر آن مشکل است. مثلاً در مدل (۱.۳) چنانچه p بزرگ باشد، ماتریس $(\mathbf{X}^\top \mathbf{X})^{-1}$ به راحتی قابل محاسبه نیست و در حالت $p > n$ وجود ندارد. در این گونه موارد استفاده از یک روش انتخاب متغیر می تواند کارساز باشد. اما از آنجایی که هدف نهایی، برآورد بردار ضرایب رگرسیونی β است، بهتر است از روش هایی استفاده کنیم که هم انتخاب متغیر انجام داده و هم β را برآورد کند. برای این منظور، می توان از رگرسیون تاوانیده^۴ استفاده کرد و به هدف مورد نظر رسید. دقت کنید چنانچه تعداد پارامترها زیاد باشد تابعی از قدر مطلق β نیز بزرگ می شود. بنابراین اگر در این گونه مسائل بزرگی این تابع را تاوان کنیم و کاهش دهیم، عملاً می توان به طور غیرمستقیم تعداد متغیرها را کاهش داد. به عنوان نتیجه از این بحث می توان به جای کمینه کردن محک LS، محک LS همراه با بزرگی این نتایج تاوان را هم زمان کمینه کرد. این عمل باعث می شود به طور هم زمان هم کاهش بعد انجام داد و هم بردار β را برآورد کنیم. لذا در حالت کلی در رگرسیون تاوانیده، هدف کمینه کردن تابع زیر است:

$$\begin{aligned} \Phi(\beta) &= S(\beta) + p_\lambda(\beta) \\ &= (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + p_\lambda(\beta) \end{aligned} \quad (۴.۳)$$

که در آن $p_\lambda(\beta)$ تابع تاوان بر حسب β و λ ثابتی است که میزان تاوان را کنترل یا تنظیم می کند. در اینجا پارامتر تنظیم کننده^۵ $\lambda > 0$ است. لازم به ذکر است λ در اینجا، موازنه ای بین خطاهای پیش گویی^۶ و پیچیدگی مدل^۷ انجام می دهد و انتخاب درست آن حائز اهمیت است. تعیین مقدار مناسب برای پارامتر تنظیم کننده در آمار نیز، یک مسئله چالش برانگیز است. برای جزئیات بیشتر درباره ارتباط بین مسائل بهینه سازی ریاضی با آمار به ویژه روش های رگرسیونی، [۳۰]، [۳۲] و [۲] را ببینید.

همان طور که در بخش (۲) توضیح داده شد، روش کاکس شبیه به رگرسیون چند گانه است با این تفاوت که متغیر پاسخ در رگرسیون کاکس، تابع مخاطره است. در مدل رگرسیون کاکس (۲.۲)، وقتی تعداد متغیرهای پیش گو در مسئله مورد بررسی زیاد است، برای پیش گویی زمان بقا در حالت $p > n$ ، مدل کاکس نمی تواند به طور مستقیم کاربردی باشد، از این رو از صورت تاوانیده مدل رگرسیون کاکس استفاده می کنیم. در این حالت با توجه به تابع لگاریتم درستنمایی جزئی در مدل رگرسیون کاکس (۵.۲)، برآوردگر درستنمایی تاوانیده از رابطه زیر حاصل می شود:

$$\hat{\beta} = \arg \max_{\beta \in \mathbb{R}^p} \{l(\beta) - p_\lambda(\beta)\} \quad (۵.۳)$$

می توان مسئله (۵.۳) را مشابه روش کمترین توان های دوم در رگرسیون چند گانه، به یک مسئله کمینه سازی تبدیل کرد. پس می توان نوشت:

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \{-l(\beta) + p_\lambda(\beta)\} \quad (۶.۳)$$

که $p_\lambda(\beta) = \lambda \sum_{i=1}^p |\beta_i|^\gamma$ است و صورت های مختلف توابع تاوان به مقدار γ بستگی دارد.

در ادامه رابطه ریاضی برخی از توابع تاوان معروف بیان شده است.

لاسو، [۲۶]

$$p_\lambda(\beta) = \lambda \sum_{i=1}^p |\beta_i|$$

ریج، [۱۵]

$$p_\lambda(\beta) = \lambda \sum_{i=1}^p \beta_i^2$$

لاسوی سازوار^۸، [۳۴]

³Least squares

⁴Penalized regression

⁵Tuning parameter

⁶Prediction errors

⁷Model complexity

⁸Adobtive Lasso

$$p_{\lambda, \gamma}(|\beta|) = \lambda \left(\gamma + \frac{\gamma}{\pi} \right) \arctan \left(\frac{|\beta|}{\gamma} \right)$$

سلو^{۱۶}، [۹]

$$p_{\lambda}(\beta) = \lambda \sum_{j=1}^p \hat{w}_j |\beta_j|$$

شبکه ارتجاعی^۹، [۳۵]

$$p_{\lambda, \tau}(\beta_j) = \frac{\lambda}{\log(\tau)} \log \left(\frac{|\beta_j|}{|\beta_j| + \tau} + 1 \right)$$

لاسوی مشتق‌پذیر، [۱۴]

$$p_{\lambda}(\beta) = \lambda_1 \sum_{i=1}^p |\beta_i| + \lambda_2 \sum_{i=1}^p \beta_i^2$$

شبکه ارتجاعی سازوار^{۱۰}، [۳۶]

$$p_{\lambda}(\beta) = \lambda \sum_{i=1}^p p(\beta_i, s)$$

MIC^{۱۷}، [۲۵]

$$p_{\lambda}(\beta) = \lambda_1^* \sum_{j=1}^p \hat{w}_j |\beta_j| + \lambda_2 \sum_{i=1}^p \beta_i^2$$

اسکد^{۱۱}، [۱۰]

$$p(\beta) = -\frac{\gamma}{n} L(\beta) + \frac{\ln(n_0)}{n} \sum_{j=1}^p p_n(\beta_j)$$

$$p'_{\lambda}(|\beta|) = \lambda \left\{ \mathbb{I}_{\{|\beta| \leq \lambda\}} + \frac{(a\lambda - |\beta|)}{(a-1)\lambda} \mathbb{I}_{\{|\beta| > \lambda\}} \right\} \quad a > 1, \beta \neq 0$$

MCP^{۱۸}، [۳۳]لاسوی ذوب‌شده^{۱۲}، [۲۸]

$$p_{\lambda}(\beta) = \begin{cases} \lambda |\beta| - \frac{\beta^2}{\gamma} & |\beta| \leq \gamma\lambda \\ \frac{1}{\gamma} \gamma\lambda^2 & |\beta| > \gamma\lambda \end{cases}$$

$$p_{\lambda}(\beta) = \lambda_1 \sum_{i=1}^p |\beta_i| + \lambda_2 \sum_{i=2}^p |\beta_i - \beta_{i-1}|$$

اسکار^{۱۳}، [۵]

