اصلاح روش رگرسیون وارون ورقه شده نوع دو برای داده‌های بقای سانسور شده

اعظم راستین، محمدرضا فریدروحانی

گروه آمار، دانشگاه شهید بهشتی تهران

چکیده: روش‌شناسی کاهش بعد بسنده یک راهکار مؤثر برای تسهیل در تحلیل رگرسیونی با داده‌های با بعد بالاست. هنگامی که پاسخ‌ها سانسور شده باشند، برآوردگرهای موجود را نمی‌توان به‌کار برد یا به شرایط محدودکننده‌ای نیاز است. در این مقاله، برای کاهش بعد داده‌های رگرسیونی سانسور شده غیرخطی، اصلاحی از روش رگرسیون وارون ورقه شده نوع دو پیشنهاد شده است. روش پیشنهادی، اولاً به هیچ مدل از پیش تعیین شده‌ای نیاز ندارد، ثانیاً اطلاعات کامل رگرسیونی را حفظ کرده و مجموعه کوچکی از ترکیب پیشگوها را ارائه می‌دهد که فرمول‌بندی مدل و پیش‌بینی براساس این مجموعه انجام می‌گیرد. در انتها عملکرد این روش، علاوه بر داده‌های شبیه‌سازی شده، برای مجموعه داده‌های واقعی سیروز صفراوی اولیه کبد مورد بررسی قرار گرفته و نتایج روش معرفی شده با روش رگرسیون وارون ورقه شده نوع یک مقایسه شده است.

واژه‌های کلیدی: رگرسیون سانسور شده، تحلیل بقا، کاهش بعد بسنده، زیر فضای کاهش بعد، رگرسیون وارون ورقه شده .

۱ مقدمه

داده‌های بقا اغلب با پدیده سانسور همراه هستند. هنگامی که این مسئله رخ می‌دهد، ناکامل بودن داده‌های مشاهده شده ممکن است منجر به اریبی قابل توجهی در نمونه شود (لو و لی، ۲۰۱۱)، روش‌هایی مانند مدل خطرهای متناسب کاکس، مدل بخت‌های متناسب، مدل زمان شکست شتابیده از روش‌های متداول بررسی پاسخ‌های سانسور بشمار می‌روند. با افزایش بعد متغیرهای کمکی، علاوه بر پیچیده‌تر شدن تحلیل‌های بقا، انتخاب یک مدل مناسب قبل از تحلیل داده‌ها نیازمند اطلاعاتی است که اغلب ناکافی بوده و هنگامی که بعد

°آدرس الکترونیک مسئول مقاله: اعظم راستین، rastinstat@gmail.com

°کد موضوع‌بندی ریاضی (۲۰۱۰): 62N01.

پیشگو بالاست، تعیین و تشخیص آن دشوارتر می‌شود (لی و همکاران، ۱۹۹۹)، کاهش بعد بسنده بدون نیاز به یک مدل از پیش تعیین شده و نیز بدون از دست دادن اطلاعات مرتبط با رگرسیون، بعد بردار پیشگو را کاهش می‌دهد. از این رو کاهش بعد بسنده، اغلب یک مسیر مطلوب برای تحلیل داده‌هایی با پیشگوهای با بعد بالا ارائه می‌دهد (کوک، ۱۹۹۸)، ایده کاهش بعد بسنده اولین بار توسط لی (۱۹۹۱) تحت عنوان روش رگرسیون وارون ورقه شده [۱] (SIR) و بر پایه گشتاور شرطی مرتبه اول، معرفی شد. روش رگرسیون وارون متداول‌ترین رده از روش‌های کاهش بعد بسنده است. ایده این روش، وارون کردن رابطه بین متغیر پاسخ اسکالر و متغیرهای پیشگو است. مزیت این تغییر نقش در این است که مسئله رگرسیون چند بعدی را به حل مسئله رگرسیون وارون یک بعدی تبدیل می‌کند تا نتیجه بهتری حاصل شود. لی و همکاران (۲۰۰۳) روش‌های کاهش بعد بسنده را برای پاسخ‌های چند متغیره بسط دادند. با به‌کار بردن روش‌های کاهش بعد بسنده، داده‌های رگرسیون سانسور شده را می‌توان بدون فرض شکل تابعی، تحلیل کرد. اما استفاده از روش‌های کاهش بعد بسنده برای تحلیل داده‌های بقا، به رابطه بین زمان سانسور و متغیرهای تبیینی بستگی دارد. لی و همکاران (۱۹۹۹) الگوهای مختلف سانسور را بررسی کردند و روش رگرسیون وارون ورقه کردن مضاعف [۲] (DSIR) را در رگرسیون وارون برای تحلیل داده‌های سانسور شده به‌کار گرفتند.

