تحلیل داده‌های بقای سانسوریده با استفاده از روش‌های کاهش بعد بسنده: مطالعهٔ قند و لیپید تهران

اعظم راستین،[۱] محمدرضا فریدروحانی،[۲] امیرعباس مومنان،[۳] فاطمه اسکندری،[۴] داود خلیلی[۵]

تاریخ دریافت: ۱۳۹۶/۱۲/۱۸
تاریخ پذیرش: ۱۳۹۷/۹/۱

چکیده:
بیماری‌های قلبی-عروقی شایع‌ترین علت مرگ و میر در سراسر جهان است. از سوی دیگر برای تعیین یک مدل بقای مناسب به‌منظور پیشگویی خطر بروز بیماری‌های قلبی و شناسایی عوامل خطرساز مهم در بروز این بیماری‌ها باید شکل تابعی که زمان بقا و عوامل خطرساز را به هم مرتبط سازد را مشخص کرد. در این مطالعه یک روش کاهش بعد بسنده با استفاده از یک مدل کلّی که مدل‌های بقای متداول را به‌عنوان موارد خاص شامل می‌شود، به‌منظور پیشگویی خطر بروز بیماری‌های قلبی پیشنهاد شده است.
روش‌های کاهش بعد بسنده بر مبنای رگرسیون وارون که با مدل خطرهای متناسب کاکس ترکیب شده، در مجموع یک عملکرد پیشگویانهٔ خوبی برای بقای آینده افراد دارد.

واژه‌های کلیدی: بیماری‌های قلبی-عروقی، مدل‌های پیشگو، کاهش بعد بسنده، رگرسیون وارون.

۱ مقدمه

با وجود پیشرفت‌های قابل توجهی که در کاهش میزان مرگ و میر ناشی از بیماری‌های عروق کرونر قلب صورت گرفته است، هنوز این بیماری‌ها مهم‌ترین عامل مرگ و میر در کشورهای پیشرفته و در حال توسعه به شمار می‌آیند [۴]، در کشورهای در حال توسعه و بخصوص در کشور ما به دلیل پایین بودن فرهنگ عمومی در زمینهٔ آشنایی با عوامل خطرزای بیماری‌های قلبی-عروقی و عدم رعایت اصول پیشگیری از این بیماری‌ها، میزان مرگ و میر ناشی از آن در حال افزایش است [۲].
در مطالعهٔ داده‌های بقا اغلب با پدیدهٔ سانسور مواجه هستیم. هنگام بروز این پدیده، ناکامل بودن داده‌های مشاهده‌شده ممکن است منجر به اریبی قابل توجه در نمونه شود. از طرف دیگر انتخاب یک مدل مناسب قبل از تحلیل داده‌ها نیازمند اطلاعاتی است که اغلب ناکافی بوده و هنگامی‌که تعداد پیشگوها زیاد است، تعیین و تشخیص آن دشوارتر می‌شود. در صورتی که به توان تعداد پیشگوها را به‌گونه‌ای مناسب کاهش داد، تحلیل‌های مفیدتری از مدل‌بندی نهایی می‌توان ارائه کرد [۷]؛ یک شیوهٔ مناسب برای کاهش تعداد متغیرهای کمکی در مدل‌های رگرسیونی روش عمومی موسوم به کاهش بعد بسنده است. کاهش بعد بسنده، به هیچ مدل از پیش‌تعیین‌شده‌ای نیاز ندارد [۱۰]؛ این روش اطلاعات کامل رگرسیونی را حفظ کرده و مجموعهٔ کوچکی از ترکیب متغیرها را ارائه می‌کند که فرمول‌بندی مدل و پیش‌بینی بر اساس

[۱] دانش آموخته کارشناسی ارشد آمار، دانشگاه شهید بهشتی تهران، ایران
[۲] هیأت علمی گروه آمار، دانشگاه شهید بهشتی تهران، ایران
[۳] پزشک پژوهشگر ارشد، مرکز تحقیقات پیشگیری از بیماری‌های متابولیک، پژوهشکده علوم غدد درون ریز و متابولیسم دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران
[۴] پزشک پژوهشگر، مرکز تحقیقات پیشگیری از بیماری‌های متابولیک، پژوهشکده علوم غدد درون ریز و متابولیسم دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران
[۵] متخصص اپیدمیولوژی مرکز تحقیقات پیشگیری از بیماری‌های متابولیک، پژوهشکده علوم غدد درون ریز و متابولیسم دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران

این مجموعه انجام می‌گیرد. از این رو کاهش بعد بسنده اغلب یک مسیر مطلوب برای تحلیل داده‌هایی با پیشگوهای با بعد بالا ارائه می‌کند [۶، ۱۰]. یکی از پرکاربردترین روش‌های کاهش بعد بسنده روش رگرسیون وارون ورقه[۶] (SIR) است که اوّلین بار توسط لی در سال ۱۹۹۱ معرفی شد [۸]. ایدهٔ این روش، وارون کردن رابطهٔ بین متغیر پاسخ و متغیرهای کمکی یا تبیینی است. مزیت این تغییر نقش در این است که مسئلهٔ رگرسیون چندبعدی را به حل مسئلهٔ رگرسیون وارون یک‌بعدی تبدیل می‌کند تا نتیجهٔ بهتری حاصل شود. روش SIR برای داده‌های سانسورشده بسط داده شده است [۱۰].