۴ برآوردگرهای توانیده در مدل کاکس

در تحلیل داده‌های بقا وقتی تعداد متغیرهای پیش‌گو زیاد است، مثلاً ژن‌هایی با سطوح مختلف که با آزمایش‌های ریزآرایه^{۱۹} اندازه‌گیری شده‌اند، روش‌های کلاسیک برآورد پارامتر نمی‌توانند به‌طور مستقیم استفاده شوند، زیرا در برآورد ضرایب مدل کاکس، هیچ‌یک از ضرایب رگرسیون برآورد شده دقیقاً برابر با صفر نیست و تمام متغیرها در مدل نهایی حضور دارند. در

$$p_{\lambda}(\beta) = \lambda \sum_{j=1}^p |\beta_j| + c \sum_{j < k} \max(|\beta_j|, |\beta_k|)$$

هورسز^{۱۴}، [۱۷]

$$p_{\lambda}(\beta) = \lambda_1 \alpha \sum_{j=1}^p |\beta_j| + \lambda_2 (1 - \alpha) \sum_{j < k} |\beta_j - \beta_k|$$

آتان^{۱۵}، [۳۱]⁹Elastic net¹⁰Adaptive Elastic net¹¹Smoothly clipped absolute deviation (SCAD)¹²Fused Lasso¹³Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR)¹⁴Hexagonal Operator for Regression with Shrinkage and Equality Selection (HORSES)¹⁵Arctangent (Atan)¹⁶seamless- L_0 (SELO)¹⁷Minimum Information Criterion¹⁸Minimax Concave Penalty¹⁹Microarray

می‌آید. این برآوردگر در اولین مرحله، از برآوردگر کاکس که با روش بیشینه درستنمایی به دست آمده است، استفاده می‌کند. بنابراین برآورد مرحله اول ریح، برابر است با

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

ویژگی‌های برآوردگر ریح در صورتی که درجه همخطی شدید باشد، به صورت زیر است:

۱. اگر $\lambda = 0$ ، آنگاه $\hat{\beta}_\lambda = \hat{\beta}_1 = \hat{\beta}$ که همان برآوردگری است که از

رگرسیون کاکس حاصل می‌شود.

۲. اگر $\lambda > 0$ ، آنگاه هر دو برآوردگر $\hat{\beta}_\lambda$ و $\hat{\beta}_1$ اریب هستند.

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

۴. کاربرد مدل انقباضی ریح در برآورد ضرایبی است که متغیرها نسبت به هم همبستگی بالایی داشته باشند.

۵. وقتی تعداد پارامترها بیشتر از تعداد مشاهدات است، کاربرد دارد.

۶. عمل انتخاب متغیر توسط برآوردگر ریح انجام نمی‌شود.

۲.۴ رگرسیون کاکس با تاوان لاسو

اگرچه به کار بردن تاوان ریح در رگرسیون تاوانیده، سبب بهبود در عملکرد پیش‌گوا می‌شود، اما از طرف دیگر، با توجه به حضور همه متغیرها در آن، تفسیرپذیری به سادگی امکان‌پذیر نیست. در این راستا [۲۶]، روش کمترین قدر مطلق انقباض و انتخاب عملگر را پیشنهاد کرد که منجر به مدل‌های دقیق، پایا و تفسیرپذیری می‌شود که هم انقباض و هم انتخاب متغیر را انجام می‌دهد و آن را لاسو نامید.

[۲۷] از تاوان لاسو برای انتخاب متغیر و انقباض در مدل رگرسیون کاکس استفاده کرد. طرحی که او ارائه کرد، لگاریتم درستنمایی جزئی آزمودنی را با تابع تاوان مجموع مقادیر قدر مطلق پارامترها بیشینه می‌کند. به عبارت دیگر،

$$\hat{\beta}^{\text{LASSO}} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ l(\beta) + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (2.4)$$

[۲۷] در مطالعات شبیه‌سازی نشان داد که لاسو می‌تواند سبب کاهش مقدار ضرایب و ارائه بعضی از ضرایب دقیقاً صفر شود.

[۲۷] دو الگوریتم برای روش لاسو براساس تکنیک‌های بهینه‌سازی ارائه داد.

الگوریتم‌های تکراری و شامل راه‌حل مکرر، از مشکلات توان‌های دوم مطلق می‌باشد. استراتژی برای حل (۲.۴) این است که از روش نیوتن-رافسون به روز

شده^{۲۰} استفاده شود.

فرض کنید $\mathbf{X}$ ماتریس طرح متغیرهای پیش‌گو را مشخص می‌کند و هم‌چنین

نتیجه، قادر به انتخاب متغیرهای مهم و مدیریت حالتی که 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 بیشینه کردن لگاریتم درستنمایی جزئی برآورد کرد. اما برخی تحقیقات نشان داده‌اند که مدل کاکس به دلیل فضای بالای بعد پیش‌گوا و همخطی بالای برخی از ژن‌ها نمی‌تواند به طور 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 هزاران بیان ژن درست می‌شود. بنابراین به دلیل این که ممکن است (۱) مدل پیچیده باشد، (۲) مدل کامل و درست، ناشناخته باشد و (۳) وجود متغیرهای اضافی در مدل باعث بیش برآوردی شود، نمی‌توان از همه متغیرهای پیش‌گو در مدل استفاده کرد و از آنجایی که مدل‌ها با تعداد متغیرهای کم‌تر همواره دلخواه تحلیل‌گران آماری بوده است بنابراین مدلی انتخاب شود که هم آسان و قابل فهم و هم قابل توضیح برای دیگران باشد.

از این رو همان‌طور که در بخش ۲ توضیح داده شد، از صورت تاوانیده مدل رگرسیون کاکس، استفاده می‌کنیم. برای آگاهی بیشتر [۱] را ببینید. در ادامه به معرفی اجمالی برخی از توابع تاوان پر کاربرد می‌پردازیم.

۱.۴ رگرسیون کاکس با تاوان ریح

[۱۵] روش ریح را پیشنهاد کردند. اگر در رگرسیون کاکس، متغیرهای پیش‌گو از همبستگی بالایی برخوردار باشند، روش‌های برآورد استاندارد مانند بیشینه درستنمایی، برآوردهای ناپایدار و واریانس بزرگ تولید می‌کند، در برخی مواقع از روش ریح می‌توان به عنوان راه‌حلی در این مشکل استفاده کرد. روش‌های رگرسیون ریح در مدل‌های خطی، به منظور پایداری بیشتر پارامتر و تقویت برآورد واریانس، یک اریبی ناچیز در برآوردها می‌پذیرد. [۲۹] از درستنمایی تاوانیده در رگرسیون کاکس استفاده کردند که لگاریتم درستنمایی جزئی تاوانیده به صورت زیر است:

$$l_\lambda^{\text{Ridge}}(\beta) = l(\beta) + \lambda \sum_{j=1}^p \beta_j^2 \quad (1.4)$$

λ که مقدار انقباضی برآورد (β) را کنترل می‌کند، پارامتر ریح نامیده می‌شود. فرض کنید $U(\beta) = \frac{\partial}{\partial \beta} l(\beta)$ و $\Omega(\beta) = \mathbf{X}^T \mathbf{X}$ ماتریس اطلاع باشد، بنابراین برای تابع لگاریتم جزئی با تاوان ریح، فرض کنید

$$U_\lambda(\beta) = U(\beta) + \lambda \beta \quad \Omega_\lambda(\beta) = \mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p$$

که $\mathbf{I}_p$ یک ماتریس همانی $p \times p$ است. با روش نیوتن-رافسون، بیشینه تابع $l_\lambda(\beta)$ که از رابطه (۱.۴) تعریف شده است طبق الگوریتم تکراری به دست

²⁰Update Newton-Raphson method

می تواند انتخاب کند.

