پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان

15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan

انجمن آمار ایران

بررسی رفتار مجانبی رهیافت PTE در حضور همخطی چندگانه

سید امیرحسین طباطبائی شیرازی [1]، مهدی عمادی، محمد آرشی، سولماز سیف‌اللهی

گروه آمار، دانشگاه فردوسی مشهد

چکیده: رگرسیون خطرات متناسب کاکس یکی از مدل‌های پرکاربرد در تحلیل داده‌های بقا با متغیرهای کمکی است. با این حال، در حضور هم‌خطی چندگانه، کارایی این مدل کاهش می‌یابد و برآوردهای حاصل از روش استاندارد درستنمایی جزئی ممکن است غیرقابل اعتماد شوند. در این مقاله، برآوردگر پیش‌آزمون را بر اساس برآوردگر لیو بسط می‌دهیم تا دقت برآورد ضرایب افزایش یابد. ویژگی‌های نظری حدی این برآوردگر را تعیین و عملکرد آن از طریق شبیه‌سازی‌های گسترده مونت‌کارلو ارزیابی می‌گردد. همچنین، نحوه کاربرد این استراتژی برآورد بر داده‌های بقا مورد بررسی قرار می‌گیرد. نتایج نشان‌دهنده بهبودهای چشمگیر بوده و مزایای عملی این برآوردگر را برای پژوهشگران به‌خوبی آشکار می‌سازد.

واژه‌های کلیدی: مدل خطرات متناسب کاکس، برآوردگر لیو، همخطی، برآوردگر پیش‌آزمون.

کد موضوع‌بندی ریاضی (۲۰۲۰): 62N02، 62F30.

۱ مقدمه

در عصر حاضر، تحلیل داده‌های پیچیده و حجیم، به‌ویژه در حوزه‌هایی مانند پزشکی، زیست‌فناوری و مهندسی، به یکی از چالش‌های اساسی پژوهشگران تبدیل شده است. در این میان، مدل‌های رگرسیونی ابزارهای مهمی برای تفسیر روابط میان متغیرها به‌شمار می‌روند. یکی از پرکاربردترین این مدل‌ها در تحلیل داده‌های بقا، مدل خطرات متناسب کاکس (CPHM) است که با امکان بررسی هم‌زمان تأثیر چندین عامل بر نرخ وقوع یک رویداد، ابزاری قدرتمند در مطالعات طولی به‌حساب می‌آید (کاکس ، ۱۹۷۲).

مدل کاکس با فرض ثبات ضرایب در طول زمان و بهره‌گیری از روش درست‌نمایی جزئی برای برآورد پارامترها، ساختاری انعطاف‌پذیر برای تحلیل زمان وقوع رویدادهای حیاتی فراهم می‌آورد. با این حال، در عمل با چالش‌هایی روبه‌روست؛ یکی از مهم‌ترین آن‌ها، هم‌خطی

[1] سخنران، tabatabaeishirazi.se@um.ac.ir

چندگانه یا همبستگی شدید میان متغیرهای مستقل است. این پدیده می‌تواند منجر به ناپایداری ضرایب، افزایش واریانس برآوردها و کاهش دقت و قابلیت تفسیر نتایج شود. راهکارهایی نظیر حذف یا گروه‌بندی متغیرهای همبسته، اگرچه مرسوم‌اند، اما ممکن است موجب افت کارایی مدل و بروز اریبی در نتایج گردند. از این رو، توجه به این چالش و ارائه راهکارهای مؤثر برای مقابله با آن، ضرورتی اجتناب‌ناپذیر است. در راستای مقابله با هم‌خطی در CPHM، برآوردگرهای متعددی پیشنهاد شده‌اند؛ از جمله برآوردگر ریج، برآوردگر نوع–لیو و برآوردگر کیبریا–لکمان. در این میان، برآوردگر لیو به‌سبب ساختار ساده، برخورداری از تنها یک پارامتر تنظیم و سهولت در تعیین آن، مورد توجه ویژه قرار گرفته است. مطالعات متعدد نشان داده‌اند که این برآوردگر، به‌ویژه در شرایط وجود هم‌خطی، از نظر معیار میانگین مربعات خطا، عملکرد بهتری نسبت به برآوردگر حداکثر درست‌نمایی جزئی (MPLE) دارد.

