Archive of SID

دانشور
پزشکی

تحلیل داده‌های رتبه‌ای و همبسته پزشکی به کمک معادلات برآوردگر تعمیم یافته

نویسندگان: فرید زایری[۱]، دکتر انوشیروان کاظم نژاد[۲]، دکتر غلامرضا بابایی[۲]، دکتر مجتبی گنجعلی[۳] و دکتر محمدجواد خرازی‌فرد[۴]

۱. دانشجوی دکترای آمار زیستی، دانشکده پزشکی، دانشگاه تربیت مدرس
۲. دانشیار گروه آمار زیستی، دانشکده پزشکی، دانشگاه تربیت مدرس
۳. استادیار گروه آمار، دانشکده علوم ریاضی، دانشگاه شهید بهشتی
۴. دندان‌پزشک و اپیدمیولوژیست، دانشکده دندان‌پزشکی، دانشگاه علوم پزشکی تهران

چکیده

سابقه: مدل‌سازی داده‌های چند متغیره، همبسته و رتبه‌ای پزشکی معمولاً دشوارتر از تحلیل داده‌های پیوسته یا دو حالتی است. برآورد پارامترهای رگرسیونی در چنین مدل‌هایی به سبب ماهیت همبسته و رتبه‌ای داده‌ها، با روش‌های معمول، نظیر روش حداکثر درستنمایی بسیار زمان‌بر و مستلزم طراحی برنامه‌های کامپیوتری پیچیده است.

روش‌بررسی: در این مقاله، طریقه به‌کارگیری یک مدل رگرسیون حاشیه‌ای بخت‌های متناسب برای تحلیل این گونه داده‌ها و همچنین استفاده از روش معادلات برآوردگر تعمیم یافته را به‌منظور برآورد پارامترهای این مدل مورد بررسی قرار می‌دهیم. این روش مبتنی بر برآورد شبه درستنمایی بوده، انجام آن بسیار ساده‌تر از دیگر روش‌های معرفی شده در این زمینه است. همچنین، ساختارهای مختلفی برای توصیف همبستگی بین متغیرهای پاسخ ارائه و روش برآورد هر یک معرفی می‌شود.

یافته‌ها: روش تشریح شده در داده‌های به‌دست آمده از مطالعه وضعیت پریودنتال دانش‌آموزان ۱۹-۱۵ ساله تهرانی مورد استفاده قرار گرفت. با مقایسه نتایج حاصل از برازش مدل به کمک معادلات برآوردگر تعمیم‌یافته و روش حداکثر درستنمایی، مشخص شد که این روش برازش مناسب‌تری نسبت به روش حداکثر درستنمایی به‌دست می‌دهد.

بحث: روش معادلات برآوردگر تعمیم‌یافته، روشی مناسب و ساده برای تحلیل داده‌های همبسته و رتبه‌ای پزشکی است که می‌توان آن را جانشین روش‌های پیچیده‌تری نظیر حداکثر درستنمایی کرد.

واژه‌های کلیدی: مدل‌های خطی تعمیم یافته، معادلات برآوردگر تعمیم یافته، داده‌های رتبه‌ای همبسته، وضعیت پریودنتال

دوماهنامه علمی-پژوهشی دانشگاه شاهد
سال دوازدهم - شماره ۵۴
دی ۱۳۸۳

www.SID.ir

Archive of SID

مقدمه

در دو دهه اخیر، توجه بسیاری از آماردانان به تجزیه و تحلیل داده‌های چند متغیره و همبسته جلب شده است. این گونه داده‌ها در مطالعات مربوط به اندام‌های زوجی بدن و همچنین مطالعات طولی به کرات دیده می‌شوند و از آن‌ها به‌عنوان اندازه‌های تکراری (repeated measurements) یاد می‌شود. این گونه اندازه‌گیری‌ها معمولاً روی فرد و یا واحد اندازه‌گیری در زمان‌های مختلف در طول مطالعه انجام می‌شوند. به هر فرد یا واحد نمونه‌گیری که برای وی در طول زمان چند اندازه‌گیری انجام شده است، یک خوشه (cluster) گفته می‌شود. بدیهی است مشاهدات حاصل از هر فرد یا خوشه که در واقع، اندازه‌های تکراری را تشکیل می‌دهند، با یکدیگر همبسته بوده، روش‌های معمول برای تحلیل داده‌های مستقل در مورد آن‌ها کارایی لازم را دارا نخواهند بود. این مطلب در مورد داده‌های به‌دست آمده از اندام‌های زوجی بدن نظیر چشم، گوش، دست، پا و ... نیز صادق است.

یکی از مهم‌ترین روش‌هایی که در سالیان اخیر برای تحلیل داده‌های چند متغیره و همبسته پیشنهاد شده، روش معادلات برآوردگر تعمیم یافته (GEE) است. این روش برای نخستین بار توسط لیانگ و زیگر [۱] معرفی شد و سپس توسط پرنتیس[۲] و ژائو و پرنتیس [۳] تعمیم یافت. دراین روش‌ها که به ترتیب به GEE1 و GEE2 مشهورند، اندازه‌های تکراری به‌صورت دوحالتی (۱و۰) در نظر گرفته می‌شوند، با این تفاوت که در روش GEE1 همبستگی بین مشاهدات به‌عنوان یک پارامتر مزاحم در نظر گرفته می‌شود، اما در GEE2 پارامتر ارتباط یا همبستگی به اندازه پارامترهای رگرسیونی مهم تلقی می‌شود و برآوردی از آن به‌دست می‌آید. مطالعات مروری مختلفی در مورد این دو روش صورت گرفته است. به‌عنوان نمونه، خواننده می‌تواند به مقالات لیانگ و همکاران او [۴]، زیگر و لیانگ[۵] و مارتوس [۶] مراجعه کند. همچنین لیپشتیز و همکارانش [۷] روشی مبتنی بر برآوردهای حداکثر درستنمایی برای داده‌های زوج شده دو حالتی پیشنهاد کرده‌اند. تعمیم روش‌های مبتنی بر برآورد حداکثر درستنمایی برای داده‌های رتبه‌ای همبسته به سادگی امکان‌پذیر نیست. مشکل عمده در این زمینه، تشکیل معادلات درستنمایی به‌صورتی است که همبستگی و رتبه‌ای بودن داده‌ها در آن لحاظ شده باشد. اگرستی[۸] در فصل یازدهم کتاب خود، بخش مربوط به برآورد حداکثر درستنمایی را با این جمله آغاز می‌کند:

ML fitting of marginal logit models is awkward.