در سال‌های اخیر نیز بسط روش‌های کاهش بعد بسنده برای داده‌های سانسورشده مطرح شده است که به‌عنوان نمونه، کستلی [۸] همزمان با ورقه کردن زمان بقا، ضمن اختصاص وزن‌های مساوی به مشاهدات سانسورشده، یک ماتریس وزن را تشکیل داد. همچنین کوک [۱۰] با یک روش تبدیل، زمان بقا و زمان سانسور را به یک تک‌متغیره تبدیل کرد. برای به‌کارگیری روش SIR در تحلیل بقا، چیرسوهو و دیگران [۹] روش رگرسیون وارون‌ورقه شدهٔ ترکیب‌شده را مطرح کرد.

در این مقاله نشان داده می‌شود که داده‌های رگرسیونی سانسورشده[۷] (CHD) نیز می‌توانند بدون آن‌که نیازی به شکل تابعی از پیش مشخص‌شده‌ای داشته باشند، تحلیل شوند.

۲ روش رگرسیون وارون ورقه ورقه

کاهش بعد، بدون از دست دادن اطلاعات یک موضوع کلیدی در رگرسیون است. اگر بتوان بعد بردار پیشگوی $\mathbf{X}$ را بدون از دست دادن اطلاعات مرتبط با رگرسیون کاهش داد، به پیروی از اصطلاحات آمار کلاسیک، آن را کاهش بعد بسنده می‌نامند. نمونهٔ تصادفی $(Y_n, X_n), \ldots, (Y_1, X_1)$ را در نظر بگیرید که در آن Y_i پاسخ یک‌متغیره و X_i بردار p-بعدی از متغیرهای تبیینی است. فرض کنید Y بردار پاسخ‌های تک‌متغیرهٔ نمونه و $\mathbf{X}$ ماتریس طرح $p \times n$ باشد، روش رگرسیون وارون ورقه ورقه، بعد پیشگوی p-بعدی $\mathbf{X}$ را به‌وسیلهٔ شناسایی بردار سوهای تصویر p-بعدی $(k < p)$ $\beta_k, \ldots, \beta_1$ در مدل زیر، کاهش می‌دهد:

$$Y = g(\beta_1'\mathbf{X}, \beta_2'\mathbf{X}, \ldots, \beta_k'\mathbf{X}, \epsilon) \qquad (1)$$

که در آن g یک تابع نامعلوم و ϵ خطای تصادفی نامعلوم و مستقل از $\mathbf{X}$ است. زیرفضای تنیده‌شده توسط β_jها را برای رگرسیون Y روی $\mathbf{X}$ به شرط (۱) یک زیرفضای کاهش بعد و هر یک از ترکیبات خطی β_jها یک سوی کاهش بعد مؤثر[۸] نامیده می‌شوند [۸]. روش رگرسیون وارون مسیر منحنی میانگین وارون $E(\mathbf{X}\,|Y = y)$ را بررسی می‌کند که به آن منحنی وارون در R^p گویند.

الگوریتم SIR

۱. ابتدا $\mathbf{X}$ را با یک تبدیل $Z_i = \hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-\frac{1}{2}}(X_i - \bar{X})$ استاندارد کنید، که در آن $\hat{\mathbf{\Sigma}}_{\mathbf{X}}$ و $\bar{X}$ به‌ترتیب ماتریس کوواریانس نمونه‌ای و میانگین نمونهٔ $\mathbf{X}$ هستند.

۲. برد Y را به H ورقهٔ نامتداخل، $I_H, \ldots, I_1$ افراز کنید. اگر n_h تعداد مشاهدات درون ورقه h-ام $h = 1, \ldots, H$، باشد در این صورت داریم:

$$n_h = \sum_{i=1}^{n} I_h(y_i)$$

که I_h تابع نشانگر ورقهٔ h-ام است.

۳. درون هر ورقهٔ h، میانگین نمونه‌ای Z_iها که با $\bar{Z}_h$ نشان داده می‌شود، به‌صورت زیر محاسبه کنید:

$$\bar{Z}_h = \frac{1}{n_h}\sum_{i=1}^{n} Z_i I_h(y_i).$$

۴. تحلیل مؤلفهٔ اصلی موزون را برای داده‌های $\bar{Z}_h$ به روش زیر به کار گیرید:

- ماتریس کوواریانس موزون $\hat{\mathbf{\Sigma}}_\eta = \sum_{h=1}^{H} \frac{n_h}{n}\bar{Z}_h\bar{Z}_h^T$ را تشکیل دهید.
- ویژه‌مقدارها و ویژه‌بردارهای $\hat{\mathbf{\Sigma}}_\eta$ را بیابید.