در سال‌های اخیر نیز بسط روش‌های کاهش بعد بسنده برای داده‌های سانسور شده مطرح شده است که به عنوان نمونه، ون و کوک (۲۰۰۹) روش‌های کاهش بعد را برای داده‌های سانسور شده پیشنهاد دادند. همچنین لی و لی (۲۰۰۴) روش‌های کاهش بعد بسنده را برای کاهش بعد داده‌های ریزآرایه به‌کار گرفتند، اما به‌علت بعد بالای این داده‌ها، روش رگرسیون وارون ورقه شده را نمی‌توان مستقیماً برای آنها به‌کار بست. بدین منظور آنها تحلیل مؤلفه‌های اصلی را همراه با رگرسیون وارون ورقه شده به‌کار گرفتند تا به کاهش بعد این داده‌ها دست یابند. کوک و نی (۲۰۰۵) برآوردگر رگرسیون وارون را پیشنهاد دادند که به طور مجانبی از روش‌های مبتنی بر گشتاورهای شرطی اول کاراتر بوده و نیز برای داده‌های سانسور شده بسط داده شده است (نادکارنی و همکاران، ۲۰۱۱)، لو و لی (۲۰۱۱) برآوردگر موزون وارون احتمال سانسور را معرفی کردند که روش رگرسیون وارون ورقه شده بر پایه کمترین توان‌های دوم انجام می‌شود. شیولیاکوف و مورگنتالیر (۲۰۱۴) همزمان با ورقه کردن زمان بقا، ضمن اختصاص وزن‌های مساوی به مشاهدات سانسور شده، یک ماتریس وزن را تشکیل دادند. همچنین یوو و همکاران (۲۰۱۶) با یک روش تبدیل، زمان بقا و زمان سانسور را به یک تک متغیره تبدیل

[1] Sliced Inverse Regression

[2] Double Slicing Inverse Regression

کردند. برای به‌کارگیری روش SIR در تحلیل بقا، یوو (۲۰۱۷) نیز روش رگرسیون وارون ورقه شده ترکیب شده را مطرح کرد.

در این مقاله ابتدا مروری بر روش‌های رگرسیون وارون ورقه شده نوع یک و دو که به ترتیب بر پایه گشتاورهای شرطی مرتبه اول و دوم هستند، صورت می‌گیرد. در واقع روش SIR-II برای فائق آمدن بر مشکل ناتوانی SIR-I در یافتن انواع مشخصی از روابط رگرسیونی غیرخطی پیشنهاد شده است (لی، ۱۹۹۱). سپس، روش SIR-II برای داده‌های رگرسیونی سانسور شده غیرخطی، بدون نیاز به شکل تابعی رگرسیونی از پیش مشخص شده، بسط داده می‌شود.

۲ کاهش بعد بسنده برای داده‌های کامل

نمونه تصادفی $(Y_1, X_1), \ldots, (Y_n, X_n)$ را در نظر بگیرید، که در آن Y_i پاسخ یک متغیره و X_i بردار p-بعدی از متغیرهای تبیینی است. فرض کنید Y بردار پاسخ‌های تک متغیره نمونه و $\mathbf{X}$ ماتریس طرح باشد.

۱.۲ روش رگرسیون وارون ورقه شده نوع یک

رگرسیون وارون ورقه شده، بعد پیشگوی p-بعدی $\mathbf{X}$ را با شناسایی بردار سوهای تصویر p-بعدی $(\beta_1, \ldots, \beta_k)$، $(k < p)$ در مدل

$$Y = g(\beta_1'\mathbf{X}, \ldots, \beta_k'\mathbf{X}, \varepsilon), \qquad (۱)$$

کاهش می‌دهد، که در آن g تابعی نامعلوم و ε خطای تصادفی مستقل از $\mathbf{X}$ است. زیرفضای تنیده شده توسط β_jها برای رگرسیون Y روی $\mathbf{X}$ تحت مدل (۱) یک زیرفضای کاهش بعد و هر یک از ترکیبات خطی β_jها یک سوی کاهش بعد مؤثر[3] نامیده می‌شود (لی، ۱۹۹۱)، تحت شرط خطی که بیان می‌کند برای سوهای $\mathbf{B}_{p\times k} = (\beta_1, \ldots, \beta_k)$ در مدل (۱) و هر ثابت $b \in R^p$، ثابت‌های $c_0 \in R^1$ و $c \in R^k$ وابسته به b وجود دارند به طوری که

$$E(b^T\mathbf{X} \,|\, \mathbf{B}^T\mathbf{X}) = c_0 + c^T\mathbf{B}^T\mathbf{X}, \qquad (۲)$$

[3] Effective Dimension Reduction (EDR)

منحنی وارون $E(\mathbf{X}\,|Y=y)$ در زیرفضای تنیده شده توسط β_jها، $j=۱,\ldots,k$ قرار می‌گیرد. همانطور که کوک و ویسبرگ (۱۹۹۱) نشان دادند، مهم‌ترین خانواده توزیع‌هایی که در شرط خطی بودن صدق می‌کنند، توزیع متقارن بیضی‌وار مانند توزیع نرمال است. در روش SIR-I ابتدا تکیه‌گاه متغیر پاسخ Y به h $(h>k)$ ورقه مجزا تقسیم می‌شود و در نهایت سوهای تصویر $\beta_۱,\ldots,\beta_k$، ویژه بردارهای ماتریس $Cov[E(\mathbf{X}\mid Y)]$ خواهند بود (لی، ۱۹۹۱)، تعداد سوهای تصویر معنی‌دار که با k نشان داده می‌شود می‌تواند بوسیله یک آزمون مجانبی خی‌دو تعیین شود. در واقع آزمون فرض‌های $k=m$ در مقابل $k>m$، برای $m=۰,\ldots,p-۱$، به صورت دنباله‌ای انجام می‌شود.