البته، در این زمینه پیشنهادهای دیگری به‌خصوص در تحلیل داده‌های دو متغیره همبسته ارائه شده است، اما تعمیم این روش‌ها به حالت چند متغیره حداقل از نظر طراحی نرم‌افزار کامپیوتری بسیار پیچیده به نظر می‌رسد. برای درک بهتر علت پیچیده بودن تعمیم روش حداکثر درستنمایی برای داده‌های رتبه‌ای همبسته، خواننده می‌تواند به مقاله کیم[۹] مراجعه کند.

در این مقاله، هدف ما بررسی روش‌هایی مبتنی بر مشاهدات دوحالتی نیست، بلکه ما حالتی را در نظر می‌گیریم که در آن مشاهدات به‌صورت طبقه‌ای چند حالته (multinomial) باشند. این روش که در واقع تعمیمی از روش لیانگ و زیگر است و به‌وسیله آن می‌توان اندازه‌های تکراری همبسته را در حالت‌های دو و چند حالتی و حتی رتبه‌ای مورد تجزیه و تحلیل قرار داد، برای نخستین بار توسط لیپشتیز و همکاران او [۱۰] برای داده‌های طبقه‌ای تکراری معرفی شده است. هدف اصلی ما از ارائه این روش، معرفی روش‌هایی به مراتب ساده‌تر، هم از نظر ساختار و هم از بعد نرم‌افزاری، نسبت به روش‌هایی نظیر روش حداکثر درستنمایی است. ما علاوه بر معرفی این روش، چگونگی استفاده از آن را برای تحلیل داده‌های رتبه‌ای تشریح کرده، مثالی کاربردی از داده‌های چند متغیره، همبسته و رتبه‌ای دندان‌پزشکی ارائه خواهیم کرد.

۲. روش کار

از آن‌جا که روش معادلات برآوردگر تعمیم یافته بر روش شبه درستنمایی استوار است، پیش از ارائه آن

www.SID.ir

Archive of SID

به‌طور اجمالی، روش برآورد شبه درستنمایی را تشریح می‌کنیم. خواننده علاقه‌مند می‌تواند برای آگاهی از جزئیات این روش و روش حداقل مربعات تعمیم یافته به کتاب مدل‌های خطی تعمیم یافته نوشته مایرز و همکاران او [۱۱] مراجعه کند.

۲-۱. روش شبه درستنمایی

روش حداقل مربعات، تکنیکی معمول برای برآورد پارامترهای مدل‌های رگرسیونی است. در حالتی که مشاهدات پاسخ دارای واریانس‌های نامساوی باشند، می‌توان روش حداقل مربعات تعمیم یافته را جانشین روش حداقل مربعات معمولی کرد. فرض کنیم، y نشان‌دهنده بردار پاسخ باشد و $\mu = E(y)$. اگر $g(\mu) = X\beta$ مدلی به فرم GLM باشد و V یک ماتریس مثبت معین را نمایش دهد، آنگاه تابع حداقل مربعات تعمیم یافته را می‌توان به صورت زیر تشکیل داد:

$$u = (y-\mu)'V^{-1}(y-\mu)$$

با مشتق‌گیری از رابطه بالا، تابع امتیاز زیر نتیجه می‌شود:

$$D'V^{-1}(y-\mu) = 0$$

که در آن D ماتریسی از مشتقات به‌صورت $D = d\mu/d\beta$ است.

در روش شبه درستنمایی، به‌طور معمول فرض بر این است که مشاهدات پاسخ ناهمبسته است و در نتیجه V ماتریسی غیرقطری خواهد بود. با حل عددی تابع امتیاز بالا به کمک روش‌های تکراری مانند نیوتن-رافسون یا شبه نیوتن می‌توان برآورد پارامترهای مدل رگرسیونی را به‌دست آورد.

۲-۲. ساختار داده‌ها

فرض کنیم در یک مطالعه طولی، برای هر خوشه یا فرد مورد مطالعه T زمان اندازه‌گیری وجود داشته باشد. بنابراین فرد iام $(i = 1,...,N)$ در موقعیت‌های $t = 1,2,...,T_i$ ، $T_i \le T$ مشاهده می‌شود. برای ساده‌تر شدن اندیس‌ها فرض می‌کنیم $T_i = T$. از طرفی فرض کنیم متغیر پاسخ در هر زمان دارای $k = 1,...,L$ سطح باشد. بنابراین، متغیر پاسخ مربوط به فرد iام، در زمان t را می‌توان به‌صورت y_{it} نمایش داد که ممکن است مقادیر ۱ تا L را اختیار کند. حال برای پاسخ y_{it} می‌توان L متغیر تصادفی نشانگر y_{itk} را به‌صورت زیر تعریف کرد:

$$y_{itk} = \begin{cases} 1 & ; \quad y_{it} = k \\ 0 & ; \quad o.w \end{cases}$$

بنابراین برای فرد iام در زمان t برداری $(L-1)\times 1$ از پاسخ‌ها به‌صورت زیر در اختیار خواهیم داشت:

$$y_{itk} = \left[y_{it1}, y_{it2}, ..., y_{it,L-1}\right]'$$

لازم به توضیح است از آن‌جا که $y_{it+} = \sum_{K=1}^{L} y_{itk} = 1$، متغیر y_{itL} در بردار y_{it} وارد نمی‌شود. همچنین وقتی متغیر پاسخ در هر زمان دوحالتی باشد (یعنی L=۲)، y_{it} یک اسکالر خواهد بود و روش ذیل به معادلات برآوردگر تعمیم یافته معرفی شده توسط لیانگ و زیگر تبدیل می‌شود. هر فرد دارای T بردار متغیرهای کمکی به‌صورت x_{it} با بُعد P×۱ است. بنابراین ماتریس T×P به‌صورت $X_i = [x_{i1},...,x_{iT}]'$ ماتریسی از متغیرهای کمکی برای فرد iام خواهد بود.