[۶] Sliced inverse regression

[۷] Coronary heart disease

[۸] Effective dimension reduction(E. D. R)

۵. اگر $\hat{\gamma}_1, \ldots, \hat{\gamma}_k$، ویژه‌بردارهای متناظر با بزرگ‌ترین ویژه‌مقدارهای $\hat{\boldsymbol{\Sigma}}_\eta$ باشند آن‌گاه ویژه‌بردارهای حاصل را به مقیاس $\mathbf{X}$ تبدیل کنید تا برآورد سوهای تصویر SIR به‌صورت زیر به دست آیند:

$$\hat{\beta}_j = \boldsymbol{\Sigma}_{\mathbf{X}}^{-\frac{1}{2}} \hat{\gamma}_j, \quad j = 1, \ldots, k$$

تا کنون فرض شده که بعد ساختاری k معلوم است. اما در عمل k نامعلوم است و باید آن را برآورد کرد. استنباط روی k اغلب بر مبنای آزمون فرضیهٔ آماری دنباله‌ای به این صورت است که دنبالهٔ فرض‌های $k = m$ در مقابل $k > m$ با آغاز از $m = 0$ آزمون می‌شوند. اگر در هر مرحله فرض صفر رد شود m را یک واحد افزایش داده و آزمون فرضیهٔ آماری جدید انجام می‌شود. این دنبالهٔ آزمون‌ها در صورت لزوم تا $m \leq p-1$ ادامه می‌یابند. لی [۸] آمارهٔ زیر را برای آزمون فرضیهٔ آماری دنباله‌ای فوق پیشنهاد داد:

$$\hat{\Lambda}_m = n \sum_{j=m+1}^{p} \hat{\lambda}_j$$

که در آن $\hat{\lambda}_j$ها ویژه‌مقدارهای غیر صفر $var(E(\mathbf{Z}|Y))$ هستند. او ثابت کرد هنگامی‌که Z نرمال باشد، این آماره دارای توزیع خی‌دو با $(p-m)(h-m-1)$ درجهٔ آزادی است.

توجه شود که روش SIR هیچ فرض اضافی روی توزیع $Y|X$ اعمال نمی‌کند از این رو این روش یک انعطاف‌پذیری کامل در ساختاربندی مدل‌های بعدی ایجاد می‌کند. روش SIR به شرط خطی، یک شرط روی توزیع حاشیه‌ای X که به‌صورت زیر است، نیاز دارد:

$$E[X|\beta' X = u] = A_0 + A_1 u$$

که در آن $A_0 \in R^p$ و A_1 یک ماتریس $p \times k$ بعدی است. هنگامی‌که X از توزیع نرمال پیروی می‌کند، شرط خطی برقرار است و چون تصویرهای با بعد پایین حاصل از داده‌های با بعد بالا اغلب از توزیع نرمال پیروی می‌کنند، بنا بر این شرط خطی در روش SIR محدودیت شدیدی محسوب نمی‌شود [۷].

خوانندگان علاقه‌مند، برای اطلاعات بیشتر در مورد روش SIR می‌توانند به [۲۸] مراجعه کنند.

۱.۲ اصلاح روش SIR برای داده‌های بقای سانسوریده

در مطالعات بقا و بسیاری از تحقیقات پزشکی، متغیر پاسخ، زمان بقای یک بیمار، T، در نظر گرفته می‌شود که معمولاً تحت تأثیر برداری متشکل از متغیرهای کمکی، $\mathbf{X}$، هستند که اغلب فاکتورهای تشخیص بیماری نامیده می‌شوند. اما در عمل همهٔ زمان‌های بقای $T_n, \ldots, T_1$ نمونهٔ تصادفی n تایی مشاهده نمی‌شوند و بعضی از زمان‌های بقا به علت اتمام دورهٔ پیگیری سانسور (راست) شده و بعضی دیگر مفقود می‌شوند.

فرض کنید T را زمان بقای واقعی (غیرقابل مشاهده)، C را زمان سانسور، δ را نشانگر سانسور که $\delta = I(T \leq C)$ است و Y زمان بقای مشاهده‌شده و به‌صورت $Y = min(T, C)$ در نظر گرفته می‌شود. متغیرهای تصادفی پیوسته T و C قابل مشاهده نیستند. به علت سانسور راست، فقط Y_i و δ_i قابل مشاهده هستند. بنا بر این نمونهٔ مشاهده‌شده شامل سه‌تایی‌های مستقل و هم‌توزیع (Y_i, X_i, δ_i)، $i = 1, \ldots, n$، از توزیع $(Y, \mathbf{X}, \delta)$ خواهند بود.

چگونگی تأثیر سانسور بر SIR به رابطهٔ بین زمان سانسور C و متغیر تبیینی $\mathbf{X}$ بستگی دارد. لای [۸] دو مکانیزم سانسور زیر را مورد بررسی قرار داد:

۱.

$$T \perp C | \mathbf{X} \qquad (۲)$$

۲.

$$(T, \mathbf{X}) \perp C \qquad (۳)$$

شرط (۲)، فرض استقلال متداول برای شناسایی‌پذیری تحت سانسور تصادفی است. اگر این شرط برقرار نباشد آن‌گاه اطلاعات بیشتری در بارهٔ مکانیزم سانسور لازم است تا یک مدل مناسب ساخته شود. این مسئله در این‌جا در نظر گرفته نمی‌شود. تحت شرط (۳)، نظریهٔ عمومی SIR را بدون تغییر می‌توان به کار برد و سوهای برآوردشده همچنان سازگار باقی می‌مانند. بنا بر این با فرض استقلال، سانسور در برآوردگرهای SIR اریبی ایجاد نمی‌کند. اما SIR تحت مکانیزم‌های دیگر سانسور که شرط (۳) برای آنها برقرار نیست، باید اصلاح گردد.