تابع تاوان شبکه ارتجاعی به صورت زیر تعریف می شود:

$$p_{\alpha, \lambda}(|\beta_j|) = \lambda \times (\alpha |\beta_j| + (1 - \alpha) \beta_j^2)$$

در اینجا $\alpha \in [0, 1]$ تعیین تأثیر تاوان لاسو نسبت به ريج است. تابع تاوان شبکه ارتجاعی دو پارامتر تنظیم کننده دارد که به داده ها وابسته هستند.

برای بیشینه کردن تابع درستمایی جزئی تاوانیده باید یک مجموعه از پارامترهای تنظیم کننده (α, λ) پیدا کنیم که با چالش مواجه هستیم. معمولاً از یک روش جستجو شبکه ای استفاده می شود که به محاسبات سیستمی در انحراف تاوان لگاریتم درستمایی در هر نقطه از شبکه نیاز دارد که به این معنی است که تراکم شبکه در دقت و پیچیدگی زمانی الگوریتم کارآمد است. [۳]

[۳۶] روش شبکه ارتجاعی سازوار را پیشنهاد دادند. [۱۹] نشان دادند که برآوردگر شبکه ارتجاعی سازوار در مدل رگرسیون کاکس خاصیت پیش گویی دارد. اضافه کردن تابع تاوان تغییر یافته شبکه ارتجاعی منجر به برآوردگر شبکه ارتجاعی سازوار برای مدل رگرسیون کاکس می شود:

$$\hat{\beta}^{AEN} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ l(\beta) + \lambda_1^* \sum_{j=1}^p \hat{\omega}_j |\beta_j| + \lambda_2^* \sum_{j=1}^p \beta_j^2 \right\} \quad (5.4)$$

که $\lambda_1^* \geq 0$ و $\lambda_2^* \geq 0$ پارامترهای تنظیم کننده هستند که مقادیر نامنفی را اختیار می کنند و اگر $\hat{\omega}_j = (\hat{\beta}^{AEN})^{-1}$ را برآوردگر شبکه ارتجاعی سازوار قرار دهیم، برآوردگرهای ريج، لاسو و یا شبکه ارتجاعی را شامل می شود. بدین منظور، (۱) اگر $\lambda_1^* = 0$ ، آنگاه رابطه (۵.۴) برآوردگر کاکس ريج را نتیجه می دهد. (۲) اگر $\lambda_2^* = 0$ و $\lambda_1^* = 1$ (به ازای هر j)، آنگاه برآوردگر لاسو در مدل رگرسیونی کاکس را نتیجه می دهد. (۳) اگر $\lambda_1^* \neq 0$ و $\lambda_2^* \neq 0$ (به ازای هر j) باشد، در این صورت برآوردگر شبکه ارتجاعی سازوار به برآوردگر شبکه ارتجاعی معمولی تبدیل می شود.

برخی از خواص برآوردگر شبکه ارتجاعی عبارت است از:

۱. برآوردگر شبکه ارتجاعی خاصیت پیش گویی ندارد.

۲. از شبکه ارتجاعی، برای دسته بندی متغیرهای مستقل به خصوص زمانی

که تعداد آن ها بزرگ است، استفاده می شود.

۳. ناحیه تاوان در شبکه ارتجاعی، ترکیب محدبی از نواحی تاوان

روش های لاسو و ريج است.

۴. روش شبکه ارتجاعی برای مدل هایی با تعداد متغیرهای متوسط یا کم،

در حالتی که همبستگی بین پیش گوها زیاد باشد و همچنین برای

حالتی که همبستگی چندان قابل توجه نباشد، عملکرد قابل قبولی از

خود نشان داده است.

بنابراین $\eta = \mathbf{X}^T \beta$ ، $U = \frac{\partial l(\beta)}{\partial \eta}$ و $A = -\frac{\partial^2 l(\beta)}{\partial \eta \partial \eta^T}$ باشد. بنابراین جمله اول سری بسط تیلور برای $l(\beta)$ به صورت زیر است:

$$(Z - \eta)^T A (Z - \eta) \quad (3.4)$$

قبل از بیان الگوریتم، بهتر است یافتن برآوردگر لاسو را به صورت یک مسئله بهینه سازی نوشت. به عبارت دیگر

$$\hat{\beta}_n^{\text{LASSO}} = \max \{l(\beta)\} \quad \text{s.t.} \quad \sum_{j=1}^p |\beta_j| \leq s \quad (4.4)$$

از این رو برای یافتن برآوردگر لاسو از الگوریتم زیر استفاده می شود:

۱. مقدار ثابت دلخواهی در نظر گرفته شود و قرار دهید $\hat{\beta} = 0$.

۲. Z, U, A, η را براساس مقدار کنونی $\hat{\beta}$ محاسبه کنید.

۳. $(Z - \eta)^T A (Z - \eta)$ را با شرط $\sum_{j=1}^p |\beta_j| \leq s$ کمینه کنید.

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

کمینه سازی در مرحله ۳ به صورت یک مسئله برنامه ریزی درجه دو انجام می شود.

خواص رگرسیون لاسو عبارت است از:

۱. رگرسیون لاسو هم زمان هم انتخاب متغیر انجام می دهد و هم ضرایب را منقبض می کند.

۲. در حالت $p > n$ ، روش رگرسیون لاسو، حداکثر n متغیر انتخاب می کند.

۳. تاوان لاسو در کاهش برخی از ضرایب به صفر برتری بیشتری به نسبت تاوان ريج دارد.

۴. لاسو برای انتخاب متغیر در مدل کاکس یک رقیب شایسته نسبت به انتخاب گام به گام است.

از روش اعتبارسنجی برای به دست آوردن پارامتر تنظیم کننده پس از پایداری خطای اعتبارسنجی استفاده می شود.

۳.۴ رگرسیون کاکس با تاوان شبکه ارتجاعی

تابع تاوان شبکه ارتجاعی یکی از توابع تاوان جدید است که ترکیبی از تاوان های لاسو و ريج است و اولین بار توسط [۳۵] معرفی شد. این تابع تاوان نیز همچون لاسو با صفر قرار دادن برخی از ضرایب، انتخاب متغیر اتوماتیک انجام می دهد، ولی شبکه ارتجاعی به انتخاب متغیرهای بیشتری در مقایسه با لاسو تمایل دارد. به همین جهت در شرایطی که متغیرها نسبت به یکدیگر خیلی همبسته باشند برآوردگر شبکه ارتجاعی کاربردی تر است زیرا برآوردگر لاسو فقط یکی از متغیرها را از بین مجموعه متغیرهای همبسته

۴.۴ رگرسیون کاکس با تاوان اسکد

[۱۰] سه شرط برای یک تابع تاوان خوب ارائه کردند که عبارت‌اند از:

۱. نااریبی^{۲۱}: برای جلوگیری از اریبی مدل، باید پارامتر برآورد شده به مقادیر بزرگ پارامتر نزدیک باشد.
۲. تنگی^{۲۲}: برآوردگر به دست آمده در آستانه واقعی است، زمانی که به طور خودکار ضرایب برآوردگر کوچک را صفر می‌کند تا پیچیدگی مدل را کاهش دهد.
۳. پیوستگی^{۲۳}: برآوردگر به دست آمده در داده‌ها پیوسته است تا از ناپایداری مدل جلوگیری کند.