۲.۲ روش رگرسیون وارون ورقه شده نوع دو

در برخی موارد روش SIR-I قادر به یافتن سوهای EDR نیست که با جایگزین کردن $Cov(\mathbf{X}|Y)$ به جای $E(\mathbf{X}|Y)$ می‌توان بر این مشکل فائق آمد. برای مثال، وقتی $(X_۱,X_۲)'\sim N(۰,I_۲)$ و $Y=X_۱^۲$، روش‌های بر پایه $E(\mathbf{X}|Y)$، نمی‌توانند سوهای تصویر را تعیین کنند. اما واریانس شرطی $Var(X_۱|Y=y)=E(X_۱^۲|Y=y)=y$ روشی جایگزین برای پیدا کردن سوهای EDR ارائه می‌دهد (لی، ۱۹۹۱)، ایده SIR-II بر پایه کواریانس شرطی

$$V=E\{Cov(\mathbf{X}|Y=y)-E(Cov(\mathbf{X}|Y=y))\}^۲, \qquad (۳)$$

است (لی، ۱۹۹۱و ۱۹۹۲)، روش SIR-II بسیار مشابه روش SIR-I است و تنها با این تفاوت که به جای محاسبه میانگین، کواریانس هر ورقه محاسبه می‌شود.

۱.۲.۲ الگوریتم SIR-II

۱. ابتدا X_iها را با تبدیل $Z_i=\hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-\frac{۱}{۲}}(X_i-\bar{X})$ استاندارد کنید، که در آن $\hat{\mathbf{\Sigma}}_{\mathbf{X}}$ و $\bar{X}$ به ترتیب ماتریس کواریانس نمونه‌ای و میانگین نمونه‌ای X_iها هستند.

۲. برد Y را به H ورقه نامتداخل $I_۱,\ldots I_H$ افراز کنید. در این صورت تعداد مشاهدات درون ورقه hام $(h=۱,\ldots,H)$، برابر $n_h=\sum_{i=۱}^{n} I_h(y_i)$، است، که در آن I_h تابع نشانگر ورقه hام است.

۳. درون هر ورقه h، میانگین نمونه‌ای Z_iها را به صورت $\bar{Z}_h=\frac{۱}{n_h}\sum_{i=۱}^{n} Z_iI_h(y_i)$، محاسبه کنید.

۴. ماتریس کواریانس ورقه‌ای $\hat{\mathbf{V}}_h = \frac{۱}{n_h - ۱}\sum_{i=۱}^{n} I_h(y_i) z_i z_i' - n_h \bar{z_h}\bar{z_h'}$ را محاسبه کنید.

۵. میانگین همه ماتریس‌های کواریانس ورقه‌ای $\bar{\mathbf{V}} = \frac{۱}{n}\sum_{h=۱}^{H} n_h \hat{\mathbf{V}}_h$ را محاسبه کنید.

۶. ویژه‌مقدارها و ویژه‌بردارهای ماتریس $\hat{\mathbf{V}} = \frac{۱}{n}\sum_{h=۱}^{H} n_h(\hat{\mathbf{V}}_h - \bar{\mathbf{V}})^۲ = \frac{۱}{n}\sum_{h=۱}^{H} n_h \hat{\mathbf{V}}_h{}^۲ - \bar{\mathbf{V}}^۲$ را بیابید. اگر $\hat{\gamma}_j$ها بردار ویژه‌های این ماتریس باشند، آنها را به مقیاس $\mathbf{X}$ تبدیل کنید:

$$\hat{\beta}_j = \mathbf{\Sigma}_{\mathbf{X}}^{-\frac{۱}{۲}} \hat{\gamma}_j \qquad j = ۱, \ldots, k$$

۳ کاهش بعد بسنده برای داده‌های بقای سانسورشده

وقتی پاسخ‌ها سانسور شده باشند، روش رگرسیون وارون ورقه شده را نمی‌توان به‌کار برد یا برای استفاده از آنها شرایط محدودکننده‌ای لازم است. در ادامه، T به عنوان زمان بقای واقعی (غیرقابل مشاهده)، C زمان سانسور، δ نشانگر سانسور، $\delta = I(T \leq C)$ و Y زمان بقای مشاهده شده $Y = \min(T, C)$ در نظر گرفته شده است. فرض کنید زمان بقای T از الگوی کاهش بعد (۱) پیروی می‌کند و نیز $T \perp C|\mathbf{X}$، که فرض استقلال متداول برای شناساپذیری تحت سانسور تصادفی است.

یک روش ایده‌آل برای رفع مشکل ناشی از سانسور ورقه کردن زمان بقای واقعی T است. اما از آنجا که T غیرقابل مشاهده است، امید ریاضی شرطی Z، متغیرهای پیشگوی تبدیل‌یافته در گام ۱ الگوریتم SIR-II در هر ورقه را با زمان مشاهده شده Y و نشانگر سانسور مرتبط می‌سازد.