در واقع هنگامی‌که پاسخ‌ها سانسورشده باشند، روش رگرسیون وارون ورقه ورقه را نمی‌توان به کار برد و یا برای استفاده از آنها شرایط محدودکننده‌ای لازم است. لی و دیگران [۱۱] پیشنهاد دادند که Y را می‌توان به‌ترتیب درون هر زیرنمونه $\delta = ۰$ و $\delta = ۱$ افراز کرد، آن‌گاه سایر مراحل مانند روش SIR معمولی باقی می‌مانند.

پس از کاهش بعد، بسیاری از روش‌های هموارسازی و نمایش‌های گرافیکی با موفقیت بیشتری انجام خواهند گرفت [۵]. مدل‌های پیشگو نیز می‌توانند بر اساس داده‌های کاهش‌یافته ساخته شوند. برای مثال مدل‌های بقا مانند مدل مخاطره متناسب کاکس را می‌توان با سوهای تصویرشده به‌عنوان متغیرهای جدید تبیینی برازش داد و از این مدل برازش داده شده برای پیش‌بینی احتمال بقای افراد استفاده کرد. همچنین با استفاده از ویژه مقدارها و نمودارهای با بعد پایین می‌توان مناسب بودن مدل بقای متداول را بررسی کرد. این نمودارها اطلاعات با ارزشی را در بارهٔ الگوی کلّی سانسور، وجود داده‌های پرت یا دور افتاده ممکن و شکل سطح رگرسیون ارائه می‌کند [۹].

۳ تحلیل داده‌های CHD

مطالعهٔ قند و لیپید تهران مطالعه‌ای است که به‌طور جامع عوامل خطرساز آترواسکلروزرا (با توجه ویژه به اختلالات متابولیسم قند و لیپید) در جمعیت شهری تهرانی شناسایی نموده، هدف از آن ارائهٔ راه حل‌هایی برای تغییر در شیوهٔ زندگی این جمعیت و پیشگیری از بروز دیابت شیرین و دیسلیپیدمی است. این مطالعه از سال ۱۳۷۸ شروع شده است و در آن حدود ۱۵۰۰۰ نفر از مردم شرق تهران مورد بررسی قرار گرفته‌اند. داده‌ها با استفاده از پرسشنامه، آزمایش و معاینهٔ بالینی گردآوری شدند. جزئیات کامل روش تحقیق این طرح در [۲] است.

تحلیل حاضر، به‌منظور بررسی عوامل مؤثر در بروز بیماری‌های عروق کرونر قلب با استفاده از روش‌های کاهش بعد بسنده انجام گرفته است. در این پژوهش همهٔ افراد بالای ۳۰ سال مطالعهٔ قند و لیپید تهران که مبتلا به دیابت نبوده و سابقهٔ بیماری‌های قلبی نیز نداشته‌اند، در نظر گرفته شده‌اند. متغیر زمان بقا، زمان از ورود به مطالعه تا رخداد اوّلین CHD یا سانسور شدن بر حسب روز را نشان می‌دهد. از مجموع ۶۱۵۸ نفر تحت پوشش مطالعهٔ قند و لیپید تهران، ۵۴۱۷ نفر دارای اطلاعات کامل بودند که در مطالعهٔ حاضر صرفاً این افراد مورد توجه قرار گرفته‌اند. از این تعداد ۲۳۸۳ نفر (۴۴ درصد) را مردان و ۳۰۳۴ نفر (۵۶ درصد) را زنان تشکیل داده‌اند. از ۵۴۱۷ نفر مورد مطالعه ۳۳۲ نفر در طول این مدت دچار بیماری‌های قلبی شدند که ۲۱۶ نفر آنها مرد و ۱۱۶ نفر زن بودند، البته ۵۰۸۵ نفر همچنان زنده مانده‌اند یا اطلاعات دقیقی از وضعیت بقای آنها موجود نیست و به‌عنوان مشاهدات سانسورشدهٔ راست در نظر گرفته شدند که این امر باعث وجود تعداد زیادی دادهٔ سانسورشده (بیش از ۹۰ درصد) در مشاهدات شد. با استفاده از برآورد کاپلان-مه‌یر میانگین بقای در طی میانهٔ ۱۰ سال پیگیری کلّ افراد ۷۵/۳۹۵۰ روز با انحراف معیار ۷۷/۶ به دست آمد که این میانگین در زنان ۳۹۹۶ و در مردان ۳۸۹۷ روز بود. $p-$ مقدار حاصل از آزمون لگ‌رتبه‌ای برای بررسی وجود تفاوت بین بقای دو گروه نشان می‌دهد که اختلاف معنی داری بین بقای مردان و زنان وجود دارد ($p = ۰/۰۰۱$). به‌عبارت دقیق‌تر بقای زنان نسبت به بقای مردان بیشتر بوده و مردان بیشتر از زنان تحت تأثیر عوامل مخاطره بیماری‌های قلبی هستند.