آن‌ها بیان داشتند توابع تاوان محدب این سه شرط را دارا نیستند و اگر مقعر باشد این شرایط برقرار است، و برای این منظور تابع تاوان اسکد را که مقعر است معرفی کردند که بدین صورت می‌باشد:

$$p_{\lambda}(\beta) = \begin{cases} \lambda |\beta_j| & |\beta_j| \leq \lambda \\ -\frac{|\beta_j|^2 - 2\alpha\lambda|\beta_j| + \lambda^2}{2(\alpha-1)} & \lambda < |\beta_j| < \alpha\lambda \\ \frac{(\alpha+1)\lambda^2}{2} & |\beta_j| > \alpha\lambda \end{cases}$$

در عمل مقدار بهینه $\alpha = 3/7$ استفاده می‌شود. تابع تاوان اسکد به طور پیوسته روی $\mathbb{R} - \{0\}$ مشتق پذیر است ولی در صفر مشتق پذیر نمی‌باشد و دارای شروط سه گانه تابع تاوان خوب می‌باشد. برآوردگر اسکد به صورت زیر به دست می‌آید:

$$\hat{\beta}^{\text{SCAD}} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ l(\beta) + \sum_{j=1}^p p_{\lambda}(\beta_j) \right\} \quad (۶.۴)$$

برخی از خواص برآوردگر اسکد عبارت است از:

۱. پارامتر تنظیم α برای برآوردگر اسکد، سرعت تاوان را کنترل می‌کند که با چه سرعتی صفر می‌شود.
۲. اگر α مقدار پایین باشد، متغیر زیادی تولید می‌کند، اما اریبی کم‌تری را تخمین می‌زند.

۵.۴ رگرسیون کاکس با تاوان MIC

مدل MIC، برای به حداقل رساندن معیارهای اطلاعات تقریبی، و برآورد تنگی در مدل رگرسیونی کاکس توسط [۲۵]، بر اساس [۲۴] طراحی شده است. MIC انتخاب بهترین زیرمجموعه‌ها (BSS^{۲۴}) را برای سناریوهای با p

بزرگ انجام می‌دهد. مسئله اصلی در این روش، تابع نشانگر است، $l(\beta \neq 0)$ که عملکرد تاوان L_0 را درگیر می‌کند و منجر به بهینه‌سازی گسسته می‌شود. برای غلبه بر این مشکل، MIC پیشنهاد می‌کند که تابع نشانگر را با استفاده از یک تابع پیوسته یا تابع واحد هموار بریده شده تقریب زد. یک تقریب معقول، تابع تانزانت هایپربولیک است که بدین صورت می‌باشد:

$$\omega(\beta) = \tanh(a\beta^2) = \frac{\exp(a\beta^2) - \exp(-a\beta^2)}{\exp(a\beta^2) + \exp(-a\beta^2)}$$

a یک پارامتر مقیاس نامنفی است که دقت تقریب را کنترل می‌کند. تابع تاوان MIC به این صورت خواهد بود

$$p_n(\beta) = -\frac{2}{n}L(\beta) + \frac{\ln(n_0)}{n} \sum_{j=1}^p \rho_n(\beta_j)$$

که $\rho_n(\beta_j)$ تابع تاوان است که از بازپارامتری کردن $\beta_j = \gamma_j \omega(\gamma_j)$ و $\rho_n(\beta_j) = \omega(\gamma_j) = \tanh(a_n \gamma_j^2)$ تعریف می‌شود که فرض می‌کنیم $P(\lim_{n \rightarrow \infty} \frac{a_n}{n} < \infty) \rightarrow 1$ باشد.

پس برآوردگر MIC در مدل کاکس به این صورت خواهد بود:

$$\hat{\beta} = \arg \min_{\gamma \in \mathbb{R}^p} -2l(\beta) + \ln(n_0) \sum_{j=1}^p \omega_j \quad (۷.۴)$$

که $\omega_j = \tanh(n_0 \gamma_j^2)$ و $\beta = (\beta_j) = (\gamma_j \omega_j)$ می‌باشد.

برخی از خواص برآوردگر MIC عبارت است از:

۱. MIC از نظر محاسباتی از سایر رقبا کارآمدتر است زیرا به دلیل فرمول‌بندی خاص خود، بدون تنظیم است.
۲. MIC به دلیل بهینه‌سازی یک BIC تقریبی، بدون کم کردن فضای جستجو، یک روش منظم‌سازی سومند است.

۵ مثال کاربردی

در این بخش در قالب یک مثال کاربردی بر روی مجموعه داده‌های پزشکی mgus2 [۱۸]، چگونگی استفاده و تفسیر پارامترهای مدل کاکس را همراه با جزئیات شرح می‌دهیم. این داده‌ها از بسته Survival نرم‌افزار R3.5.2 قابل دسترسی هستند.

در ابتدا ناگزیر از توضیح مختصری درباره این مجموعه داده‌های پزشکی بیماران مونوکلونال گامپاتی کلینیک مایو یا به اختصار mgus2 هستیم. در این بیماری پروتئینی غیرطبیعی به نام پروتئین M در خون وجود دارد که این پروتئین در پلاسماهای مغز استخوان ساخته می‌شود. در این مجموعه داده ۱۳۸۴ مشاهده داریم که ۴۶ داده گمشده، از بررسی حذف می‌شود و از بین ۱۳۳۸ مشاهده، ۹۳۸ نفر در طول زمان مطالعه از بین رفته‌اند.

²¹Unbiasedness

²²Sparsity

²³Continuity

²⁴Best Subset Selection

age سن برحسب سال در زمان تشخیص،
sex جنسیت (۷۲۸ نفر مرد (M) و ۶۱۰ نفر زن (F))،
hgb تعداد هموگلوبین،
creat میزان کراتین در خون،
mspike اندازه تجزیه سرم مونوکلونال،

جدول ۱، آماره‌های توصیفی متغیرهای کمی مجموعه داده‌های mgus2 را نشان می‌دهد

هدف رگرسیون خطرات کاکس بررسی اثرهای متغیرهای پیش‌گو بر بقا بیماران است. برای تعیین برآورد ضرایب مدل خطرات کاکس از تابع درستنمایی جزئی استفاده می‌کنیم که در این بین هر کدام که مقدار $\exp\{\beta\}$ از یک بزرگ‌تر باشد، باعث افزایش خطر مرگ در بیمار است.

با توجه به نتایج به دست آمده از جدول ۲، با افزایش عوامل sexM، age، mspike، خطر مرگ در بیماران مونوکلونال گاموپاتی افزایش می‌یابد.