۲-۳. معادلات برآوردگر تعمیم یافته

با تعاریف بالا، توزیع حاشیه‌ای بردار y_{it} چند جمله‌ای (با حجم نمونه ۱= y_{it+}) با بردار پارامتر $\pi_{it} = (\pi_{it1},...,\pi_{it,L-1})'$ خواهد بود. وقتی y_{it} دوحالتی باشد، π_{itk} معمولاً با یکی از توابع لجستیک، پروبیت یا لگ – لگ مکمل مدل‌سازی می‌شود. وقتی $L>2$ و طبقات متغیر پاسخ اسمی باشند، می‌توان از تابع لجستیک چند جمله‌ای برای مدل‌سازی استفاده کرد. در حالتی که طبقات به‌صورت رتبه‌ای باشند، دو انتخاب محتمل عبارت خواهد بود از لجستیک یا پروبیت تجمعی.

طبق معمول ما علاقه‌مند به استنباط در مورد پارامترهای β مرتبط با بردار $(L-1)\times 1$ بعدی زیر هستیم:

www.SID.ir

Archive of SID

$$E(y_{it}/X_i)=\pi_{it}(\beta)=[\pi_{it1},..,\pi_{it,L-1}]'$$

می‌توان این بردارهای احتمالات حاشیه‌ای را به‌صورت بردار $T(L-1)\times 1$ بعدی زیر نوشت:

$$E(y_{it}/X_i)=\pi_i=[\pi'_{i1},..,\pi'_{iT}]'$$

که در آن $y_i=[y'_{i1},...,y'_{iT}]'$.

حال برای برآورد β، معادلات برآوردگر تعمیم یافته را به‌صورت زیر تعریف می‌کنیم:

$$u(\hat{\beta})=\sum_{i=1}^{N}D_i'\hat{V}_i^{-1}[y_i-\hat{\pi}_i]=0 \qquad (۲-۱)$$

به‌طوری‌که:

$$D_i'=D_i(\beta)'=\frac{d[\pi_i(\beta)]'}{d\beta}$$

همچنین، V_i ماتریس کوواریانس عملی بردار متغیرهای پاسخ y_i است، یعنی:

$$V_i\approx \mathrm{var}(y_i)$$

ماتریس V_i علاوه بر این‌که تابعی از بردار ضرایب رگرسیونی β است، تابعی از پارامترهای مزاحم α نیز در نظر گرفته می‌شود، یعنی $V_i=V_i(\beta,\alpha)$. در واقع پارامترهای مزاحم α نشان‌دهنده همبستگی بین عناصر بردارهای y_{is} و y_{it} است. از طرفی، از آن‌جا که توزیع حاشیه‌ای y_{it} چندجمله‌ای است، بلوک‌های قطری $(L-1)\times(L-1)$ در ماتریس V_i، همان ماتریس‌های کوواریانس چند جمله‌ای هستند، یعنی:

$$V_{it}=V_{it}(\beta)=\mathrm{var}(y_{it})=Diag[\pi_{it}]-\pi_{it}\pi'_{it}$$

به‌طوری که $Diag[\pi_{it}]$ ماتریسی قطری با عناصر π_{it} روی قطر اصلی است. توجه داریم که V_{it} تابعی از β است نه α. بنابراین، تنها عناصر خارج از بلوک‌های واقع در قطر اصلی ماتریس V_i که نشان‌دهنده کوواریانس بین عناصر بردارهای y_{is} و y_{it} هستند، تابعی از پارامتر مزاحم α خواهند بود. در عمل به جای ماتریس کوواریانس V_i، از ماتریس همبستگی یعنی $Corr(y_i)$ درمعادلات برآوردگر تعمیم یافته استفاده می‌شود. برای تبدیل ماتریس V_i به $Corr(y_i)$ می‌توان از فرمول ساده زیر استفاده کرد:

$$V_i=\mathrm{var}(y_i)=A_i^{1/2}Corr(y_i)A_i^{1/2}$$

که در آن $A_i=Diag\{A_{i1},...,A_{iT}\}$ وعناصر A_{it} یک ماتریس قطری با واریانس‌های دوحالتی $\mathrm{var}(y_{itk})$ روی قطر اصلی هستند، یعنی:

$$A_{it}=Diag\{\pi_{it1}(1-\pi_{it1}),...,\pi_{it,L-1}(1-\pi_{it,L-1})\}$$

و $A_i^{1/2}$ ماتریسی است که عناصر آن ریشه دوم عناصر ماتریس A_i هستند.

۲- ٤. فرم کلی ماتریس همبستگی

دیدیم در معادلات برآوردگر تعمیم یافته (۲-۱) به جای ماتریس کوواریانس $V_i=\mathrm{var}(y_i)$ از ماتریس همبستگی $Corr(y_i)$ استفاده‌می‌شود. اگر ماتریس $R_i(\alpha)$ مدلی برای ماتریس همبستگی $Corr(y_i)$ باشد، داریم:

$$R_i(\alpha)=\begin{bmatrix} A_{i1}^{-1/2}V_{i1}A_{i1}^{-1/2} & \rho_{i12} & \cdots & \rho_{i1T} \\ \rho_{i21} & A_{i2}^{-1/2}V_{i2}A_{i2}^{-1/2} & \cdots & \rho_{i2T} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{iT1} & \rho_{iT2} & \cdots & A_{iT}^{-1/2}V_{iT}A_{iT}^{-1/2} \end{bmatrix}$$

که در آن tامین بلوک قطری $(L-1)\times(L-1)$ ماتریس $R_i(\alpha)$ برابر $A_{it}^{-1/2}V_{it}A_{it}^{-1/2}$ است و ماتریس غیرقطری $(L-1)\times(L-1)$ یعنی ρ_{ist} را می‌توان‌به‌صورت زیر نوشت:

$$\rho_{ist}(\alpha)=A_{is}^{-1/2}E[(y_{is}-\pi_{is})(y_{it}-\pi_{it})']A_{it}^{-1/2}$$

برای محاسبه ρ_{ist} در ماتریس فوق باید بردار پارامترهای α را برآورد کرد. در بخش بعد روش‌های پیشنهادی برای محاسبه برآوردهای α و β را مورد بحث قرار می‌دهیم.