به دنبال مراحل شرح داده شده در قبل، روش SIR سانسورشده را روی داده‌های CHD و گروه مردان اجرا می‌شود. با معنی‌دار شدن سه ویژه-مقدار اوّل از دنبالهٔ ویژه-مقدارهای خروجی حاصل از این روش، تصویر داده‌ها در سه سوی تصویرهای $SIR_۱ = \beta_۱'X$، $SIR_۲ = \beta_۲'X$ و $SIR_۳ = \beta_۳'X$ در نظر گرفته می‌شود.

برآوردهای پایهٔ $\beta_۱$، $\beta_۲$ و $\beta_۳$ در جدول ۱ آمده است. همان‌طور که از جدول مشاهده می‌شود متغیرهای مصرف داروی کاهندهٔ چربی خون و مصرف داروی کاهندهٔ فشار خون، استعمال دخانیات، نمایهٔ تودهٔ بدنی در هر سه سو در مقایسه با سایر متغیرها ضرایب بالاتری دارند. در حالی‌که متغیرهای سن، فشارخون سیستولیک، اندازهٔ دور کمر، HDL سرم خون، در دو سو و متغیرهای فشار خون دیاستولیک، قند خون ناشتا تنها در یک سو ضرایب بالایی را منعکس می‌کنند. در جدول ۱ برآوردهای

حاصل از برازش مدل مخاطره متناسب کاکس نیز آمده است. تحلیل مشابهی را برای زنان انجام داده و با توجه به آزمون بعد این بار تصویر داده‌ها در دو سوی تصویر $SIR_1 = \beta_1' X$ و $SIR_2 = \beta_2' X$ در نظر گرفته می‌شود. برآوردهای پایهٔ β_1 و β_2 در جدول ۲ آمده است. از جدول ۲، متغیرهای مصرف داروی کاهندهٔ چربی خون و مصرف داروی کاهندهٔ فشار خون، اندازهٔ دور کمر، نمایهٔ تودهٔ بدنی، فشار خون دیاستولیک، سن در هر دو سوی تصویر ضرایب بالاتری دارند، در حالی که متغیرهای فشارخون سیستولیک، HDL سرم خون فقط در یک سوی تصویر ضرایب بالایی را منعکس می‌کنند که این موضوع با برآوردهای حاصل از رگرسیون کاکس سازگار است.

۱.۳ پیشگویی و اعتبارسنجی

از آن‌جا که کاهش بعد بسنده هیچ فرض مدل‌بندی را در مرحلهٔ کاهش بعد اعمال نمی‌کند مدلی را می‌توان بر مبنای متغیرهای کمکی شناسایی‌شده برازش داد. با توجه به این‌که در برازش داده‌های بقای سانسورشده، یکی از مدل‌های بسیار مختلف و مناسب، مدل خطرهای متناسب کاکس است، این مدل روی متغیرهای کاهش‌یافتهٔ حاصل از روش SIR سانسورشده برازش داده می‌شود. بر این اساس بهترین مدل‌های برازش‌یافته برای مردان و زنان به‌ترتیب زیر به دست می‌آیند:

$$\lambda(t) = \lambda_0(t)\exp(-3/23 SIR_1 - 0/215 SIR_1^2)$$

و

$$\lambda(t) = \lambda_0(t)\exp(-9/58 SIR_1 - 0/43 SIR_1^2).$$

به‌عبارت دیگر در بهترین مدل باقیمانده از برازش مدل خطرهای متناسب کاکس به سوهای معرفی شده در بخش قبل به دو گروه مردان و زنان اوّلین سوی متناظر با β_1 در مدل باقی مانده است. به‌منظور ارزیابی عمومی مجموعهٔ عوامل خطرزا در بیماری‌های عروق کرونر قلب افراد، امتیازات خطر CHD را محاسبه کرده تا بر این اساس برآوردی از بقای افراد به دست آیند.

جدول ۱. برآورد سوهای تصویر (β_1، β_2 و β_3) حاصل از روش رگرسیون وارون ورقه ورقه و برآورد ضریب رگرسیونی (β_0) حاصل از مدل کاکس برای داده‌های CHD در گروه مردان