شکل ۱، نمودار بقا را برای این مجموعه نشان می‌دهد. نمودار بقا یک تابع، غیر صعودی و از چپ پیوسته است.

این مجموعه داده که به mgus2 معروف است، بزرگ‌تر از مطالعه شرایطی است که در مجموعه داده mgus یافت می‌شود. متغیر پاسخ futime، واحد زمان مرگ یا آخرین تماس در ماه است. متغیر death، دو مقدار ۰ و ۱ را پذیرفته که زنده ماندن یا فوت بیمار نشان می‌دهد. اگر ptime زمان تا پیشرفت بدخیمی سلول پلاسما (PCM) یا آخرین تماس در ماه است. pstat وقوع PCM، اگر مقدار ۱ را بپذیرد یعنی بدخیم می‌شود و اگر مقدار ۰ بپذیرد بدخیم نیست. برای ساختن متغیر پاسخ در مدل کاکس، باید به این نکته دقت داشت که اگر pstat مقدار ۱ را بپذیرد یعنی PCM به وقوع پیوندد، متغیر جدید event مقدار ۱ و در غیر این صورت، مقدار death * ۲ را می‌پذیرد. هم‌چنین اگر pstat مقدار ۱ را بپذیرد، آنگاه متغیر جدید مربوط به زمان، etime برابر با مقدار متغیر ptime و در غیر این صورت برابر با مقدار متغیر futime است. با استفاده از کد ساده زیر می‌توان این متغیرها را ساخت.

```
event = with(mgus2, ifelse(pstat==1, 1, 2*death))
event = factor(event, 0:2, c("censored", "progression", "death"))
```

```
etime = with(mgus2, ifelse(pstat==1, ptime, futime))
```

متغیرهای پیش‌گو این مجموعه داده عبارت‌اند از:

جدول ۱: آماره‌های توصیفی متغیرهای کمی داده mgus2

متغیرها	کمینه	چارک اول	میانه	میانگین	چارک سوم	بیشینه
age	۲۴,۰۰	۶۳,۲۵	۷۲,۰۰	۷۰,۶۶	۷۹,۰۰	۹۶,۰۰
hgb	۵,۷۰	۱۲,۲۰	۱۲,۵۰	۱۳,۲۲	۱۴,۷۰	۱۸,۱۹
creat	۰,۴۰۰	۰,۹۰۰	۱,۱۰۰	۱,۲۹۱	۱,۳۰۰	۲۲,۰۰۰
mspike	۰,۰۰	۰,۶۰	۱,۲۰	۱,۱۶	۱,۵۰	۳,۰۰
ptime	۱,۰۰	۳۷,۰۰	۸۰,۰۰	۹۲,۵۱	۱۳۶,۰۰	۴۲۴,۰۰
pstat	۰,۰۰۰	۰,۰۰۰	۰,۰۰۰	۰,۰۸۳	۰,۰۰۰	۱,۰۰۰
future	۱,۰۰	۳۹,۰۰	۸۳,۰۰	۹۴,۷۸	۱۳۸,۷۵	۴۲۴,۰۰
death	۰,۰۰۰	۰,۰۰۰	۱,۰۰۰	۰,۷۰۱	۱,۰۰۰	۱,۰۰۰

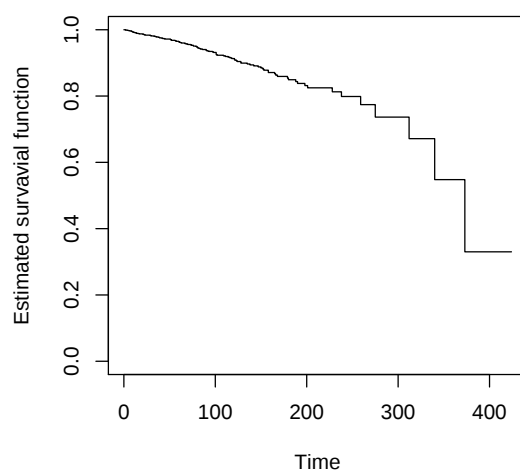
به منظور بررسی فرضیه متناسب بودن خطرات کاکس، از روش مبتنی بر آزمون [۱۳] و نمودار باقی‌مانده‌های شوینفلد^{۲۵} استفاده می‌کنیم که نتایج آزمون نیکویی برازش در جدول ۳ آمده است و نشان می‌دهد فرضیه متناسب بودن خطرات برای تمامی متغیرهای پیش‌گو در مجموعه داده mgus2 برقرار است.

برای این که بدانیم آیا مدل رگرسیون کاکس برای این نوع داده مناسب است، لازم است که ابتدا فرضیه متناسب بودن نرخ خطر بررسی شود. پس H_0 : فرضیه متناسب بودن خطرات کاکس برای همه متغیرهای پیش‌گو مجموعه داده mgus2 برقرار است و فرضیه مقابل H_1 : فرضیه متناسب بودن خطرات کاکس برای همه متغیرهای پیش‌گو مجموعه داده mgus2 برقرار نیست.

²⁵Schoenfeld residuals

جدول ۲: نتایج مدل رگرسیون خطرات کاکس برای داده‌های mgus2

متغیرها	β	$\exp\{\beta\}$	s.e (β)	z	p-مقدار
age	۰/۰۱۱	۱/۰۱۱	۰/۰۰۸۵	۱/۳۰۹	۰/۱۹۰۵
sexM	۰/۰۹۸۷	۱/۱۰۴	۰/۲۰۵۷	۰/۴۸۰	۰/۶۳۱۴
hgb	-۰/۱۳۵	۰/۸۷۴	۰/۰۵۵۶	-۲/۴۲۳	۰/۰۱۵۴
creat	-۰/۱۴۵	۰/۸۶۵	۰/۱۸۱۹	-۰/۷۹۷	۰/۴۲۵۲
mspike	۰/۹۱۲	۲/۴۹۰	۰/۱۶۵۱	۵/۵۲۶	۱۰-۸