۲- ۵. برآورد α، β و $V(\beta)$

با تعاریف بالا، برآورد ضرایب رگرسیونی β را می‌توان با الگورتیم امتیازدهی فیشر به‌صورت زیر به‌دست آورد:

$$\hat{\beta}^{(m+1)}=\hat{\beta}^{(m)}+\left[\sum_{i=1}^{N}D_i(\hat{\beta}^{(m)})'[V_i(\hat{\beta}^{(m)},\hat{\alpha}^{(m)})]\right]^{-1}[D_i(\hat{\beta}^{(m)})]^{-1}$$

$$\sum_{i=1}^{N}D_i(\hat{\beta}^{(m)})'[V_i(\hat{\beta}^{(m)},\hat{\alpha}^{(m)})]^{-1}[y_i-\pi_i(\hat{\beta}^{(m)})]$$

و تکرارها تا زمانی ادامه می‌یابند که $\hat{\beta}^{(m+1)}=\hat{\beta}^{(m)}$ و $\hat{\alpha}^{(m+1)}=\hat{\alpha}^{(m)}$.

www.SID.ir

Archive of SID

برای محاسبه $\hat{\alpha}^{(m)}$، یعنی برآورد α در mامین گام، ابتدا باقیمانده y_{itk} در زمان t را از فرمول زیر محاسبه می‌کنیم:

$$e_{itk}=\frac{y_{itk}-\pi_{itk}}{\{\pi_{itk}(1-\pi_{itk})\}^{1/2}}$$

در حالت کلی، عناصر e_{itk}، برداری به‌صورت e_{it} تشکیل می‌دهند. این بردار را می‌توان به فرم ماتریس زیر نوشت:

$$e_{it}=A_{it}^{-1/2}[y_{it}-\pi_{it}]$$

به کمک بردار باقیمانده‌های e_{it} می‌توان روش‌های مختلفی برای برآورد پارامتر α پیشنهاد کرد. برخی از مشهورترین این روش‌ها برای برآورد پارامتر α در mامین گام از روش تکراری عبارتند از:

۱- روش تبادل *(exchangeable)*: در این حالت، فرض براین است که برای تمامی زوج‌های is و it یک α ثابت وجود دارد. برآورد α را می‌توان به‌صورت زیر در هر گام محاسبه کرد:

$$\hat{\alpha}=\rho_{ist}(\hat{\alpha})=\frac{\sum_{i=1}^{N}\sum_{t>s}\hat{e}_{is}\hat{e}'_{it}}{[\sum_{i=1}^{N}\frac{1}{2}T(T-1)]-P}$$

به‌طوری‌که $\hat{e}_{it}=\hat{A}_{it}^{-1/2}[y_{it}-\hat{\pi}_{it}]$، با قرار دادن $\hat{\beta}$ از گام پیشین الگوریتم امتیازدهی فیشر در A_{it} و π_{it} برآورد می‌شود.

۲- روش *1*- همبستگی *(1-dependence)* که در آن هر مشاهده در زمان t، تنها با مشاهده زمان بعدی همبستگی دارد و سایر همبستگی‌ها برابر صفر فرض می‌شوند. با نماد ریاضی داریم: $\rho_{it,t+1}=\alpha_t$ برای $t=1,...,T\text{-}1$ و برای سایر مقادیر $\rho_{ist}=0$. این فرض منجر به برآورد زیر برای α می‌شود:

$$\hat{a}_t=\frac{\sum_{i=1}^{N}\hat{e}_{it}\hat{e}'_{i,t+1}}{N-P}$$

۳- روش متحد *(Banded)*: اگر همبستگی بین مشاهدات به فاصله زمانی بین آن‌ها وابسته باشد، یعنی $\rho_{ist}=\alpha_u$ به‌طوری‌که $|t-s|=u$، آنگاه:

$$\hat{\alpha}_u=\frac{\sum_{i=1}^{N}\sum_{S=1}^{T-u}\hat{e}_{is}\hat{e}'_{i,s+u}}{N(T-u)-P}$$

برای $u=1,...,T\text{-}1$.

۴- روش بی ساختار *(Unstructured)*: دراین حالت، الگوی مشخص برای توصیف همبستگی بین مشاهدات وجود ندارد، یعنی $\rho_{ist}=\alpha_{st}$، و در نتیجه:

$$\hat{a}_{st}=\frac{\sum_{i=1}^{N}\hat{e}_{is}\hat{e}'_{it}}{N-P}$$

بنابراین در حالت کلی، برای برآورد α و β می‌توان کار را با یک حدس ابتدایی برای β یعنی $\hat{\beta}^{(1)}$ آغاز و سپس با یکی از روش‌های یاد شده مقدار $\hat{\alpha}^{(1)}$ را محاسبه کرد. این عمل تا همگرایی برآوردهای $\hat{\alpha}^{(m)}$ و $\hat{\beta}^{(m)}$ ادامه پیدا می‌کند.