اگر $۰ = t_۱ < \ldots < t_H < \infty$ افرازی روی زمان بقا باشد، $\mathbf{m}_j^{۱} = E\{\mathbf{Z}|T \in [t_j, t_{j+۱})\}$ مقدار مورد انتظار $\mathbf{Z}$ روی ورقه jام ($j = ۱, \ldots, H$) خواهد بود. لی و همکاران (۱۹۹۹) نشان دادند که $\mathbf{m}_j^{۱}$ را می‌توان به‌صورت

$$\mathbf{m}_j^{۱} = \frac{E\{\mathbf{Z}I(T \in [t_j, t_{j+۱}))\}}{P\{T \in [t_j, t_{j+۱})\}} = \frac{E\{\mathbf{Z}I(T \geq t_j)\} - E\{\mathbf{Z}I(T \geq t_{j+۱})\}}{E\{I(T \geq t_j)\} - E\{I(T \geq t_{j+۱})\}}, \qquad (۴)$$

بازنویسی کرد، که در آن $I(\cdot)$ تابع نشانگر است. توجه شود که در رابطه (۴) برابری‌های

$$E\{\mathbf{Z}I(T \geq t)\} = E\{\mathbf{Z}I(Y \geq t)\} + E\{\mathbf{Z}I(Y < t, \delta = ۰)w(Y, t, z)\}, \quad (۵)$$

$$E\{I(T \geq t)\} = E\{I(Y \geq t)\} + E\{I(Y < t, \delta = ۰)w(Y, t, z)\}, \qquad (۶)$$

برقرار هستند، که در آن برای $t^{'} < t$

$$w(t^{'}, t, z) = \frac{\bar{F}_T(t|\mathbf{Z})}{\bar{F}_T(t^{'}|\mathbf{Z})} = \exp\{-\Lambda(t^{'}, t|\mathbf{Z})\}, \quad (۷)$$

و $\bar{F}(.|\mathbf{Z})$ تابع بقای شرطی $T|\mathbf{Z}$، یعنی $\bar{F}(t|\mathbf{Z}) = P(T \geq t|\mathbf{Z})$ است. به‌علاوه

$$\Lambda(t^{'}, t|\mathbf{Z}) = E\{\frac{I(t^{'} < Y < t, \delta = ۱)}{\bar{F}_Y(Y|\mathbf{Z})}|\mathbf{Z}\}.$$

به‌طور مشابه، می‌توان عبارت $\mathbf{m}_j^{۲} = E\{\mathbf{Z}\mathbf{Z}'|T \in [t_j, t_{j+۱})\}$, را در نظر گرفت. بنابراین

$$\mathbf{m}_j^{۲} = \frac{E\{\mathbf{Z}\mathbf{Z}'I(T \geq t_j)\} - E\{\mathbf{Z}\mathbf{Z}'I(T \geq t_{j+۱})\}}{E\{I(T \geq t_j)\} - E\{I(T \geq t_{j+۱})\}}, \quad (۸)$$

که در آن

$$E\{\mathbf{Z}\mathbf{Z}'I(T \geq t)\} = E\{\mathbf{Z}\mathbf{Z}'I(Y \geq t)\} + E\{\mathbf{Z}\mathbf{Z}'I(Y < t, \delta = ۰)w(Y, t, z)\}. \quad (۹)$$

با استفاده از

$$Cov(\mathbf{Z}|Y) = E(\mathbf{Z}\mathbf{Z}'|Y) - E(\mathbf{Z}|Y)E(\mathbf{Z}'|Y),$$

و قرار دادن روابط (۵)-(۱۰) در (۳) می‌توان تجزیه ویژه‌مقدار را در الگوریتم SIR-II برای بدست آوردن برآوردهای SIR-II اصلاح شده، اجرا کرد.

اینک با استفاده از روابط (۴) و (۸) و با جایگزینی امید ریاضی‌ها با گشتاورهای نمونه‌ای، برآوردگرهای

$$\hat{\mathbf{m}}_j^{۱} = \frac{\hat{E}\{\mathbf{Z}I(T \geq t_j)\} - \hat{E}\{\mathbf{Z}I(T \geq t_{j+۱})\}}{\hat{P}\{T \geq t_j\} - \hat{P}\{T \geq t_{j+۱}\}}, \quad (۱۰)$$

$$\hat{\mathbf{m}}_j^{۲} = \frac{\hat{E}\{\mathbf{Z}\mathbf{Z}'I(T \geq t_j)\} - \hat{E}\{\mathbf{Z}\mathbf{Z}'I(T \geq t_{j+۱})\}}{\hat{P}\{T \geq t_j\} - \hat{P}\{T \geq t_{j+۱}\}}, \quad (۱۱)$$

بدست می‌آیند، که در آن‌ها

$$\hat{E}\{\mathbf{Z}I(T \geq t)\} = n^{-۱} \sum_{i:Y_i \geq t}^{n} Z_i + n^{-۱} \sum_{i:Y_i < t, \delta_i = ۰}^{n} Z_i \hat{w}(Y_i, t, Z_i),$$

$$\hat{E}\{\mathbf{Z}\mathbf{Z}'I(T \geq t)\} = n^{-۱} \sum_{i:Y_i \geq t}^{n} Z_i Z_i{}' + n^{-۱} \sum_{i:Y_i < t, \delta_i = ۰}^{n} Z_i Z_i{}' \hat{w}(Y_i, t, Z_i),$$

$$\hat{P}\{T \geq t\} = \#\{i : Y_i \geq t\}/n + n^{-۱} \sum_{i:Y_i < t, \delta_i = ۰}^{n} \hat{w}(Y_i, t, Z_i),$$