در برخی از تحلیل‌های رگرسیونی، اطلاعات پیشین درباره پارامترهای مدل ممکن است در دسترس باشد. به عنوان مثال، نتایج تحلیل‌های مقدماتی می‌توانند نشان دهند که برخی از متغیرها از لحاظ آماری معنادار نیستند و در نتیجه باید از مدل حذف شوند. این قبیل اطلاعات را می‌توان به صورت قیود خطی از نوع $\boldsymbol{C\beta} = \boldsymbol{b}$ در ساختار مدل اعمال کرد. برآوردگری که تحت این قیود به‌دست می‌آید، برآوردگر مقید نامیده می‌شود. با این وجود، در بسیاری از موارد، نسبت به صحت این قیود تردید وجود دارد. در چنین شرایطی، استفاده از برآوردگر پیش‌آزمون پیشنهاد می‌شود. این برآوردگر، ترکیبی از برآوردگر معمول و برآوردگر مقید است که تصمیم‌گیری درباره استفاده از هر یک، بر اساس نتایج آزمون فرض آماری انجام می‌گیرد. مطالعات نشان داده‌اند که در صورت اعتبار اطلاعات پیشین، برآوردگر پیش‌آزمون می‌تواند نسبت به برآوردگرهای کلاسیک عملکرد بهتری از لحاظ میانگین مربعات خطا داشته باشد.

هدف این مقاله، معرفی برآوردگر پیش‌آزمون در (CPHM) با استفاده از برآوردگر لیو و بررسی ویژگی‌های مجانبی آن است. در ادامه، با بهره‌گیری از شبیه‌سازی مونت‌کارلو، عملکرد این برآوردگر با برآوردگر لیو در شرایط مختلف، مقایسه خواهد شد و در نهایت، کارایی آن در تحلیل داده‌های مربوط به سرطان ریه مورد ارزیابی قرار خواهد گرفت.

۲ مدل خطرات متناسب کاکس

مدل CPHM که نخستین‌بار توسط کاکس (۱۹۷۲) معرفی شد، به‌عنوان یکی از ارکان اصلی در تحلیل بقا شناخته می‌شود. این مدل، به‌ویژه به دلیل توانایی‌اش در برقراری ارتباط میان زمان وقوع یک رویداد و یک یا چند متغیر کمکی، بدون نیاز به فرض شکل مشخصی برای تابع خطر پایه، از محبوبیت فراوانی برخوردار است. فرض کنید $\tau_i = \min(t_i, c_i)$ که در آن، t_i زمان بقا در صورت مشاهده رویداد و c_i زمان سانسور در صورت عدم مشاهده رویداد است. همچنین، بردار متغیرهای کمکی را به صورت $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ در نظر بگیرید. در این صورت، مدل CPHM به صورت $h(t_i \mid \mathbf{x}_i) = h_0(t_i)\exp(\mathbf{x}_i^\top \boldsymbol{\beta})$ تعریف می‌شود که در آن، $h_0(t)$ تابع خطر پایه و $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ بردار ضرایب مجهول مدل می‌باشد. برآوردگر حداکثر درست‌نمایی جزئی (MPLE) برای برآورد $\boldsymbol{\beta}$ به صورت زیر محاسبه شده است:

$$\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} = (\hat{\boldsymbol{Z}}^\top \hat{\boldsymbol{Z}})^{-1} \hat{\boldsymbol{Z}} \hat{\boldsymbol{V}}, \tag{۱.۲}$$

بطوریکه $\hat{\boldsymbol{Z}}$ ماتریسی به ابعاد $\sum(m_i + 1) \times p$ است، m_i اندازه مجموعه‌ی ریسک آزمودنی i است و $\hat{\boldsymbol{V}}$ برداری با ابعاد $p \times 1$ می باشدکه در $\hat{\boldsymbol{\beta}}_{PMPLE}$ محاسبه می‌شود تحت شرایط نظم، نشان داده شده که برآوردگرهای MPLE دارای توزیع مجانبی نرمال به صورت زیر می‌باشد: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{\beta}) \sim N_p(0, \boldsymbol{F}^{-1})$ که در آن، $\boldsymbol{F}^{-1} = \hat{\boldsymbol{Z}}^\top \hat{\boldsymbol{Z}}$ ماتریس اطلاع فیشر است (ماگنوس ، ۱۹۹۷).

فرض کنید اطلاعات پیشین درباره ضرایب مدل را بتوان در قالب q قید خطی مستقل بیان کرد. برای بررسی صحت این قیود در مدل، می‌توان آن‌ها را بر اساس فرضیه‌های آماری $H_0 : \boldsymbol{C\beta} = \boldsymbol{b}\ vs\ H_1 : \boldsymbol{C\beta} \neq \boldsymbol{b}$ آزمود. در این رابطه، $\boldsymbol{C}$ و $\boldsymbol{b}$ به ترتیب ماتریس و برداری معلوم با ابعاد $q \times p$ و $q \times 1$ می‌باشند. بر اساس توزیع مجانبی MPLE، آماره‌ای که برای آزمون فرض صفر H_0 به کار می‌رود، به صورت زیر تعریف می‌شود:

$$T_n = \left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{b}\right)^{\top}\left(\boldsymbol{C}\boldsymbol{I}^{-1}\boldsymbol{C}^{\top}\right)^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{b}\right). \tag{۲.۲}$$