یک برآورد ساده (naive) برای واریانس بردار ضریب رگرسیونی β را می‌توان به‌صورت زیر به‌دست آورد:

$$\hat{V}(\beta)=\lim_{N\to\infty} N[\sum_{i=1}^{N}\hat{D}_i'\hat{V}_i^{-1}\hat{D}_i]^{-1}$$

که در آن D_i' با جایگذاری برآوردهای α و β به جای مقادیر واقعی محاسبه می‌گردد، همچنین:

$$\hat{V}_i=(y_i-\hat{\pi}_i)(y_i-\hat{\pi}_i)'$$

۳. نتایج حاصل از یک مثال کاربردی

به‌عنوان یک مثال کاربردی، داده‌های مربوط به وضعیت پریودنتال یک نمونه تصادفی به حجم ۸۶۷ دانش‌آموز دبیرستانی ۱۹-۱۵ ساله در شهر تهران را به کمک روش یاد شده مورد تجزیه و تحلیل قرار دادیم. برای انتخاب این نمونه، به کمک روش نمونه گیری چند مرحله‌ای، ابتدا شش منطقه از مناطق شهر تهران به تصادف انتخاب و سپس در هر منطقه دو دبیرستان - یکی پسرانه و دیگری دخترانه - به‌طور تصادفی برگزیده شد. در هر یک از این مدارس، تعدادی دانش‌آموز از روی فهرست کل دانش‌آموزان به تصادف انتخاب و اطلاعات مورد نیاز به کمک یک پرسشنامه و معاینات پزشکی جمع‌آوری گردید.

www.SID.ir

Archive of SID

وضعیت پریودنتال افراد تحت مطالعه با محاسبه شاخص CPTIN در شش ناحیه از فک‌های بالا و پایین هر دانش‌آموز مشخص شد. این شاخص دارای یک مقیاس رتبه‌ای ۵ سطحی به‌صورت زیر است:

۰= سالم، ۱= خونریزی، ۲ = جرم سخت، ۳ = پاکت با عمق ۶-۴ میلی‌متر، ۴= پاکت با عمق بیش از ۶ میلی‌متر. در فرایند مدل‌سازی به‌ترتیب کدهای ۱ تا ۵ را برای متغیر پاسخ CPTIN در نظر می‌گیریم.

با این تعاریف، برای هر دانش‌آموز، شش متغیر پاسخ به‌صورت پنج حالته رتبه‌ای در اختیار خواهیم داشت. بدیهی است، پاسخ‌های مربوط به هر دانش‌آموز با یکدیگر همبستگی نسبتاً بالایی خواهند داشت و در نتیجه مدل ارائه شده باید قابلیت تحلیل پاسخ‌های چند متغیره، رتبه‌ای و همبسته را دارا باشد.

جدول۱ مقادیر شاخص CPITN را برای‌دانش‌آموزان نمونه نشان می‌دهد. به کمک این جدول می‌توان شیوع و شدت بیماری‌های پریودنتال را در بین دانش‌آموزان ۱۹-۱۵ ساله تهران محاسبه کرد. عواملی که در این مطالعه به‌عنوان ریسک فاکتور یا متغیرهای کمکی در نظر گرفته شدند، عبارتند از: جنس (۱= مرد، ۲= زن)، میزان تحصیلات والدین (۱= تحصیلات پایین، ۲= تحصیلات متوسط، ۳= تحصیلات بالا)، مسواک زدن (۱= هرگز، ۲= به‌طور نامرتب، ۳= یکبار در روز، ۴= بیش از یکبار در روز)، استفاده از نخ دندان (۱= هرگز، ۲= به‌طور نامرتب، ۳= یکبار در روز، ۴= بیش از یکبار در روز) و ویزیت توسط دندان‌پزشک (۱= به‌طور مرتب‌به‌منظورپیشگیری، ۲=هرگز یا در مواقع‌اضطراری).

برای تحلیل داده‌ها، از یک مدل رگرسیون بخت‌های متناسب چند متغیره حاشیه‌ای استفاده کردیم. خواننده برای به‌دست آوردن اطلاعات پایه در مورد حالت یک متغیره این مدل می‌تواند به مک کلاف [۱۲] یا اگرستی [۸] مراجعه کند. همچنین به علت ماهیت رتبه‌ای متغیرهای پاسخ، یک تابع پیوند لجستیک تجمعی برای برقراری ارتباط بین میانگین پاسخ‌های حاشیه‌ای و متغیرهای کمکی مورد استفاده قرار گرفت.

در فرایند مدل‌سازی، ابتدا مدلی با آثار اصلی و همچنین آثار متقابل مرتبه دوم متغیرهای کمکی به کار برده شد، اما هیچ‌یک از آثار متقابل جواز ورود به مدل را کسب نکردند (مقادیر P.value برای تمامی این آثار متقابل بیش از ۰/۴ بود). بنابراین در مرحله دوم، ما از مدلی که تنها آثار اصلی متغیرهای کمکی یاد شده را شامل می‌شد، استفاده کردیم. با این تفاسیر، مدل نهایی را می‌توان به صورت زیر نوشت:

$$Log(\frac{F'_{itk}}{1-F'_{itk}}) = \theta_k + \beta_1 sex_i + \beta_2 father's\_education1_i + \beta_3 father's\_education2_i + \beta_4 mother's\_education1_i + \beta_5 mother's\_education2_i + \beta_6 brush1_i + \beta_7 brush2_i + \beta_8 brush3_i + \beta_9 floss1_i + \beta_{10} floss2_i + \beta_{11} floss3_i + \beta_{12} visit_i,$$

که در آن F'_{itk} احتمال قرارگرفتن مقدار CPITN در سطوح K $(K=۱،...،۴)$ یا بالاتر (بدتر) در اندازه‌گیری tام $(t=۱،...،۶)$ مربوط به فرد iام $(i=۱،...،۸۹۷)$ است. نتایج حاصل از برازش این مدل در جدول ۲ گزارش شده است. لازم به تذکر است برای تمامی متغیرهای کمکی، طبقه آخر (وضعیت بهتر) به‌عنوان طبقه مرجع (reference) در نظر گرفته شده است.