و $\hat{w}(\cdot,\cdot,\cdot)$ برآوردی از تابع وزن (۷) است، که برای به‌دست آوردن آن از روش‌های هموارسازی استفاده شده است. در این مقاله از هموارسازی هسته‌ای (بیران، ۱۹۸۱) بهره گرفته شده است. فرض کنید $K_p(\cdot)$ یک تابع هسته روی R^p و h_n پهنای باند یا پارامتر هموارسازی، در هر یک از مختصات آن باشد. تابع هسته باید در شرط $\int K(u)\,du = 1$ صدق کند و معمولاً این تابع حول نقطه صفر متقارن است. برآورد هسته‌ای (۷) به‌صورت

$$\hat{\Lambda}(t',t|\mathbf{Z}) = \frac{\sum_{i:t'<Y_i<t,\delta_i=1}^{n} (\hat{\bar{F}}_Y(Y_i|Z_i))^{-1} h_n^{-p} K_p(h_n^{-1}(Z_i-\mathbf{Z}))}{n\hat{f}(\mathbf{Z})},$$

تعریف می‌شود، که در آن

$$\hat{\bar{F}}_Y(Y_i|Z_i) = \frac{\sum_{j:Y_j>Y_i}^{n} h_n^{-p} K_p(h_n^{-1}(Z_j-Z_i))}{n\hat{f}(Z_i)},$$

$$\hat{f}(\mathbf{Z}) = n^{-1}\sum_{i}^{n} h_n^{-p} K_p(h_n^{-1}(Z_i-\mathbf{Z})),$$

بعد از برآورد میانگین ورقه‌ها از (۱۰) و (۱۱)، می‌توان $\hat{\mathbf{V}}$ را به روش معمول به صورت

$$\hat{\mathbf{V}} = \sum_{j=1}^{H} \hat{p}_j\{\hat{m}_j^2 - \hat{m}_j^1\hat{m}_j^1\} - \sum_{j=1}^{H} \hat{p}_j[\hat{m}_j^2 - \hat{m}_j^1\hat{m}_j'^{1}]\}^2,$$

به‌دست آورد، که در آن

$$\hat{p}_j = \hat{P}\{Y \geq t_j\} - \hat{P}\{Y \geq t_{j+1}\}.$$

۴ مطالعه شبیه‌سازی

در این بخش با استفاده از شبیه‌سازی ضمن ارزیابی برآوردگر اصلاح شده DSIR-II این برآوردگر با برآوردگر متداول DSIR-I (لی و همکاران، ۱۹۹۹) در رگرسیون‌های سانسور شده، مقایسه می‌شود. برای این منظور دو مدل شکست شتابیده و کاکس در نظر گرفته شده است. زمان بقای واقعی از دو مدل

$$T = \exp(x_1 + x_3)\varepsilon_1,$$

$$T = -\log(\varepsilon_2)/\exp(x_1 + x_3),$$

تولید می‌شود، که در آن ε_1 از توزیع نمایی با پارامتر ۱ و ε_2 از توزیع یکنواخت روی بازه $[0, 1]$ پیروی می‌کنند. همچنین $X = (x_1, \ldots, x_6)$ از توزیع نرمال استاندارد تولید می‌شود. از این رو، با توجه به اینکه فقط x_1 و x_3 در تولید زمان T دخالت دارند، $\gamma^T = (1, 0, 1, 0, 0, 0)$، و بعد ساختاری $d = 1$. زمان سانسور C نیز از مدل $C = \exp(x_1 + x_2 + x_3)^4$ تولید می‌شود و نسبت سانسور در سطح ۴۵ درصد کنترل شده است. اندازه نمونه ۵۰، ۱۰۰، ۲۰۰، ۴۰۰ و ۸۰۰ تغییر می‌یابد تا تاثیر اندازه نمونه بر برآوردها نیز بررسی شود. همچنین برای بررسی روش پیشنهادی در وقتی بعد بردار پیشگو افزایش می‌یابد، p از ۶ به ۱۰، ۱۵ و ۲۰ تغییر می‌یابد و در این حالت اندازه نمونه برابر ۲۰۰ ثابت نگه داشته می‌شود. پیشگوهای اضافه شده همچنان از توزیع نرمال استاندارد تولید می‌شوند.

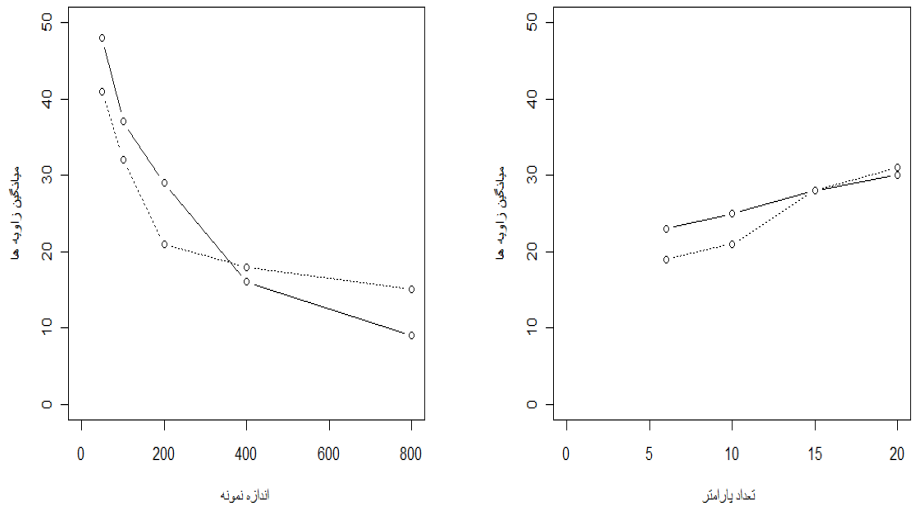
بردار ضرایب همبستگی برای ارزیابی تفاوت بین زیرفضاهای کاهش بعد واقعی و برآورد شده به‌کار می‌رود. مقادیر این معیار بین صفر و یک قرار دارند و مقدارهای بزرگ ضریب همبستگی نشان‌دهنده دقت بالای برآورد است. در شکل ۱، که میانه بردار همبستگی را در ۱۰۰ تکرار نشان می‌دهد، ملاحظه می‌شود که روش اصلاحی DSIR-II در هر دو مورد، برتر از روش DSIR-I عمل می‌کند.