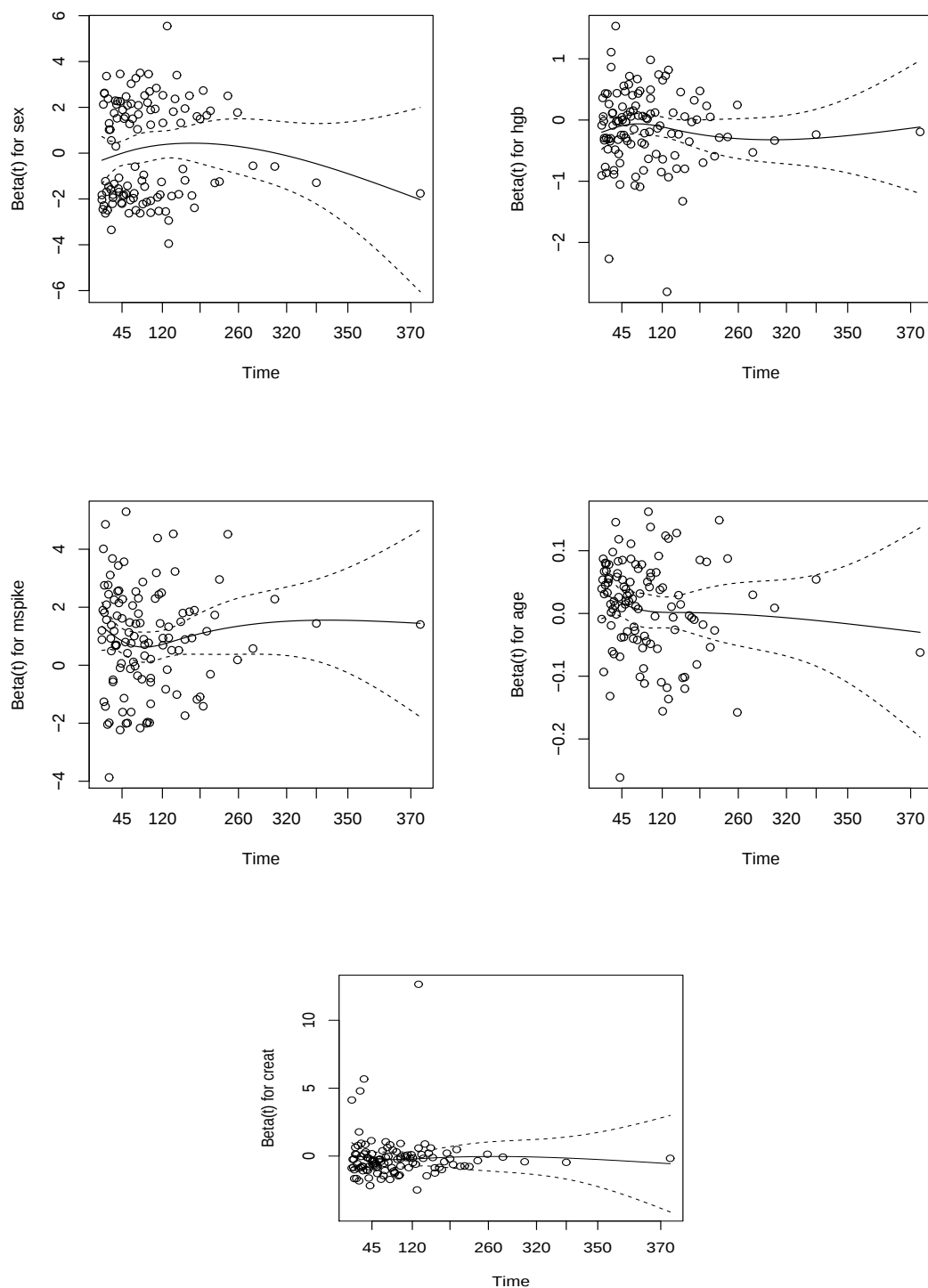
شکل ۱: نمودار بقا

جدول ۳: نتایج آزمون نیکویی برازش برای بررسی فرضیه متناسب بودن متغیرهای پیش‌گو در مجموعه داده mgus2

متغیرها	χ^2	p-مقدار
age	۱/۴۲۸۴۱	۰/۲۳
sex	۰/۰۰۷۱۱	۰/۹۳
hgb	۰/۱۹۹۹۱	۰/۶۵
creat	۰/۱۰۳۸۹	۰/۷۵
mspike	۰/۳۳۶۷۵	۰/۵۶

شونفلد به صورت تصادفی در طول زمان پراکنده باشند و باقی مانده‌ها مستقل از زمان باشند، آنگاه فرضیه متناسب بودن خطرات برقرار است. با بررسی نمودار باقی مانده‌های شونفلد برای متغیرهای این مجموعه داده، همان‌طور که در شکل ۲ مشاهده می‌کنید، فرضیه متناسب بودن خطرات برای

در نمودار باقی مانده‌های شونفلد، محور Y، طرح باقی مانده‌های شونفلد را مقیاس‌بندی می‌کند و محور X، زمان بقا را نشان می‌دهد، خط مشکی وسط، $\hat{F}$ است و خط‌های بریده بریده شده در نمودار، بازه اطمینان ۹۵٪ را برای اطراف این برازش نشان می‌دهد. اگر داده‌ها در نمودار باقی مانده‌های



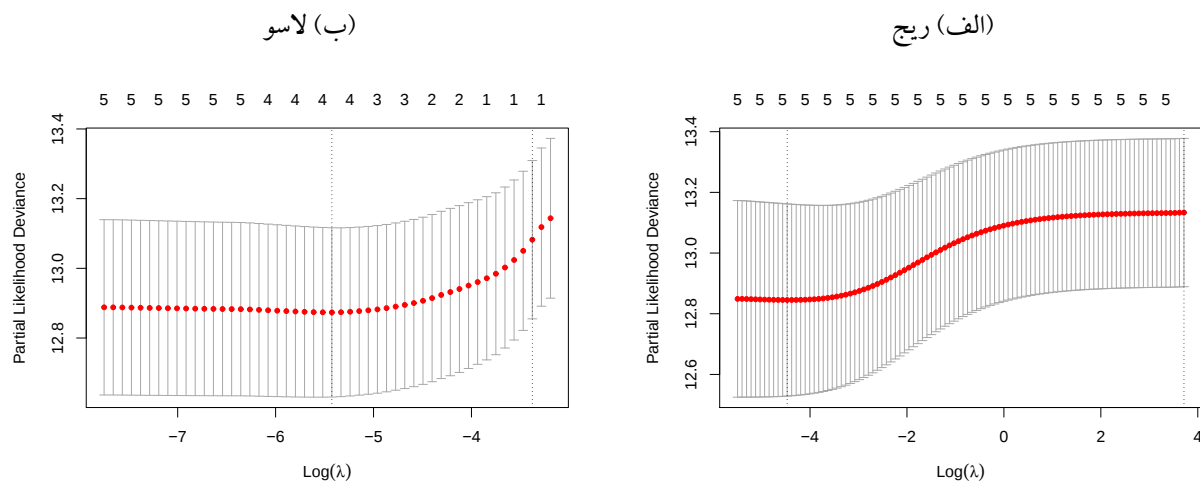
شکل ۲: نمودار باقی مانده های شوینفلد برای متغیرهای پیش گو در مجموعه داده های mgus2

براساس داده های بیماری مونوکلونال گامپاتی است.

متغیرها برقرار می باشد.

رگرسیون کاکس با تاوان ریج را بر روی مجموعه داده mgus2 اجرا

هدف این مطالعه، پیدا کردن مدل آگاهی بخش برای زمان بقای بیماران



شکل ۳: نمودار اعتبارسنجی متقابل انحراف لگاریتمی جزئی مربوط به برآورد گرهای ریج و لاسو برای مجموعه داده‌های mgus2. خطوط عمودی نقطه چین مقادیر λ با کمترین انحراف (خط چپ) و بزرگ‌ترین مقدار λ با یک انحراف استاندارد از حداقل انحراف‌ها (خط راست)

۲۷۹۵ و ۱۹۰٪ است. با توجه به ستون مربوط به اسکد در جدول ۴، این تاوان متغیر mspike را انتخاب می‌کند.