مقادیر برآوردها در جدول ۲ و احتمال‌های مربوط نشان می‌دهد که بجز سطح سوم متغیرهای استفاده از مسواک و نخ دندان، سایر متغیرها دارای تأثیری معنادار بر وضعیت پریودنتال دانش‌آموزان بوده‌اند. به عبارت دقیق‌تر، دانش‌آموزان دختر، دانش‌آموزانی که والدین آن‌ها تحصیلات پایینی دارند، دانش‌آموزانی که هرگز از مسواک و نخ دندان استفاده نمی‌کنند یا این‌که از مسواک و نخ دندان به‌طور نامرتب استفاده می‌کنند و همچنین دانش‌آموزانی که تنها در مواقع اضطراری به دندان‌پزشک مراجعه می‌کنند، بخت بیش‌تری برای ابتلا به بیماری‌های پریودنتال دارند.

برای تفسیر برآوردهای جدول ۲ می‌توان از نسبت بخت‌های گزارش شده در ستون آخر استفاده کرد. به‌عنوان مثال در این جدول، نسبت بخت‌ها معادل ۱/۸۳ برای متغیر کمکی جنس نشان می‌دهد که بخت دانش‌آموزان دختر برای داشتن شاخص CPITN سطح K

www.SID.ir

Archive of SID

جدول ۱. شمای کلی وضعیت پریودنتال دانش‌آموزان ۱۹ - ۱۵ ساله تهرانی

CPITN	ناحیه فک					
	بالا راست	بالا جلو	بالا چپ	پایین راست	پایین جلو	پایین چپ
۰	۳۶۸	۵۹۸	۴۳۸	۴۵۸	۳۴۷	۴۴۵
۱	۱۲۲	۱۵۲	۱۰۴	۱۴۷	۱۲۱	۱۵۳
۲	۲۵۳	۷۰	۲۵۷	۱۵۹	۳۳۲	۱۷۱
۳	۱۲۲	۴۷	۱۵۸	۱۰۳	۶۶	۹۸
۴	۲	۰	۰	۰	۱	۰

جدول ۲. نتایج برازش GEE برای داده های مطالعه اپیدمیولوژیک وضعیت پریودنتال دانش‌آموزان تهرانی

پارامتر	برآورد	SE	Z	P	نسبت بخت‌ها
θ_1	۳،۴۹۹۷	۰،۷۲۹۴	۴،۸۰	<۰،۰۰۰۱	_
θ_2	۴،۱۹۱۹	۰،۷۲۷۹	۵،۷۶	<۰،۰۰۰۱	_
θ_3	۵،۷۴۱۵	۰،۷۳۱۳	۷،۸۵	<۰،۰۰۰۱	_
θ_4	۱۱،۲۳۷۱	۰،۹۱۸۹	۱۲،۲۳	<۰،۰۰۰۱	_
β_1/ جنس(زن)	۰،۶۰۴۲	۰،۰۸۳۷	۷،۲۲	<۰،۰۰۰۱	۱،۸۳
جنس(مرد)	طبقه مرجع				
β_2/ تحصیلات پدر۱	۰،۷۹۲۱	۰،۲۵۱۳	۳،۱۵	۰،۰۰۱۶	۴،۸۸
β_3/ تحصیلات پدر۲	۰،۳۷۷۱	۰،۱۶۳۹	۲،۳۰	۰،۰۲۱۴	۱،۴۶
تحصیلات پدر۳	طبقه مرجع				
B_4/ تحصیلات مادر۱	۰،۸۲۳۳	۰،۲۲۴۶	۳،۶۷	۰،۰۰۰۲	۵،۱۹
B_5/ تحصیلات مادر۲	۰،۶۳۸۱	۰،۲۹۲۶	۲،۱۸	۰،۰۲۹۲	۱،۸۹
تحصیلات مادر۳	طبقه مرجع				
β_6/ مسواک ۱	۰،۶۵۰۱	۰،۲۵۳۲	۲،۵۷	۰،۰۱۰۲	۷،۰۳
β_7/ مسواک ۲	۰،۳۱۰۱	۰،۱۲۱۲	۲،۵۶	۰،۰۱۰۵	۱،۸۶
β_8/ مسواک ۳	۰،۰۶۱۲	۰،۱۰۰۳	۰،۶۱	۰،۵۴۱۷	۱،۰۶
مسواک ۴	طبقه مرجع				
β_9/ نخ دندان ۱	۰،۷۴۱۵	۰،۳۰۱۶	۲،۴۶	۰،۰۱۳۹	۹،۲۵
β_{10}/ نخ دندان ۲	۰،۷۱۶۴	۰،۳۲۵۷	۲،۲۰	۰،۰۲۷۸	۴،۱۹
β_{11}/ نخ دندان ۳	۰،۱۱۲۵	۰،۰۹۵۲	۱،۱۸	۰،۲۳۷۳	۱،۱۲
نخ دندان ۴	طبقه مرجع				
β_{12}/ ویزیت ۱	۰،۶۰۳۹	۰،۱۱۷۲	۵،۱۵	<۰،۰۰۰۱	۱،۸۳
ویزیت ۲	طبقه مرجع				
آماره نسبت درستنمایی	۶۰۶۹،۱۴۶۹				

www.SID.ir

Archive of SID

جدول ۳: نتایج برازش MLE برای داده های مطالعه اپیدمیولوژیک وضعیت پریودنتال دانش‌آموزان تهرانی