۵ تحلیل داده‌های PBC

مجموعه داده‌های PBC مربوط به آزمایش کلینیکی مایو از بیماری سیروز صفراوی اولیه کبد است که در یک مطالعه ۱۰ ساله بین سالهای ۱۹۸۴-۱۹۷۴ به‌دست آمده است (فلیمینگ و هارینگتون، ۱۹۸۴). در این بخش به پیروی از مطالعه لی و همکاران (۱۹۹۹) روی ۳۰۶ نفر و ۶ متغیر کمکی که در مدل‌بندی زمان بقای این افراد مهم هستند، تمرکز شده است. برای تحلیل این داده‌ها با روش DSIR-II ، فرض کنید δ نشانگر سانسور و Y زمان بقای مشاهده شده باشد که بیانگر تعداد روزهای بین ثبت نام و قبل از مرگ یا سانسور است. متغیرهای تبیینی عبارتند از: سن (Age) برحسب سال، وجود ورم $(Edema)$ که در آن ۰ : نبود ورم یا نبود ورم دیورتیک، $\frac{1}{2}$: وقتی دیورتیک درمانی داده نشده یا ورم با دیورتیک درمانی رفع شده است، ۱: وجود ورم علی‌رغم دیورتیک درمانی، بیلی روبین سرم $(Bilirubin)$ برحسب mg/dl، آلبومین $(Albumin)$ بر حسب mg/dl، زمان پروترومبین $Prothrombin(protime)$، برحسب ثانیه، تعداد پلاکت‌ها $(platelet\,count)$ برحسب $ml/1000$.

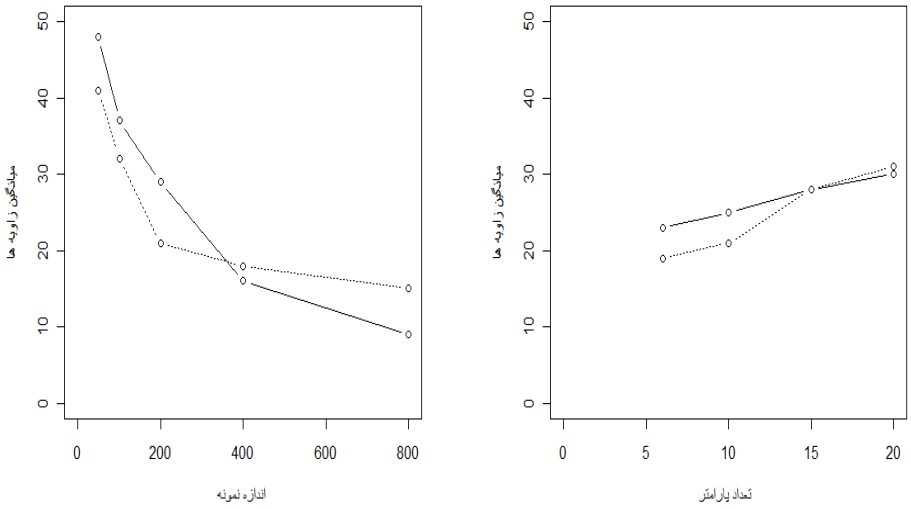
تیبشیرانی (۱۹۹۷) این مجموعه داده‌ها را در بررسی انتخاب متغیر، در مدل مخاطره‌های متناسب استفاده

کرد. او با بکارگیری روش بازگشتی، ۸ متغیر کمکی را انتخاب کرد که در بین آنها بجز متغیر تعداد پلاکت‌ها، ۵ متغیر کمکی فوق حضور دارند. آزمون بعد نشان می‌دهد که بعد زیرفضای کاهش بعد، دو است، بنابراین تصویر داده‌ها در دو سوی تصویرهای $DSIRII_۱ = \beta_۱'\mathbf{X}$ و $DSIRII_۲ = \beta_۲'\mathbf{X}$ در نظر گرفته شده است. برآوردهای پایه $\beta_۱$ و $\beta_۲$ متغیرهای پیشگوی متناظر و برآوردهای حاصل از برازش مدل رگرسیونی مخاطره متناسب کاکس نیز در جدول ۱ ارائه شده است. همان‌طور که ملاحظه می‌شود متغیرهای آلبومین و بیلی روبین در هر دو سوی تصویر ضرایب بالاتری دارند که با برآوردهای حاصل از رگرسیون کاکس ($\beta_۰$) سازگار است. در مقایسه با روش DSIR–I ، روش پیشنهادی نقش مهم‌تری برای متغیر بیلی روبین قائل است که تاثیرگذاری این متغیر قبلا توسط ژیا و همکاران (۲۰۱۰) بیان شده است. همچنین در روش پیشنهادی، متغیر وجود ورم و زمان پروترومبین تنها در یک سو، ضریب بالایی منعکس می‌کند، البته در مطالعه ون (۲۰۱۰) بر روی همین داده‌ها که با استفاده از روش‌های کاهش بعد بسنده صورت گرفته است، متغیر وجود ورم حذف شده و مطابق انتظار متغیر تعداد پلاکت در هر دو روش بی تاثیر شناخته شده است.