با افزایش اندازه نمونه (n)، آماره آزمون T_n تحت H_0 دارای توزیع مجانبی کای‌دو با q درجه آزادی است. در صورتی که H_0 برقرار باشد، می‌توان برای برآورد ضرایب مدل از برآوردگر درست‌نمایی جزئی مقید (RMPLE) استفاده کرد که به صورت زیر تعریف می‌شود:

$$\hat{\boldsymbol{\beta}}_{\mathrm{RMPLE}} = \hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{A}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{b}\right), \tag{۳.۲}$$

که در آن $\boldsymbol{A} = \boldsymbol{F}^{-1}\boldsymbol{C}^{\top}\left(\boldsymbol{C}\boldsymbol{F}^{-1}\boldsymbol{C}^{\top}\right)^{-1}$.

۳ برآوردگرهای پیشنهادی در مدل CPHM

برآوردگر لیو یکی از روش‌های شناخته‌شده برای برآورد ضرایب در حضور همخطی در مدل‌های رگرسیونی است. در CPHM، این برآوردگر توسط احمد و همکاران (۲۰۲۳)به صورت $\hat{\boldsymbol{\beta}}_{\mathrm{CLE}} = \left(\hat{\boldsymbol{Z}}^{\top}\hat{\boldsymbol{Z}} + I_p\right)^{-1}\left(\hat{\boldsymbol{Z}}^{\top}\hat{\boldsymbol{Z}} + dI_p\right)\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}}$ تعریف شده است که در آن $d \in [0, 1]$ پارامتر لیو می‌باشد. برای سادگی می‌توان برآوردگر لیو را به فرم $\hat{\boldsymbol{\beta}}_{\mathrm{CLE}} = \boldsymbol{B}\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}}$ نوشت که در آن $\boldsymbol{B} = \left(\hat{\boldsymbol{Z}}^{\top}\hat{\boldsymbol{Z}} + I_p\right)^{-1}\left(\hat{\boldsymbol{Z}}^{\top}\hat{\boldsymbol{Z}} + dI_p\right)$ می‌باشد. برای برآورد پارامتر لیو از برآوردگری زیر در این مقاله استفاده می‌کنیم:

$$\hat{d} = 1 - \frac{\sum_{j=1}^{p} 1/\lambda_j(\lambda_j + 1)}{\sum_{j=1}^{p} \hat{\boldsymbol{\beta}}^{2}_{\mathrm{MPLE(j)}}/(\lambda_j + 1)^2}. \tag{۱.۳}$$

با الهام گرفتن از کیبریا و صالح (۲۰۰۴) می‌توان فرم برآوردگر مقید لیو (RCLE) به صورت، $\hat{\boldsymbol{\beta}}_{\mathrm{RCLE}} = \boldsymbol{B}\hat{\boldsymbol{\beta}}_{\mathrm{RMPLE}}$ تعریف کرد.

برآوردگری که هدف اصلی این مقاله می‌باشد، برآوردگر پیش آزمون (PTE) در مدل CPHM می‌باشد. این برآوردگر که از انواع برآوردگرهای انقباضی می‌باشد، بر اساس صحت و یا نادرستی فرض مطرح شده در مدل، بین دو برآوردگر مقید و نامقید تصمیم‌گیری می‌کند. با استفاده از برآوردگر MPLE و RMPLE برآوردگر پیش آزمون به فرم زیر می‌توان تعریف کرد

$$\hat{\boldsymbol{\beta}}_{\mathrm{PTE}} = \hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - (\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \hat{\boldsymbol{\beta}}_{\mathrm{RMPLE}})I_{\left(Tn<\chi^{2}_{(q,\alpha)}\right)}, \tag{۲.۳}$$

و بنا به رابطه ای که بین $\hat{\boldsymbol{\beta}}_{\mathrm{CLE}}$ و $\hat{\boldsymbol{\beta}}_{\mathrm{RCLE}}$ به ترتیب با $\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}}$ و $\hat{\boldsymbol{\beta}}_{\mathrm{RMPLE}}$ دارند، می‌توان برآوردگر پیش آزمون بر اساس برآوردگر لیو را به صورت $\hat{\boldsymbol{\beta}}_{\mathrm{CLPTE}} = \boldsymbol{B}\hat{\boldsymbol{\beta}}_{\mathrm{PTE}}$ تعریف کرد. یا به عبارتی دیگر،

$$\hat{\boldsymbol{\beta}}_{\mathrm{CLPTE}} = \hat{\boldsymbol{\beta}}_{\mathrm{RCLE}} - (\hat{\boldsymbol{\beta}}_{\mathrm{CLE}} - \hat{\boldsymbol{\beta}}_{\mathrm{RCLE}}) I_{\left(T_n < \chi^{۲}_{(q,\alpha)}\right)}, \tag{۳.۳}$$

فرم برآوردگر پیش‌آزمون لیو حکایت از این دارد که اگر $H_{\circ}$ پذیرفته شود، برآوردگر پیش‌آزمون برابر RCLE و در غیر این صورت، برابر برآوردگر CLE خواهد بود. این برآوردگر توسط ماگنوس (۱۹۹۷) معرفی شد و به شدت در مدل‌های رگرسیونی مورد توجه قرار گرفت.