عوامل خطرزا	برآورد مدل کاکس β_0	برآورد پایه اول β_1	برآورد پایهٔ دوم β_2	برآورد پایهٔ سوم β_3
سن	۰/۰۴۸	۰/۰۰۲	−۰/۰۴۵	−۰/۰۱۷
فشارخون سیستولیک	۰/۰۰۸	−۰/۰۰۲	−۰/۰۲۲	۰/۰۳۱
فشارخون دیاستولیک	۰/۰۰۶	۰/۰۰۴	۰/۰۰۵	−۰/۰۸۵
اندازهٔ دور کمر	۰/۰۰۳	۰/۰۰۱	−۰/۰۳۰	−۰/۰۲۷
نمایهٔ تودهٔ بدنی	−۰/۰۶۲	−۰/۰۱۰	۰/۰۷۶	۰/۱۸۲
قندخون ناشتا	−۰/۰۰۸	۰/۰۰۱	۰/۰۰۴	−۰/۰۲۹
کلسترول	۰/۰۰۵	۰/۰۰۰	۰/۰۰۴	−۰/۰۰۳
تری گلیسیرید	۰/۰۰۱	۰/۰۰۰	−۰/۰۰۱	۰/۰۰۱
سرم HDL	−۰/۰۲۱	−۰/۰۰۰	۰/۰۱۰	۰/۰۱۷
قندخون دو ساعته	−۰/۰۰۱	−۰/۰۰۰	۰/۰۰۱	−۰/۰۱۲
استعمال دخانیات	۰/۳۰۱	−۰/۲۲۳	−۰/۲۶۵	−۰/۱۵۸
مصرف داروی کاهندهٔ فشارخون	۰/۴۲۷	−۰/۱۴۶	−۰/۵۸۴	−۰/۸۸۴
مصرف داروی کاهندهٔ چربی خون	−۰/۲۹۶	۰/۸۸۹	۰/۶۹۶	۰/۸۵۶

جدول ۲. برآورد سوهای تصویر (β_1، β_2 و β_3) حاصل از روش رگرسیون وارون ورقه ورقه و برآورد ضریب رگرسیونی (β_0) حاصل از مدل کاکس برای داده‌های CHD در گروه زنان

عوامل خطرزا	برآورد مدل کاکس β_0	برآورد پایه اول β_1	برآورد پایهٔ دوم β_2
سن	۰/۰۶۶	۰/۰۶۸	−۰/۰۲۸
فشارخون سیستولیک	۰/۰۰۲	−۰/۰۰۵	−۰/۰۱۵
فشارخون دیاستولیک	۰/۰۰۶	۰/۰۳۵	۰/۰۱۸
اندازهٔ دور کمر	۰/۰۱۵	−۰/۰۹۸	−۰/۰۲۸
نمایهٔ تودهٔ بدنی	۰//۰۱۶	۰/۲۴۷	۰/۰۳۷
قندخون ناشتا	−۰/۰۱۳	۰/۰۰۹	۰/۰۱۱
کلسترول	۰/۰۰۳	۰/۰۰۳	−۰/۰۰۳
تری گلیسیرید	۰/۰۰۰	−۰/۰۰۱	−۰/۰۰۱
سرم HDL	۰/۰۱۱	−۰/۰۲۲	۰/۰۰۳
قندخون دو ساعته	۰/۰۰۶	−۰/۰۰۳	−۰/۰۰۳
استعمال دخانیات	۰/۲۲۸	−۰/۱۸۲	−۰/۰۰۵
مصرف داروی کاهندهٔ فشارخون	۰/۷۱۴	−۰/۸۹۸	−۰/۹۸۹
مصرف داروی کاهندهٔ چربی خون	−۰/۳۰۱	۰/۹۲۱	۰/۵۳۱

این امتیازات که از مدل برازش‌یافته فوق به دست می‌آیند عبارت‌اند از:

$$f(SIR) = -۳/۲۳SIR_۱ - ۰/۲۱۵SIR_۱^۲$$

برای مردان و

$$f(SIR) = -۹/۵۸SIR_۱ - ۰/۴۳SIR_۱^۲$$

برای زنان که از این طریق خطر CHD با عوامل مخاطره مرتبط می‌شود. اکنون می‌توان گروه‌های مخاطره را بر اساس امتیازاتشان تقسیم‌بندی کرد.

نمودارهای (a) و (b) در شکل (۱) برآوردهای کاپلان-مه‌یر بقای دو گروه کم‌مخاطره و پرمخاطره را به‌ترتیب برای مردان و زنان نشان می‌دهد. شایان توجه است که گروه کم‌مخاطره و پرمخاطره با معیار $f(SIR) > m$ یا $f(SIR) < m$ تعریف شده‌اند که در آن m میانهٔ همهٔ امتیازات است به‌عنوان مقدار برینش مناسب انتخاب شده است. میانگین بقا در گروه مردان کم‌مخاطره با تعداد ۱۹۱۰ نفر برابر ۳۹۵۴ و با انحراف معیار ۲/۱۱ و در گروه مردان پرمخاطره با تعداد ۴۷۳ نفر برابر۳۶۲۳ و با انحراف معیار ۳/۴۲ است که این اختلاف معنی‌دار است ($p = ۰/۰۰۰$). میانگین بقا در گروه زنان کم‌مخاطره با تعداد ۲۶۵۲ نفر برابر ۴۰۲۹ و با انحراف معیار ۱۶/۵ و در گروه زنان پرمخاطره با تعداد ۳۸۲ نفر برابر۳۷۵۷ و با انحراف معیار ۲۴/۴۰ است که این اختلاف معنی‌دار است ($p = ۰/۰۰۲$).