(الف) (ب)

شکل ۱: میانگین زوایای بین پایه واقعی زیرفضای مرکزی و برآوردهای DSIR–I (خط تیره) و DSIR–II (نقطه چین) در مدل ۱، الف: اندازه نمونه‌های متفاوت و $p = ۶$ ب: تعداد متغیرهای پیشگو و $n = ۲۰۰$

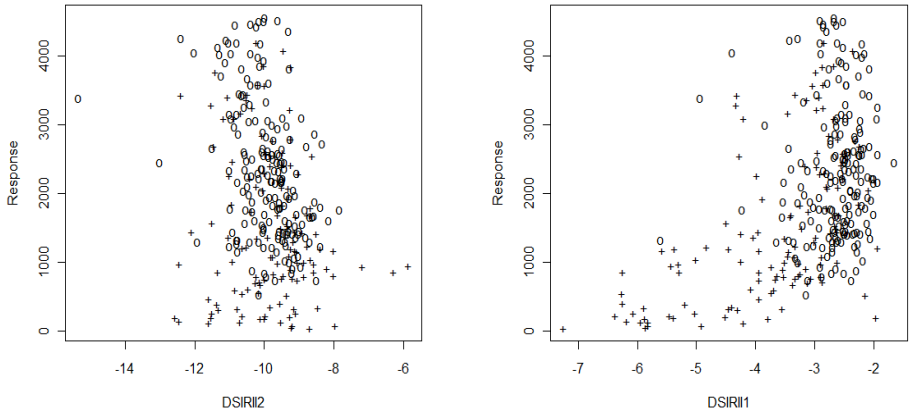


(ب) (الف)

شکل ۲: میانگین زوایایی بین پایه واقعی زیرفضای مرکزی و برآوردهای DSIR-I (خط تیره) و DSIR-II (نقطه چین) در مدل ۲، الف: اندازه نمونه‌های متفاوت و $p = ۶$ ب: تعداد متغیرهای پیشگو و $n = ۲۰۰$

جدول ۱: برآوردهای مدل کاکس و پایه زیرفضای مرکزی با بعد $d = ۲$ برای داده‌های PBC

مدل رگرسیونی کاکس	DSIR-I		DSIR-II	
متغیر کمکی	برآورد مدل کاکس	برآوردپایه اول	برآوردپایه دوم	برآوردپایه اول	برآوردپایه دوم
سن	۰/۰۳	۰/۰۲	۰/۰۳	۰/۰۱	-۰/۰۲
وجود ورم	۰/۷۸	۰/۹۰	-۲/۳	۰/۲۰	۰/۰۱
بیلی روبین	۰/۸۸	۰/۰۹	۰/۲۰	۰/۶۹	-۰/۳۷
آلبومین	-۳/۰۵	-۰/۶۲	-۰/۲۸	۰/۳۲	۰/۱۱
تعداد پلاکت	۰/۰۰	-۰/۰۰	-۰/۰۰	۰/۰۰	۰/۰۰
زمان پروترومبین	۳/۰۱	۰/۳۸	-۰/۶۸	۰/۴۱	-۰/۰۸



(ب) (الف)

شکل ۳: پراکنش پیشگوها با روش الف- DSIR-I و ب- DSIR-II ، رخداد پیشامد (+) و سانسور (o)

در شکل ۳ نمودار پراکنش مؤلفه‌های اول و دوم حاصل از روش DSIR-II در برابر متغیر پاسخ رسم شده است. همانطور که ملاحظه می‌شود مؤلفه‌های اول و دوم به ترتیب الگوهای نسبتاً خطی و غیر خطی را در داده‌ها نشان می‌دهند. ضرایب همبستگی $\beta_1'\mathbf{X}$ و $\beta_2'\mathbf{X}$ با $\beta_0'\mathbf{X}$ به ترتیب ۰/۷۹ و ۰/۰۹− هستند. از آنجا که مؤلفه خطی می‌تواند به خوبی بوسیله مدل خطرهای متناسب کاکس برآورد شود، این ضرایب دوباره شکل (۳) را تایید می‌کنند. در مطالعه لی و همکاران (۱۹۹۹) ضریب همبستگی $\beta_1'\mathbf{X}$ و $\beta_2'\mathbf{X}$ با $\beta_0'\mathbf{X}$ به ترتیب ۰/۸۵ و ۰/۸۹ گزارش شده است که تنها یک الگوی خطی را در داده‌ها پیدا می‌کند.