۱.۳ ویژگی برآوردگرهای پیشنهادی

در این بخش به بررسی ویژگی برآوردگرهای معرفی شده به خصوص CPTE می‌پردازیم. در ابتدا آماره آزمون T_n در رابطه (۲.۲) را تحت $H_{\nu} : \boldsymbol{C\beta} = \boldsymbol{b} + \boldsymbol{\nu}, \boldsymbol{\nu} \in \mathbb{R}^{+q}$ در نظر می‌گیریم. در این حالت، زمانیکه $n \to \infty$، خواهیم داشت $T_n \to \infty$، به‌گونه‌ای که $\lim_{n\to\infty} P_{\boldsymbol{\nu}}(T_n \geq x) = ۱$. به بیان دیگر، T_n یک آزمون سازگار است. در نتیجه، تحت فرض‌های جایگزین ثابت، زمانی که $n \to \infty$ داریم

$$\begin{aligned}
&n(\hat{\boldsymbol{\beta}}_{\mathrm{PTE}} - \hat{\boldsymbol{\beta}}_{\mathrm{MPLE}}) \left(\frac{۱}{n}\boldsymbol{I}\right) (\hat{\boldsymbol{\beta}}_{\mathrm{PTE}} - \hat{\boldsymbol{\beta}}_{\mathrm{MPLE}}) \\
&= n(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{b})^{\top} \left(\frac{۱}{n}\boldsymbol{C}\boldsymbol{I}^{-۱}\boldsymbol{C}^{\top}\right)^{-۱} (\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{b}) I_{\left(T_n < \chi^{۲}_{(q,\alpha)}\right)} \\
&= T_n I_{\left(T_n < \chi^{۲}_{(q,\alpha)}\right)} \leq \chi^{۲}_{(q,\alpha)} I_{\left(T_n < \chi^{۲}_{(q,\alpha)}\right)} \to ۰.
\end{aligned}$$

به عبارتی دیگر $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{PTE}} - \boldsymbol{\beta}) \overset{D}{\equiv} \sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{\beta})$ و در نتیجه $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{CLPTE}} - \boldsymbol{\beta}) \overset{D}{\equiv} \sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}} - \boldsymbol{\beta})$. بنابراین، به‌منظور تمایز مناسب میان توزیع‌های مجانبی برآوردگرها، فرض‌های جایگزین محلی $H_{\circ(n)}$ را به صورت زیر در نظر می‌گیریم:

$$H_{\circ(n)} : \boldsymbol{C\beta} = \boldsymbol{b} + \frac{\boldsymbol{\vartheta}}{\sqrt{n}}; \qquad \boldsymbol{\vartheta} = (\vartheta_۱, \ldots, \vartheta_q)^{\top} \neq ۰ \tag{۴.۳}$$

در این صورت، وقتی $n \to \infty$، T_n دارای توزیع غیر مرکزی کای‌دو با q درجه آزادی و پارامتر غیرمرکزی $\Delta = \boldsymbol{\vartheta}^{\top}\left(\boldsymbol{C}\boldsymbol{I}^{-۱}\boldsymbol{C}^{\top}\right)^{-۱}\boldsymbol{\vartheta}$ خواهد بود. حال می‌توان تحت $H_{\circ(n)}$ و زمانی که $n \to \infty$ اثبات کرد که

$$\begin{pmatrix} \sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}} - \boldsymbol{\beta}) \\ \sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{RCLE}} - \boldsymbol{\beta}) \\ \sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}} - \hat{\boldsymbol{\beta}}_{\mathrm{RCLE}}) \end{pmatrix} \sim N_{۳p}\left(\begin{pmatrix} \boldsymbol{b}_۱ \\ \boldsymbol{b}_۲ \\ \boldsymbol{b}_۳ \end{pmatrix}, \begin{pmatrix} \boldsymbol{B}\boldsymbol{I}^{-۱}\boldsymbol{B}^{\top} & \boldsymbol{B}(\boldsymbol{I}^{-۱} - \boldsymbol{J})\boldsymbol{B}^{\top} & \boldsymbol{B}\boldsymbol{J}\boldsymbol{B}^{\top} \\ \boldsymbol{B}(\boldsymbol{I}^{-۱} - \boldsymbol{J})\boldsymbol{B}^{\top} & \boldsymbol{B}(\boldsymbol{I}^{-۱} - \boldsymbol{J})\boldsymbol{B}^{\top} & ۰ \\ \boldsymbol{B}\boldsymbol{J}\boldsymbol{B}^{\top} & ۰ & \boldsymbol{B}\boldsymbol{J}\boldsymbol{B}^{\top} \end{pmatrix} \right) \tag{۵.۳}$$