به‌منظور ارزیابی توان پیشگویی روش پیشنهادی، ۱۰۰ اعتبارسنجی به‌طور جداگانه برای گروه مردان و زنان انجام شده است. در هر اعتبارسنجی به‌طور تصادفی افراد به دو گروه داده‌های آموزشی و آزمایشی (برای مردان به‌ترتیب با اندازه‌های ۱۹۰۷ و ۴۷۶ و برای زنان به‌ترتیب با اندازه‌های ۲۴۲۸ و ۶۰۶) تقسیم‌بندی شدند. از آن‌جا که نتایج یک الگوی کیفی مشابه را نشان می‌دهند نتایج فقط برای مردان گزارش می‌شوند. نتایج اعتبارسنجی برای گروه مردان نشان می‌دهد که صرفاً در ۵ مورد از ۱۰۰ تکرار اعتبارسنجی سوهای تصویر معنی‌دار نشده است ($p > ۰/۰۵$). برای ۹۵ مورد باقیمانده آزمون لگ رتبه انجام شده تا آزمون شود سوهای معنی‌دار چگونه به خوبی می‌توانند نرخ‌های بقا را در دو گروه داده‌های آموزشی و آزمایشی از هم تفکیک کنند. برای داده‌های آموزشی میانه، چندک اول و سوم p–مقدارهای آزمون لگ رتبه ۰/۰۰۰ به دست آمد. برای

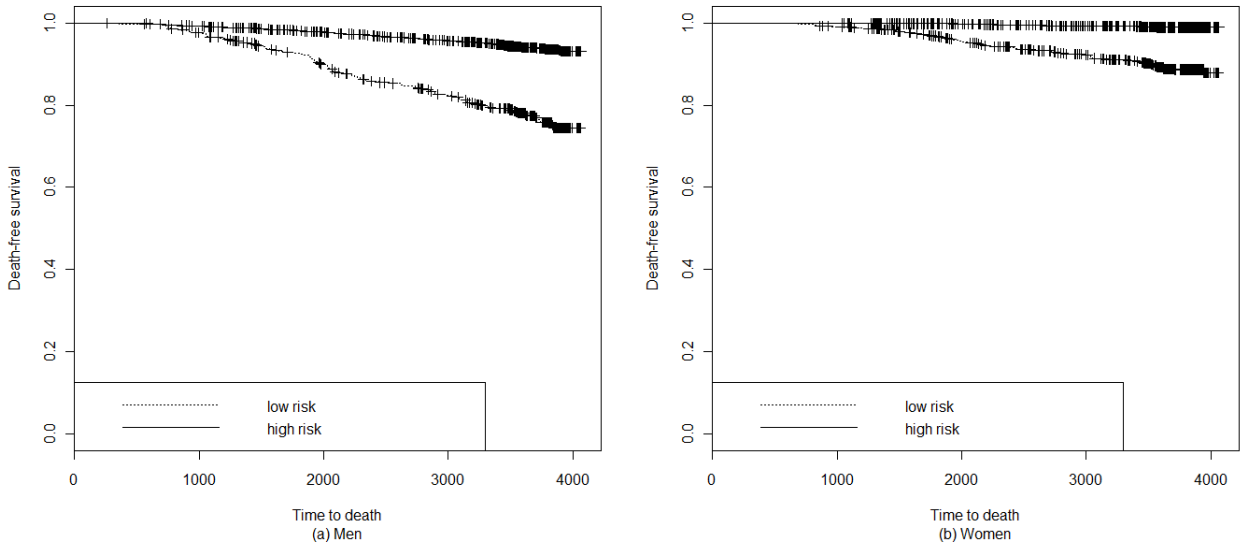
داده‌های آزمایشی میانه، چندک اول و سوم p–مقدارهای آزمون لگ رتبه نیز ۰٫۰۰۰ شده است. در میان ۹۵ اعتبار سنجی، ۸۸ مورد از آنها p–مقدارهای آزمایشی کوچک‌تر از ۰٫۰۵ شد.

بنا بر این در جمع‌بندی کلّی، نتیجه گرفته می‌شود که روش پیشنهادی حتی برای چنین نمونهٔ کوچک آموزشی با یک نرخ سانسور بالا، از توان پیشگویی بسیار خوبی برخوردار است.