برای مقایسه مدل‌های متفاوتی که به مجموعه داده‌های mgus2 برازش داده شده است، می‌توان از کارایی نسبی دو مدل (کاکس در برابر مدل توانیده) استفاده کرد [۶]. این کارایی از تقسیم مجموع توان دوم انحراف معیار برآورد ضرایب در مدل کاکس بر مجموع توان دوم برآورد از یک مدل توانیده به دست می‌آید. به عبارت دیگر،

$$RE(\hat{\beta}^{\text{Penalized}}, \hat{\beta}^{\text{Cox}}) = \frac{\sum_{j=1}^p (s.e.(\beta_j^{\text{Cox}}))}{\sum_{j=1}^p (s.e.(\beta_j^{\text{Penalized}}))}$$

اگر $RE(\hat{\beta}^{\text{Penalized}}, \hat{\beta}^{\text{Cox}}) > 1$ آنگاه برآورد گر توانیده بهتر از برآورد گر کاکس رفتار می‌کند. جدول ۵ مقادیر $RE(\hat{\beta}^{\text{Penalized}}, \hat{\beta}^{\text{Cox}})$ را برای برآورد گرهای توانیده در جدول ۴ را بیان می‌کند. بر اساس این جدول، برآورد گر لاسو بهترین رفتار را دارد.

۶ نتیجه گیری

رگرسیون خطرات متناسب کاکس و برخی از برآورد گرهای توانیده در این مدل رگرسیونی بررسی شدند. در این مقاله، این مدل‌های رگرسیونی برای مجموعه داده‌های mgus2 بررسی شد. رگرسیون کاکس با توان‌های ریج، لاسو، شبکه ارتجاعی و اسکد را بر روی مجموعه داده‌های mgus2 اعمال کردیم و مشاهده شد رگرسیون کاکس با توان ریج، انتخاب متغیر انجام نمی‌دهد و تنها متغیرها را به سمت صفر منقبض می‌کند و باعث کاهش

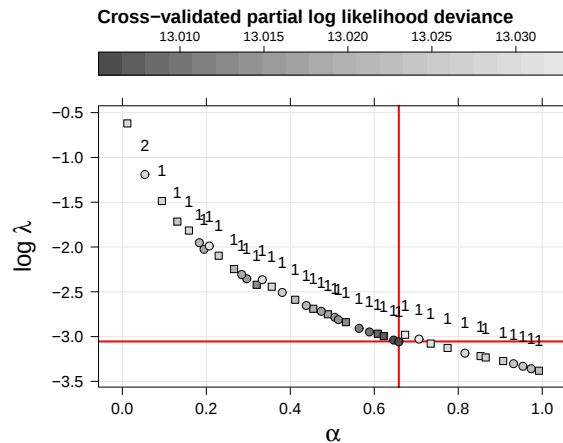
می‌کنیم که نتایج در جدول ۴ آورده شده است و خطای استاندارد ضرایب به روش خودگردان به دست آمده است. برای تعیین برآورد ضرایب مدل کاکس با توان ریج، از تابع درستنمایی جزئی استفاده می‌کنیم. با توجه به ستون مربوط به ریج در جدول ۴، هیچ انتخاب متغیری صورت نگرفته است و متغیرها، تنها به سمت صفر منقبض می‌شوند. رگرسیون کاکس با توان لاسو را بر روی این مجموعه داده اجرا می‌کنیم که نتایج در جدول ۴ آورده شده است و خطای استاندارد ضرایب به روش خودگردان به دست آمده است.

شکل ۳ نمودار اعتبارسنجی متقابل انحراف لگاریتمی جزئی، از جمله انحراف استاندارد بالا و پایین به عنوان تابعی از $\log \lambda$ در مجموعه داده‌های mgus2 نشان داده شده است که مقدار پارامتر توان ریج برای حداقل رساندن تابع زیان $\lambda = 0.1142502$ و به طور مشابه برای رگرسیون لاسو برابر 0.4402665 است.

با توجه به ستون مربوط به لاسو در جدول ۴، متغیر sex را به صفر منقبض می‌کند. رگرسیون کاکس با توان شبکه ارتجاعی را بر روی مجموعه داده mgus2 اعمال می‌کنیم. در این مدل تاوان هم‌زمان باید پارامترهای تنظیم‌کننده α و λ را مشخص کرد تا انحراف اعتبارسنجی متقابل ۱۰ تایی لگاریتمی درستنمایی توانیده کمینه شود. به این منظور از الگوریتم جستجو فاصله‌ای استفاده می‌کنیم و همان‌طور که در شکل ۴ مشاهده می‌شود، حاصل پارامتر $\alpha = 0.65844$ و $\lambda = 0.47172$ است که بر اساس جدول ۴، این برآورد گر تنها متغیر mspike را در مدل نگه می‌دارد.

رگرسیون کاکس با توان اسکد را بر روی مجموعه داده mgus2 اجرا می‌کنیم که نتایج در جدول ۴ آورده شده است و خطای استاندارد ضرایب به روش خودگردان به دست آمده است. پارامتر جریمه بهینه در این مدل برابر با

واریانس برآوردها می شود. رگرسیون کاکس با تاوان لاسو، چهار متغیر شبکه ارتجاعی و اسکد، متغیر mspike را انتخاب می کند و سایر متغیرها را انتخاب کرده است و متغیر sex را به صفر منقبض می کند. مدل با تاوان های به سمت صفر منقبض می کنند.



شکل ۴: نمودار اعتبارسنجی متقابل انحراف لگاریتم درستنمایی جزئی برای مجموعه داده های mgus2 برای (α, λ)

جدول ۴: نتایج مدل رگرسیونی کاکس با تاوان لاسو، ریج، شبکه ارتجاعی و اسکد برای مجموعه داده های استاندارد شده mgus2

متغیرها	لاسو		ریج		شبکه ارتجاعی		اسکد	
	$s.e$	β	$s.e$	β	$s.e$	β	$s.e$	β
age	۰/۰۰۰۲۹۸۰۸	۰/۰۸۳۲	۰/۰۰۰۲۳۱۰۳	۰/۱۱۹۹	۰/۰۰۰۵۵۸۶۸۶۹	۰/۰۰۰۰	۰/۰۰۰۰۸۰۳۳۹۷	۰/۰۰۰۰
sex	۰/۰۰۰۲۹۰۵۴۲	۰/۰۰۰۰	۰/۰۰۰۳۲۹۹۶۸	۰/۰۲۵۸۶	۰/۰۰۰۵۵۴۳۴۴۹۹	۰/۰۰۰۰	۰/۰۰۰۶۴۷۶۱۹۳۵	۰/۰۰۰۰
hgb	۰/۰۰۰۲۹۳۸۳۱	-۰/۱۹۰۶	۰/۰۰۰۱۶۷۵۷۲	-۰/۲۲۷۵	۰/۰۰۰۴۱۹۸۱۲۶۳	۰/۰۰۰۰	۰/۰۰۰۱۲۰۳۰۴۱۱	۰/۰۰۰۰
creat	۰/۰۰۰۸۹۴۲۶۳	-۰/۰۲۲۱	۰/۰۰۰۷۸۵۹۳۷	-۰/۱۰۱۴	۰/۰۰۰۷۵۷۶۹۰۴۸	۰/۰۰۰۰	۰/۰۰۰۸۷۵۷۵۲۷۶	۰/۰۰۰۰
mspike	۰/۰۰۰۷۳۵۸۴	۰/۴۶۹۳	۰/۰۰۰۲۸۰۸۳۳	۰/۴۵۷۷	۰/۰۰۰۰۰۰۰۰۰۰	۰/۱۱۱۱	۰/۰۰۰۰۰۰۰۰۰۰	۰/۳۵۴۳