که در آن $\boldsymbol{\gamma} = \boldsymbol{B\vartheta}$، $\boldsymbol{J} = \boldsymbol{B}\boldsymbol{C}\boldsymbol{I}^{-۱}$، $\boldsymbol{b}_۲ = \boldsymbol{b}_۱ - \boldsymbol{B\gamma}$ و $\boldsymbol{b}_۳ = \boldsymbol{B\gamma}$ (کیبریا و صالح، ۲۰۰۴).

حال اگر اریبی مجانبی را به صورت زیر تعریف کنیم

$$\mathbb{B}(\hat{\boldsymbol{\beta}}) = \lim_{n\to\infty} \mathbb{E}\big(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\big), \tag{۶.۳}$$

در این صورت با استفاده از رابطه (۵.۳) ، برای برآوردگرهای بر اساس برآوردگر لیو خواهیم داشت:

$$\mathbb{B}(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}}) = \boldsymbol{b}_1, \qquad \mathbb{B}(\hat{\boldsymbol{\beta}}_{\mathrm{RCLE}}) = \boldsymbol{b}_2, \qquad \mathbb{B}(\hat{\boldsymbol{\beta}}_{\mathrm{CLPTE}}) = \boldsymbol{b}_1 - \boldsymbol{b}_3 \Upsilon_q(\Delta; \chi^2_{(q,\alpha)}).$$

که در آن $\Upsilon_q(\Delta; z)$ نمایانگر تابع توزیع تجمعی در نقطه z برای توزیع غیر مرکزی کای‌دو با q درجه آزادی و پارامتر غیر مرکزی Δ می‌باشد. همچنین، اگر واریانس برآوردگر را به فرم

$$\mathbb{V}(\hat{\boldsymbol{\beta}}) = \lim_{n\to\infty} \mathbb{E}\big(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\big).$$

تعریف کنیم، در این صورت داریم:

$$\begin{aligned}
&\mathbb{V}(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}}) = \boldsymbol{B}\boldsymbol{I}^{-1}\boldsymbol{B}^\top + \boldsymbol{b}_1\boldsymbol{b}_1^\top,\\
&\mathbb{V}(\hat{\boldsymbol{\beta}}_{\mathrm{RCLE}}) = \boldsymbol{B}(\boldsymbol{I}^{-1} - \boldsymbol{J})\boldsymbol{B}^\top + \boldsymbol{b}_2\boldsymbol{b}_2^\top,\\
&\mathbb{V}(\hat{\boldsymbol{\beta}}_{CLPTE}) = \mathbb{V}(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}}) - \Big[2\big(\boldsymbol{b}_1 - \boldsymbol{b}_3\big)\boldsymbol{b}_3^\top + \boldsymbol{B}\boldsymbol{J}\boldsymbol{B}^\top\Big]\Upsilon_{q+2}(\Delta; \chi^2_{(q,\alpha)}) - \boldsymbol{b}_3\boldsymbol{b}_3^\top\Upsilon_{q+4}(\Delta; \chi^2_{(q,\alpha)}),
\end{aligned}$$

۴ شبیه‌سازی

در این بخش، با استفاده از شبیه‌سازی مونت کارلو به ارزیابی عملکرد برآوردگرهای پیشنهادی خواهیم پرداخت. برای بررسی عملکرد برآوردگرها در حضور همخطی در مدل، متغیرهای کمکی را مطابق مدل زیر تولید می‌کنیم:

$$x_{ij} = \sqrt{1-\rho^2}Z_{ij} + \rho Z_{i(p+1)}; \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, p, \tag{۱.۴}$$