بحث و نتیجه‌گیری

در تحلیل رگرسیونی داده‌های بقا، هدف اصلی بسط یک مدل پیشگو از طریق متغیرهای کمکی است. با وجود مدل‌های مختلفی که برای تحلیل داده‌های بقا وجود دارد اما در عمل دست‌یابی به یک مدل مناسب با تفسیرپذیری مطلوب کار آسانی نیست. در این پژوهش یک روش کاهش بعد بسنده برای تحلیل توأم داده‌های زمان بقا و عوامل مخاطره با استفاده از یک مدل کلّی که شامل مدل‌های بقای متداول به‌عنوان موارد خاص بوده مطرح شد. بسط روش رگرسیون وارون ورقه ورقه (SIR) از روش‌های کاهش بعد بسنده برای داده‌های سانسورشده اجرا شد. اعمال SIR برای داده‌های سانسورشده منجر به کاهش بعد عوامل مخاطره در قالب چند سوی تصویر شده است. امتیازات اصلی SIR برای داده‌های سانسورشده آن است که اولاً این روش هم اطلاعات زمان بقا و هم اطلاعات عوامل مخاطره را در نظر می‌گیرد. ثانیاً این روش به هیچ شکل تابعی از پیش‌تعیین‌شده‌ای برای رابطهٔ بین زمان بقا و سوهای تصویر نیاز ندارد. به‌عبارتی دیگر هیچ فرض پارامتری در روش کاهش بعد بسنده مورد نیاز نیست. به علاوه روش فوق برای بررسی کردن این‌که آیا یک مدل بقای متداول مناسب است، به‌وسیلهٔ بررسی ویژه مقدارها و نمودارهای با بعد پایین تولید شده به‌وسیلهٔ SIR به کار می‌رود. این نمودارها اطلاعات با ارزشی را در بارهٔ الگوی کلّی سانسور و وجود داده‌های دورافتاده و شکل سطح رگرسیونی ارائه می‌کنند. یک مجموعه‌داده از ۵۴۱۷ نفر از مطالعهٔ قند و لیپید تهران با استفاده از روش کاهش بعد بسنده و به‌ویژه رگرسیون وارون ورقه ورقهٔ سانسورشده تحلیل شد.

این روش که با مدل مخاطرهٔ متناسب کاکس ترکیب شد، در مجموع یک عملکرد پیشگویانهٔ خوب برای بقای آیندهٔ افراد ارائه کرد. سه سوی تصویر (ترکیب خطی از ۱۳ عامل مخاطره) برای مردان و دو سوی تصویر برای زنان شناسایی شد که به‌طور معنی‌دار با زمان بقا مربوط بود. مدل پیشگوی ساخته‌شده بر اساس مجموعه‌داده‌های آموزشی تفاوت معنی‌داری را بین گروه‌های بقا در داده‌های آزمایشی پیشگویی کرد.



شکل ۱. برآوردهای کاپلان مه‌یر نرخ‌های بقای افراد در دو گروه کم مخاطره و پرمخاطره: (a) مردان و (b) زنان.

تقدیر و تشکر

باتشکر از پژوهشکدهٔ علوم غدد درون ریز و متابولیسم دانشگاه علوم پزشکی شهید بهشتی که داده‌های مطالعه قند و لیپید تهران را برای انجام این پژوهش در اختیار قرار دادند. همچنین، نویسندگان مقاله از پیشنهادات و نظرات داوران و ویراستار محترم مجله که در بهبود مقاله مؤثر واقع شد، تقدیر و تشکر می‌نمایند.

مراجع

[۱] راستین، ا. (۱۳۹۲)، کاهش بعد بسنده برای داده‌های بقای سانسوریده، پایان نامه کارشناسی ارشد، دانشگاه شهید بهشتی، تهران.

[2] Azizi, F., Rahmani, M., Emami, H. and Madjid, M. (2000). Tehran lipid and glucose study, *Rationale and Design CVD prevention*, **3**, 242-47.

[3] Castelli, W. P. (1984). Epidemiology of coronary heart disease: The framingham study, *American Journal of Medicine*, **76**, 4-12.

[4] Chrysohoou, C., Panagiotakos, D. B., Pitsavos, C., Kokkinos, P., Marinakis, N., Stefanadis, C. and Toutouzas, P. K. (2003). Gender differences on the risk evaluation of acute coronary syndromes: the CARDIO 2000 study, *Journal of Preventive Cardiology*, **6**(**2**).

[5] Cook, R. D. (1998). Regression graphics, *Journal of the American Statistical Association*, **91**, 983-992.

[6] Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach, *Journal of the American Statistical Association*, **86**, 316-342.

[7] Hall, P. and Li, K.C. (1993). On almost linearity of low dimensional projections from high dimensional data, *Annals of Statistics*, **21**, 867-889.

[8] Li, K. C. (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 316-342.

[9] Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data, *Bioinformatics*, **20**, 3406-3412.

[10] Li, K. C., Wang, J. L. and Chen C. H. (1999). Dimension reduction for censored regression data, *The Annals of Statistics*, **27**, 1-23.

[11] Lee, E. T. and Wang, J. W. (1975). *Survival Methods for Survival Data Analysis*, John Wiley and Sons, New York.

[12] Lu, W. and Li, L. (2011). Sufficient dimension reduction for censored regressions, *Journal of the International Biometric society*, **67**, 513-523.

[13] Shevlyakova, M. and Morgenthaler, S. (2014). Sliced inverse regression for survival data, *Statistical Papers Springer*, **55**, 209-220.

[14] Yoo, J. K. (2017). Fused sliced inverse regression in survival analysis, *Communications for Statistical Applications and Methods*, **24**, 533-541.

[15] Yoo, J., Kima, S. J., Seoa, B. S., Shina, H. and Sima, S. A. (2016). Dimension reduction for right-censored survival regression: transformation approach, *Communications for Statistical Applications and Methods*, **23**, 93–103.