پارامتر	برآورد	SE	Z	P	نسبت بخت‌ها
θ_1	۳،۰۱۵۶	۰،۶۶۳۲	۴،۵۵	<۰،۰۰۰۱	_
θ_2	۳،۹۲۵۵	۰،۶۷۴۱	۵،۸۲	<۰،۰۰۰۱	_
θ_3	۵،۶۶۱۳	۰،۶۹۸۸	۸،۱۰	<۰،۰۰۰۱	_
θ_4	۱۰،۶۳۸۴	۰،۸۵۲۶	۱۲،۴۸	<۰،۰۰۰۱	_
β_1/ جنس(زن)	۰،۳۰۲۱	۰،۰۴۸۶	۶،۲۲	<۰،۰۰۰۱	۱،۳۵
جنس(مرد)	طبقه مرجع				
β_2/ تحصیلات پدر۱	۰،۴۰۲۳	۰،۱۲۲۹	۳،۲۷	۰،۰۰۱۱	۲،۲۴
β_3/ تحصیلات پدر۲	۰،۱۲۶۰	۰،۰۵۵۲	۲،۲۸	۰،۰۲۲۵	۱،۱۳
تحصیلات پدر۳	طبقه مرجع				
β_4/ تحصیلات مادر۱	۰،۷۱۲۴	۰،۱۸۰۵	۳،۹۵	۰،۰۰۰۱	۴،۱۶
β_5/ تحصیلات مادر۲	۰،۴۳۷۲	۰،۱۶۵۴	۲،۶۴	۰،۰۰۸۳	۱،۵۵
تحصیلات مادر۳	طبقه مرجع				
β_6/ مسواک ۱	۰،۵۹۴۸	۰،۲۲۳۱	۲،۶۷	۰،۰۰۷۶	۵،۹۶
β_7/ مسواک ۲	۰،۲۶۴۱	۰،۱۰۳۲	۲،۵۶	۰،۰۱۰۵	۱،۷۰
β_8/ مسواک ۳	۰،۰۵۴۷	۰،۰۵۸۷	۰،۹۳	۰،۳۵۲۴	۱،۰۶
مسواک ۴	طبقه مرجع				
β_9/ نخ دندان ۱	۰،۷۶۱۷	۰،۳۲۵۶	۲،۳۳	۰،۰۱۹۸	۹،۸۳
β_{10}/ نخ دندان ۲	۰،۷۰۶۶	۰،۳۲۷۷	۲،۱۶	۰،۰۳۰۸	۴،۱۱
β_{11}/ نخ دندان ۳	۰،۰۶۱۹	۰،۰۵۳۷	۱،۱۵	۰،۲۵۰۱	۱،۰۶
نخ دندان ۴	طبقه مرجع				
β_{12}/ ویزیت ۱	۰،۳۰۱۹	۰،۰۷۱۵	۴،۲۲	<۰،۰۰۰۱	۱،۳۵
ویزیت ۲	طبقه مرجع				
آماره نسبت درستنمایی	۶۰۸۰،۰۴۳۸				

($K=1,...,4$) یا بدتر، ۱/۸۳ برابر همین بخت برای دانش‌آموزان پسر بوده است. همچنین اولین نسبت بخت‌ها برای متغیر مسواک زدن نشان می‌دهد که بخت قرار گرفتن دانش‌آموزانی که هرگز مسواک نمی‌زنند در طبقه K شاخص CPITN یا بدتر، ۷/۰۳ برابر همین بخت برای دانش‌آموزانی است که در روز ۲ بار یا بیش‌تر مسواک می‌زنند. تفسیر سایر برآوردها به صورت مشابه امکان‌پذیر است.

بجز روش GEE، پارامترهای مدل توصیف شده را به روش حداکثر درستنمایی نیز می‌توان برآورد کرد. از آن‌جا که جزئیات این روش به‌طور کامل در فصل یازدهم کتاب اگرستی [۸] تشریح شده، از ذکر مجدد آن در این مقاله خودداری می‌کنیم. جدول ۳ برآوردهای حداکثر درستنمایی پارامترهای مدل را نمایش می‌دهد. علی‌رغم تفاوت‌های موجود در مقادیر برآوردی و خطاهای استاندارد پارامترهای مدل با نتایج حاصل از روش GEE، مقادیر p.value درج شده در این جدول، نشان از یکسان بودن عوامل خطر معنادار در دو روش دارد. همچنین با مقایسه مقادیر آماره نسبت درستنمایی دو مدل می‌توان نتیجه گرفت که برای این داده‌ها روش GEE برازشی، نسبت‌به روش حداکثر درستنمایی، بهتر نتیجه داده است.

www.SID.ir

Archive of SID

٤. بحث و نتیجه‌گیری نهایی

هرچند روش ارائه شده دراین مقاله پیچیده‌تر از روش‌های کلاسیک تحلیل داده‌های رتبه‌ای است، اما از کارایی بسیار بالاتری نسبت به این روش‌ها در نشان دادن روابط بین متغیرهای کمکی و متغیر پاسخ برخوردار است. در روش‌های تجزیه و تحلیل کلاسیک، نظیر آزمون کای دو، ناچاریم این پاسخ شش متغیره را به شکل یک متغیره تبدیل کرده، رابطه متغیرهای کمکی را به صورت یک به یک با این پاسخ بررسی کنیم. همچنین در این روش‌ها، ماهیت رتبه‌ای داده‌ها در نظر گرفته نمی‌شود. ما پیش از تحلیل GEE، داده‌ها را با روش‌های کلاسیک، مانند آزمون کای دو یا آنالیز واریانس مورد تجزیه و تحلیل قرار دادیم، اما بجز متغیر جنس، هیچ‌یک از متغیرهای کمکی، رابطه معناداری با متغیر پاسخ نشان ندادند، چرا که با تبدیل این پاسخ شش متغیره به حالت یک متغیره، اطلاعات بسیار زیادی را از دست می‌دهیم و در تجزیه و تحلیل وارد نمی‌کنیم.

در میان نرم‌افزارهای آماری موجود، نرم‌افزار SAS قادر به انجام برخی از حالت‌های خاص روش GEE است. به کمک این نرم‌افزار می‌توان پاسخ‌های دوحالتی تکراری را با چندین ساختار همبستگی متفاوت تجزیه و تحلیل کرد؛ اما روش ارائه شده دراین مقاله با این نرم‌افزار قابل انجام نیست.

برای برازش این مدل، با اِعمال تغییراتی از برنامه نوشته شده توسط ویلیامسون و همکاران [۱۳] استفاده کردیم. برنامه کامپیوتری دیگری به کمک نرم‌افزار فورترن برای تحلیل این گونه داده‌ها طراحی شده است[۱۴].

در مورد نتایج دندان‌پزشکی به‌دست آمده در این مقاله می‌توان به نکات زیر اشاره کرد:

۱. علت این‌که دانش‌آموزان دختر بیش از پسران در معرض دچار شدن به بیماری‌های پریودنتال بوده‌اند را می‌توان به تغییرات هورمونی دختران در سنین مورد مطالعه نسبت داد، چرا که این تغییرات هورمونی ممکن است سبب بروز ناهنجاری‌هایی در دهان و دندان دانش‌آموزان دختر شده باشد. از طرفی، در مطالعاتی که در گروه‌های سنی بالاتر انجام گرفته، خلاف این مطلب ادعا شده است [۱۵-۱۶].