که در آن Z_{ij} نمونه‌های تصادفی از توزیع نرمال استاندارد می‌باشد. پارامتر ρ شدت همخطی بین پمتغیرهای کمکی را کنترل می‌کند و مقدار آن را $\rho = \{0.5, 0.7, 0.95\}$ در نظر گرفته می‌گیریم. همچنین $n = \{50, 100, 200\}$، $p = \{6, 12\}$ و احتمال سانسور را $cp = \{0.2, 0.3\}$ و زمان‌های بقاء پایه را توزیع نمایی با پارامتر نرخ ۰٫۵ فرض می‌کنیم. مقادیر واقعی ضرایب را به صورت $\boldsymbol{\beta} = (0.5, 1, -0.6, 0.8, 0, 0, \ldots, 0)^\top$ تعریف می‌کنیم بدین معنا که، $p-4$ ضریب آخر بردار $\boldsymbol{\beta}$ برابر صفرند و فرض صفر مدل را بر اساس این پارامترهای صفر در مدل در نظر می‌گیریم. در نهایت، داده‌ها برای CPHM مطابق روش توضیح داده شده توسط استین (۲۰۱۲) تولید می‌کنیم. علاوه بر این، برای بررسی اثر انحراف از مقدار واقعی ضرایب واقعی، بردار $\boldsymbol{b}$ در فرض صفر را به صورت $\boldsymbol{b} = (\delta, 0_{p-5}^\top)^\top$ تعریف می‌کنیم که در آن $\delta \in [0, 4]$. این رابطه نشان می‌دهد که $\delta = 0$ متناظر با مقادیر واقعی ضرایب است و مقادیر دیگر نشان‌دهنده انحراف از مقادیر واقعی می‌باشند. برای ارزیابی عملکرد برآوردگرها، تولید داده‌ها را ۱۰۰۰ بار تکرار کرده و کارایی نسبی برآوردگرهای پیشنهادی را که به صورت زیر تعریف می‌شود، محاسبه می‌کنیم:

$$\mathrm{RE}(\hat{\boldsymbol{\beta}}) = \frac{\sum_{m=1}^{1000} \left(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}}^{(m)} - \boldsymbol{\beta}\right)^{\top} \left(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}}^{(m)} - \boldsymbol{\beta}\right)}{\sum_{m=1}^{1000} \left(\hat{\boldsymbol{\beta}}_{*}^{(m)} - \boldsymbol{\beta}\right)^{\top} \left(\hat{\boldsymbol{\beta}}_{*}^{(m)} - \boldsymbol{\beta}\right)}, \tag{۲.۴}$$

که در آن $\hat{\boldsymbol{\beta}}_{*}^{(m)}$ نشان‌دهنده هر یک از برآوردگرهای پیشنهادی در تکرار mام است. از نتایج شبیه‌سازی که در جدول ۱ گزارش شده است. با توجه به نتایج، زمانی که مقدار $\delta = 0$ باشد، برآوردگرهای پیشنهادی عملکرد بهتری نسبت به سایرین دارند. اما با فاصله گرفتن δ از مقدار واقعی، کارایی آن‌ها کاهش یافته و به‌ویژه برآوردگر لیو مقید به‌شدت افت عملکرد دارد، در حالی‌که کارایی برآوردگر پیش‌آزمون به مقدار یک میل می‌کند. همچنین، با افزایش اندازه نمونه، کارایی نسبی برآوردگرها کاهش می‌یابد، اما افزایش تعداد متغیرها موجب بهبود کارایی نسبی آن‌ها می‌شود. افزایش ضریب همبستگی بین متغیرها نیز منجر به ارتقای کارایی نسبی می‌گردد. در نهایت، با افزایش احتمال سانسور، هرچند مقادیر کارایی نسبی اندکی افزایش می‌یابد، اما روند کلی نتایج تغییر چندانی نمی‌کند.

جدول ۱: کارایی برآوردگرهای پیشنهادی نسبت به برآوردگر لیو در مدل CPHM.