بحث و نتیجه‌گیری

مسئله کاهش بعد برای رگرسیون‌های سانسور شده غیرخطی با یک بردار از پیشگوهای با بعد بالا مورد بررسی قرار گرفت. روش پیشنهادی مبتنی بر رگرسیون وارون است، که در آن داده‌های رگرسیونی سانسور شده می‌توانند بدون فرض شکل تابعی از پیش مشخص شده تحلیل شوند. گرچه روش رگرسیون وارون ورقه شده SIR-I در تعیین روند خطی عملکرد خوبی دارد، ولی برای فرم درجه دوم مناسب نیست، از این رو بهتر است روش برآورد SIR-II که مبتنی بر گشتاور شرطی مرتبه دوم است، مورد استفاده قرار گیرد. بنابراین روش SIR-II با روش

هموارسازی هسته‌ای در برآورد تابع وزن اصلاح شد تا بتوان سوهای بقای کاهش بعد را پیدا کرد. اصلاح و بسط این روش تقریباً مشابه بسط روش SIR-I است که توسط لی و همکاران (۱۹۹۹) ارائه شده است. هرچند ادعا نمی‌شود روش پیشنهادی برتر از روش‌های دیگر است، اما نشان داده شد برآوردگر DSIR-II جایگزینی بهتر و مفید برای برآوردگر DSIR-I در کاهش بعد بسنده با پاسخ‌های سانسور شده است.

تقدیر و تشکر

نویسندگان مقاله از پیشنهادات و نظرات داوران و ویراستار محترم مجله که در بهبود مقاله موثر 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] Beran, R. Z. (1981), *Nonparametric Regression with Randomly Censored Survival Data*, Technical report, Univ. California, Berkeley.

[3] Cook, R. D. (1998), Regression Graphics, *Journal of the American Statistical Association*, **91**, 983–992.

[4] 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.

[5] Cook, R. D. and Weisberg, S. (1991), Discussion of Sliced Inverse Regression for Dimension Reduction, *Journal of the American Statistical Association*, **86**, 316-342.

[6] Fleming, T. R. and Harrington, D. P. (1984), Nonparametric Estimation of Survival Distribution in Censored Data, *Comunications in Statistics*, **13**, 2469–2486.

[7] Li, K. C. (1991), Sliced Inverse Regression for Dimension Reduction (with discussion), *Journal of the American Statistical Association*, **86**, 316-342.

[8] Li, K. C. (1992), On Principal Hessian Directions for Data Visualization and Dimension Reduction: another application of Stein's lemma, *Journal of the American Statistical Association*, **87**, 1025-1039.

[9] Li, L. and Li, H. (2004), Dimension Reduction Methods for Microarrays with Application to Censored Survival Data, *Bioinformatics*, **20**, 3406–3412.

[10] Lu, W. and Li, L. (2011), Sufficient Dimension Reduction for Censored Regressions, *ournal of the International Biometric society*, **67**, 513-523.

[11] Li, K. C., Wang, J. L. and Chen, C. H. (1999), Dimension Reduction for Censored Regression Data, *The Annals of Statistics*, **27**, 1-23.

[12] Nadkarni, N. V., Zhao and Y., Kosorok, M. (2011), Inverse Regression Estimation for Censored Data, *Journal of the American Statistical Association* , **106**, 178–190.

[13] Shevlyakova, M. and Morgenthaler, S. (2014), Sliced Inverse Regression for Survival Data, *Statistical Papers, Springer*, **55**, 209-220.

[14] Tibshirani,R. (1997), The Lasso Method for Variable Selection in the Cox Model, *Statistics in Medicine*, **27**, 385-395.

[15] Wen, X. (2007), A note on Sufficient Dimension Reduction, *Statistics and Probability Letters*, **77**, 817-821.

[16] Wen, X. and Cook, D. R. (2009), New Approaches to Model-Free Dimension Reduction for Bivariate Regression, *Journal of Statistical Planning and Inference*, **139**, 734–748.

[17] Xia, Y., Zhang, D. and Xu, J. (2010), Dimension Reduction and Semiparametric Estimation of Survival Models, *Journal of the American Statistical Association*, **105**, 278–290.

[18] Yoo, J. K. (2017), Fused Sliced Inverse Regression in Survival analysis, *Communications for Statistical Applications and Methods*, **24**, 533-541.

[19] 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.

[20] Zeng, D., and Lin, D. Y. (2007), Efficient Estimation in the Accelerated Failure Time Model, *Journal of the American Statistical Association,*.

[21] Zhang, H., and Lu, W. (2007), Adaptive-Lasso for Cox's Proportional Hazards Model, *Biometrika*, **94**, 1–13.

Modification of Sliced Inverse Regression to Censored Survival Data

Rastin, A., Faridrohani, M. R.

Department of Statistics, Shahid Beheshti University, Tehran , Iran.

Abstract: The methodology of sufficient dimension reduction has offered an effective means to facilitate regression analysis of high-dimensional data. When the response is censored, most existing estimators cannot be applied, or require some restrictive conditions. In this article modification of sliced inverse, regression-II have proposed for dimension reduction for non-linear censored regression data. The proposed method requires no model specification, it retains full regression information, and it provides a usually small set of composite variables upon which subsequent model formulation and prediction can be based. Finally, the performance of the method is compared based on the simulation studies and some real data set include primary biliary cirrhosis data. We also compare with the sliced inverse regression-I estimator.

Keywords: Censored regression, Survival analysis, Sufficient dimension reduction, Dimension reduction subspace, Sliced inverse regression.

Mathematics Subject Classification (2010): 62N01.