۲. سن به‌عنوان یک عامل خطر مهم در بیماری‌های پریودنتال شناخته شده است[۱۷]؛ اما در این مطالعه به علت نزدیک بودن سن دانش‌آموزان به هم، رابطه معناداری بین سن دانش‌آموزان و ابتلای آن‌ها به این گونه بیماری‌ها به‌دست نیامد.

۳. همانند قریب به اتفاق دیگر مطالعات دندان‌پزشکی، مطالعه حاضر تأثیر مهم استفاده از مسواک و نخ دندان (حداقل یک‌بار در روز) و همچنین مراجعه مکرر به دندان‌پزشک جهت انجام معاینات پیشگیرانه را در کاهش خطر ابتلا به بیماری‌های پریودنتال نشان می‌دهد.

در این‌جا لازم می‌بینیم در مورد تصمیم‌گیری برای انتخاب یکی از روش‌های برآورد پارامترهای مدل توضیح مختصری ارائه کنیم. هنگامی که داده‌های مورد مطالعه از نوع رتبه‌ای و مستقل هستند، به علت در دسترس بودن نرم‌افزارهای مختلف، استفاده از روش برآورد حداکثر درستنمایی معمولاً ساده تر بوده، برازش بهتری به دست می‌دهد. اما وقتی داده‌های مورد مطالعه به‌صورت رتبه‌ای و همبسته هستند، روش حداکثر درستنمایی، چه از نظر تئوری و چه از نظر عملی (نرم‌افزارهای مورد نیاز) بسیار دشوار است. حتی با استفاده از برنامه‌های کامپیوتری، مثلاً در محیط نرم‌افزار S-PLUS نیز انجام این روش مستلزم صرف چندین ساعت زمان و استفاده از حدس‌های اولیه دقیق است، چرا که انتخاب حدس‌های اولیه دور از مقادیر نهایی، اغلب موجب عدم همگرایی برآوردها می‌شود. در عوض، هنگامی که داده‌ها همبسته هستند، روش GEE از نظر تئوری، روشی به مراتب ساده‌تر بوده، نرم‌افزارهای مختلفی قادر به انجام آن هستند. حتی وقتی داده‌هایی با حجم بالا در اختیار داریم، انجام

www.SID.ir

Archive of SID

روش GEE در عرض چند ثانیه امکان‌پذیر است و نتایج کلی معمولاً تفاوت قابل ملاحظه‌ای با روش حداکثر درستنمایی ندارند.

در پایان ذکر این نکته ضروری است که برازش مناسب مدل به روش GEE به انتخاب دقیق ماتریس V_i بستگی دارد. از آن‌جا که $V_i \approx \mathrm{var}(y_i)$، باید سعی کنیم ماتریس V_i را نزدیک به ماتریس $\mathrm{var}(y_i)$ انتخاب کنیم تا برآوردهای کاراتری برای β به‌دست آوریم. با این حال، برآوردهای GEE فارغ از انتخاب V_i، همواره برآوردی‌سازگار برای β خواهند بود[اثبات‌در مرجع۱].

منابع

1. Liang KY, Zeger SL: Longitudinal data analysis using generalized linear models. Biometrika; 1986; 73: 13-22.
2. Prentice RL: Correlated binary regression with covariates specific to each binary observation. Biometrics, 1988; 44: 1033-48.
3. Zhao L, Prentice RL: Correlated binary regression using quadratic exponential model. Biometrika, 1990; 77: 642-8.
4. Liang KY, Zeger SL, Qaqish B: Multivariate regression analyses for categorical data. Journal of Royal Statistical Society, Series B, 1992; 54: 3-40.
5. Zeger SL, Liang KY: An overview of methods for the analysis of longitudinal data. Statistics in Medicine, 1992; 11: 1825-39.
6. Martus P: Statistical methods for the evaluation of diagnostic measurements concerning paired organs. Statistics in Medicine, 2000; 19: 525-40.
7. Lipsitz SR, Liard NM: Maximum likelihood regression methods for paired binary data. Statistics in Medicine, 1990; 9: 1517-25.
8. Agresti A: Categorical data analysis. 2nd Edition, USA, Wiley & Sons, 2002.
9. Kim K: A bivariate cumulative probit regression model for ordered categorical data. Statistics in Medicine, 1995; 14: 1341-52.
10. Lipsitz SR, Kim K, Zhao L: Analysis of repeated categorical data using generalized estimating equations. Statistics in Medicine, 1994; 13: 1149-63.
11. Myers RH, Montgomery DC, Vining GG: Generalized linear models. USA, Wiley & Sons, 2002.
12. McCullagh P: Regression models for ordinal data (with discussion). Journal of Royal Statistical Society, Series B, 1980; 42: 109-42.
13. Williamson JM, Lipsitz SR, Kim K: GEECAT and GEEGOR: computer programs for the analysis of correlated categorical response data. Computer Methods and Programs in Biomedicine, 1999; 58:25-34.
14. Davis CS, Hall DB: A computer program for regression analysis of ordered categorical repeated measurements. Computer Methods and Programs in Biomedicine, 1996; 51: 153-69.
15. Anagnou VA, Diamanti KA, Afentoulidis N: A clinical survey of periodontal conditions in Greece. Journal of clinical period ontology, 1996; 23: 758-63.
16. Corbet EF, Holmgren CJ, Lim LP, Davies WI: Sex differences in the periodontal status of Hong Kong adults aged 35-44 years. Community Dental Health, 1989; 6: 23-30.
17. Frencken JE, Sithole WD, Mwaenga R, Htoon HM, Simon E: National oral health survey Zimbabwe 1995: periodontal conditions. International Dentistry Journal, 1999; 49: 10-14.

www.SID.ir