		$cp = 0.2$						$cp = 0.3$					
		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.95$		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.95$	
n	δ	RCLE	CLPTE	RCLE	CLPTE	RCLE	CLPTE	RCLE	CLPTE	RCLE	CLPTE	RCLE	CLPTE
$p = 6$													
50	0	1.625	1.371	1.686	1.415	1.610	1.380	1.634	1.386	1.685	1.419	1.599	1.394
	1	0.225	0.969	0.281	0.941	0.743	0.811	0.251	0.944	0.314	0.889	0.798	0.846
	2	0.064	1.000	0.082	1.000	0.309	0.872	0.072	1.000	0.094	1.000	0.350	0.835
	3	0.029	1.000	0.038	1.000	0.159	0.982	0.033	1.000	0.044	1.000	0.184	0.970
	4	0.017	1.000	0.022	1.000	0.095	1.000	0.019	1.000	0.025	1.000	0.111	0.997
100	0	1.526	1.372	1.592	1.423	1.638	1.472	1.510	1.377	1.572	1.423	1.615	1.460
	1	0.108	1.000	0.135	1.000	0.447	0.772	0.124	1.000	0.154	1.000	0.490	0.757
	2	0.029	1.000	0.036	1.000	0.147	0.993	0.033	1.000	0.042	1.000	0.167	0.987
	3	0.013	1.000	0.016	1.000	0.070	1.000	0.015	1.000	0.019	1.000	0.080	1.000
	4	0.007	1.000	0.009	1.000	0.040	1.000	0.009	1.000	0.018	1.000	0.047	1.000
200	0	1.493	1.346	1.576	1.386	1.689	1.460	1.482	1.337	1.563	1.373	1.667	1.444
	1	0.052	1.000	0.065	1.000	0.251	0.925	0.058	1.000	0.073	1.000	0.276	0.898
	2	0.013	1.000	0.017	1.000	0.072	1.000	0.015	1.000	0.019	1.000	0.081	1.000
	3	0.006	1.000	0.008	1.000	0.033	1.000	0.007	1.000	0.009	1.000	0.037	1.000
	4	0.003	1.000	0.004	1.000	0.019	1.000	0.004	1.000	0.005	1.000	0.021	1.000
$p = 12$													
50	0	4.504	2.525	4.657	2.544	3.689	2.287	4.400	2.512	4.552	2.642	3.580	2.294
	1	0.639	0.972	0.788	0.965	1.751	1.254	0.705	0.967	0.868	0.971	1.831	1.298
	2	0.185	1.000	0.236	1.000	0.754	0.936	0.208	1.000	0.267	0.999	0.829	0.937
	3	0.085	1.000	0.110	1.000	0.394	0.967	0.096	1.000	0.125	0.999	0.442	0.965
	4	0.048	1.000	0.110	1.000	0.238	1.000	0.055	1.000	0.072	1.000	0.270	0.996
100	0	3.813	2.491	4.117	2.614	4.038	2.666	3.824	2.452	4.103	2.554	3.948	2.567
	1	0.261	1.000	0.335	0.995	1.102	0.972	0.296	1.000	0.379	0.992	1.199	0.999
	2	0.069	1.000	0.090	1.000	0.363	0.976	0.079	1.000	0.103	1.000	0.410	0.958
	3	0.031	1.000	0.041	1.000	0.173	1.000	0.036	1.000	0.047	1.000	0.197	1.000
	4	0.018	1.000	0.023	1.000	0.100	1.000	0.020	1.000	0.027	1.000	0.114	1.000
200	0	3.326	2.463	3.552	2.660	3.924	2.785	3.179	2.462	3.464	2.593	3.815	2.789
	1	0.110	1.000	0.143	1.000	0.570	0.902	0.125	1.000	0.164	1.000	0.635	0.878
	2	0.028	1.000	0.037	1.000	0.164	1.000	0.032	1.000	0.043	1.000	0.109	1.000
	3	0.013	1.000	0.017	1.000	0.075	1.000	0.015	1.000	0.019	1.000	0.086	1.000
	4	0.007	1.000	0.009	1.000	0.043	1.000	0.008	1.000	0.011	1.000	0.049	1.000

۵ تحلیل داده‌های بقا

در این بخش، برای ارزیابی عملکرد برآوردگرهای پیشنهادی، داده‌های مربوط به بقای بیماران مبتلا به سرطان پیشرفته ریه از گروه درمانی سرطان North Central تحلیل خواهیم کرد. این داده‌ها، موجود در بسته آماری survival نرم‌افزار R با عنوان cancer است که شامل اطلاعاتی درباره‌ی توانایی بیماران در انجام فعالیت‌های روزمره است که با استفاده از نمرات عملکردی ارزیابی شده‌اند. پس از حذف رکوردهای دارای مقادیر گمشده، مجموعه داده نهایی شامل ۱۶۷ بیمار می‌باشد. متغیرهای موجود در این مجموعه داده شامل زمان بقا بر حسب روز، وضعیت بیمار، سن، جنسیت، نمره عملکردی کارنوفکسی بر اساس ارزیابی پزشک (ph.karno)، نمره عملکردی ECOG بر اساس نظر پزشک (ph.ecog) (۰ بدون علامت، ۱ علامت‌دار اما کاملاً قادر به حرکت، ۲ بستری کمتر از ٪۵۰از روز، ۳ بستری بیش از ٪۵۰ از روز ولی نه کاملاً بستری، ۴ کاملاً بستری) و کاهش وزن در شش ماه گذشته می‌باشد.

نتایج آزمون نسبت درست‌نمایی و آزمون والد برای برازشCPHM بر این داده‌ها، حاکی از برازش مناسب مدل با p–مقدار به ترتیب برابر با ۹×۱۰^{-۵} و $۱۰^{-۴}$ می‌باشند. بررسی همبستگی میان متغیرهای کمکی، همبستگی قوی به میزان ۰٫۸۱۱ بین متغیرهای ph.karno و ph.ecog را نشان می‌دهد. همچنین عدد شرطی محاسبه‌شده برابر با ۲۲۳٫۹۴ بوده که وجود همخطی را تایید می‌کند.