جدول ۵: جدول مقادیر کارایی برآوردهای تاوانیده

برآورد گرهای	کاکس	لاسو	ریج	شبکه ارتجاعی	اسکد
کارایی	۱/۰۰۰	۳۸/۹۸۹۰۲	۳۸/۸۵۶۷۷	۳۴/۵۰۳۰۶	۳۷/۳۴۴۱۴

برای این که بدانیم کدام مدل تاوان عملکرد بهتری دارد، کارایی نسبی متغیرها را محاسبه کردیم. با مقایسه مدل ها درمی یابیم که برآوردگر لاسو عملکرد بهتری از سایر مدل ها دارد زیرا مقدار کارایی نسبی آن بیشتر از سایر مدل ها است و نشان می دهد تاوان لاسو برای مجموعه داده mgus2 کاراتر و قابل اعتماد است. تاوان ریج در رتبه بعدی قرار می گیرد و پس از آن، اسکد و شبکه ارتجاعی قرار می گیرند.

مراجع

- [۱] گلپایگانی، م. (۱۳۹۶). مدل رگرسیون نرخ خطر کاکس جریمه شده، پایان نامه کارشناسی ارشد؛ دانشکده علوم ریاضی، دانشگاه صنعتی شاهرود.
- [۲] نوروزی راد، م. (۱۳۹۶). برآورد گرهای بهبود یافته در برخی مدل های رگرسیونی خطی جریمه شده، رساله دکتری، دانشکده علوم ریاضی، دانشگاه صنعتی شاهرود.
- [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] Askarishahi, M., Moazen, H., Akhavan, A., Bibesh, F. and Falahzadeh, H. (2017). Applying Cox proportional hazard model to identifying the factors affecting the survival of patients with brain metastases. *The Journal of Toloo-e-behdasht*, **16(1)**, 33-46.
- [5] Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection and supervised clustering of predictors with OSCAR. *Biometrics*, **64(1)**, 115-123.
- [6] Casella, G. and Berger, R. (2001). *Statistical Inference*. Wadsworth, Belmont, CA.
- [7] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, **34(2)**, 187-202.
- [8] Cox, D. R and Oakes, D. (1984). *Analysis of survival data*, Second ed, London., Champan and Hall Press.
- [9] Dicker, L., Huang, B. and Lin, X. (2013). Variable selection and estimaton with the seamless L_0 penalty. *Statistica Sinica*, **23(2)**, 929-962.
- [10] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96(456)**, 1348-1360.
- [11] Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics* **21(4)**, 826-838.
- [12] Gohari, M., Vahabi, N. and Moghadamifard, Z. (2012). Semi-parametric Cox regression for factor affecting hospitalization length. *Daneshvar Medicine*, **19(99)**, 23-30.
- [13] Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81(3)**, 515-526
- [14] Haselimashhadi, H. and Vinciotti, V. (2016). A Differentiable alternative to the Lasso penalty. *arXiv preprint arXiv:1609.04985*.
- [15] Hoerl, A. E. and Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12(1)**, 55-67.
- [16] Hosseini Teshnizi, S., Tazhibi, M. and Tavasoli Farahi, M. (2013). Comparson of Cox regression and Artificial Neural Network models in prediction of survival in acute leukemia patients. *The Scientific Journal of Iranian Blood Transfusion Organization*, **10(2)**, 154-162.
- [17] Jang, W., Lim, J., Lazar, N. A., Loh, J. M. and Yu, D. (2013). Regression shrinkage and grouping of highly correlated predictors with HORSES. *arXiv:1302.0256*.

- [18] Kyle, R., Therneau, T., Rajkumar, V., Offord, J., Larson, D., Plevak, M. and Joseph, L. (2002). Melton III, A longterms study of prognosis in monoclonal gammopathy of undertermined significance. *New England Journal of Medicine*, **346**, 564-569.
- [19] Li, C., Wei, X. and Dai, H. (2015). Adaptive Elastic net method for cox model. *arXiv:1507.06371v1*.
- [20] Machin, D., Cheung, Y. B. and Parmar, M. K. B. (2006). Survival analysis: A Practical approach, 2nd Ed., *John Wiley*.
- [21] Mentel, T. J. (1966). Study and development of simple matrix methods for inelastic structures. *Journal of Spacecraft and Rockets*, **3(4)**, 449-457.
- [22] Pourhoseingholi, M., Hajizadeh, E., Abadi, A., Safaei, A., Moghimi Dehkordi, B. and Zali, M. (2007). Comparing Cox regression and parametric models for Survival analysis of patients with Gastric cancer. *Iranian Journal of Epidemiology*, **3(1 and 2)**, 25-29.
- [23] Roudbari, M., Abbasi Asl, M., Barfei, F., Gohari, M. R. and Khodabakhshi, R. (2015). Survival analysis of Colorectal cancer patients and its Prognostic factors using Cox regression. *Razi Journal of Medical Science*, **22(130)**, 21-28.
- [24] Su, X. (2015). Variable selection via subtle uprooting, *Journal of Computational and Graphical Statistics*, **24(4)**, 1092–1113.
- [25] Su, X., Wijayasinghe, C. s., Fan, J. and Zhang, Y. (2016). Sparse estimation of Cox proportional hazards models via approximated information criteria. *Biometrics*, **72(3)**, 751-759.
- [26] Tibshirani, R. (1996). Regression Shrinkage and selection via the Lasso. *Royal Statistical Society*, **58(1)**, 267-288.
- [27] Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model, *Statistics in medicine*, **16(4)**, 385-395.
- [28] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67(1)**, 91-108.
- [29] 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.
- [30] Wang, D. Q., Chukova, S. and Lai, C. D. (2004). On the relationship between regression analysis and mathematical programming. *Journal of Applied Mathematical and Decision Sciences*, **8**, 131-140.
- [31] Wang, Y. and Zhu, L. (2016). Variable selection and parameter estimation with the Atan regularization method. *Journal of Probability and Statistics*, **2016**, 1-12.
- [32] Yao, Y. (2008). Statistical Applications of linear programming for feature selection via regularization methods. *Ph.D. disseration, The Ohio Stat University*.
- [33] Zhang, C. H. (2010). Nearly unbiased variable selection under Minimax Concave Penalty. *The Annals of Statistics*, **38(2)**, 894-942.
- [34] Zou, H. (2006). The Adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, **101(476)**, 1418-1429.
- [35] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Royal Statistical Society: Series B*, **67(2)**, 301–320.

- [36] Zou, H. and Zhang, H. H. (2009). On The adaptive Elastic-net with a diverging number of parameters. *The Annals of Statistics*, **37(4)**, 1733–1751.