با توجه به نبود اطلاعات پیشین درباره ضرایب مدل، آزمون ضرایب انجام شد که نشان داد متغیرهای سن و کاهش وزن از نظر آماری معنادار نیستند و ضرایب آن‌ها صفر در نظر گرفته شد. برای ارزیابی عملکرد برآوردگرهای پیشنهادی مبتنی بر لیو، از روش بوت‌استرپ با اندازه نمونه ۵۰ و ۱۰۰۰ تکرار استفاده شد. نتایج جدول ۲ نشان می‌دهد که این برآوردگرها دارای انحراف معیار کمتر و کارایی نسبی بالاتر از یک هستند، که به معنای برتری آن‌ها نسبت به برآوردگر لیو است؛ یافته‌ای که با نتایج شبیه‌سازی نیز هم‌خوانی دارد.

جدول ۲: برآورد، انحراف معیار sd و کارایی نسبی برآوردگرها برای داده‌های سرطان ریه.

CLPTE		RCLE		CLE		
sd	برآورد	sd	برآورد	sd	برآورد	متغیرها
۰٫۲۷۴	−۰٫۵۱۲	۰٫۲۶۴	−۰٫۴۹۵	۰٫۲۸۴	−۰٫۵۵۱	جنسیت
۰٫۰۲۳	۰٫۰۱۰	۰٫۰۱۰	۰٫۰۲۲	۰٫۰۲۴	۰٫۰۱۱	ph.ecog
۰٫۰۱۰	۰٫۳۵۲	۰٫۰۰۱	۰٫۳۴۵	۰٫۰۱۳	۰٫۳۵۶	ph.karno
۱٫۲۱۷		۱٫۲۳۲		-		RE

بحث و نتیجه‌گیری

در این مقاله به بررسی کارایی برآوردگر که براساس برآوردگر لیو در مدل کاکس با وجود همخطی بین متغیرهای کمکی پرداخته می‌شود. ضمن معرفی برآوردگر، ویژگی‌های مجانبی آن مانند اریبی و واریانس استخراج کرده و با انجام مطالعات شبیه‌سازی گسترده، عملکرد آن‌ها را در شرایط مختلف ارزیابی نمودیم. نتایج نشان داد که این برآوردگر پیشنهادی به طور پیوسته کارایی نسبی بالاتری نسبت به برآوردگر اصلی لیو کسب کرده به‌ویژه در شرایطی که همخطی چندگانه شدید باشد، وجود دارد. در نهایت، مزایای عملی رویکرد پیشنهادی با استفاده از مجموعه داده سرطان ریه NCCTG نشان داده شد. این یافته‌ها بر ارزش استفاده از برآوردگر پیش‌آزمون مبتنی بر لیو در ارائه استنباط‌های قابل اطمینان‌تر در تحلیل بقاء تأکید دارند.

بیانیه تأمین مالی

تحقیق محمد آرشی تحت حمایت مادی بنیاد ملی علم ایران INSF برگرفته شده از طرح شماره ۴۰۱۵۳۲۰ انجام شده است.

مراجع

Ahmad S., Aslam M., Ahmad S. (2023), *Extending the Liu estimator for the Cox proportional hazards regression model with multicollinearity* Commun. Stat. Simul. Comput., **53**(12), 5828–5841.

Austin PC. (2012), *Generating survival times to simulate Cox proportional hazards models with time-varying covariates*,**31**(29),:3946–3958.

Cox DR. (1972), *Regression models and life-tables (with discussion)*, J. R. Stat. Soc., B: Methodol. ;**34**(1):187–220.

Kibria B.M.G. and Saleh A. (2004), *Performance of positive rule estimator in the ill-conditioned Gaussian regression model*, Calcutta Statistical Association Bulletin, **55**, 209–240.

Magnus J.R. (1997), *The traditional pretest estimator*, Theory of Probability and Its Applications, **44**(2), 293–308.

Invastigation in PTE Asymptotic Behavior in the Present of Multicollinearity

Seyed Amirhossein Tabatabaei Shirazi, Mahdi Emadi, Mohammad Arashi, Solmaz Seifollahi

Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran

---

Abstract: The Cox proportional hazards regression model is widely used for analyzing survival data with covariates. However, its performance can deteriorate in the presence of multicollinearity, leading to unreliable estimates when using the standard partial likelihood approach. In this paper, we extend the pretest estimator based on the Liu method to improve the accuracy of coefficient estimation. We derive the asymptotic theoretical properties of the proposed estimator and evaluate its performance through extensive Monte Carlo simulations. We also demonstrate how this estimation strategy can be applied to real survival data. The results show significant improvements, highlighting the practical advantages of this estimator for researchers working in survival analysis.

Keywords: Cox proportional hazards model, Liu estimator, Multicollinearity, Pretest estimator.

Mathematics Subject Classification (2020): 62N02, 62F30.
