چهل و ششمین کنفرانس ریاضی ایران
۳–۶ شهریور ۱۳۹۴
دانشگاه یزد

پوستر | تحلیل کواریانس با فرض چوله نرمال بودن توزیع متغیر پاسخ | ص: ۴–۱

تحلیل کواریانس با فرض چوله نرمال بودن توزیع متغیر پاسخ

افشین فلاح* زهرا گودرزی
دانشگاه بین المللی امام خمینی قزوین دانشگاه بین المللی امام خمینی قزوین

چکیده

در این مقاله تحلیل کواریانس تحت فرض چوله نرمال بودن توزیع متغیر پاسخ مورد توجه قرار گرفته و برای این منظور مدلی پیشنهاد شده است. پارامترهای مدل پیشنهادی با استفاده از روش ماکسیمم درستنمایی، یک الگوریتم EM توسعه داده شده است. مدل پیشنهادی در قالب یک مطالعه شبیه سازی مورد ارزیابی قرار گرفته است.

واژه‌های کلیدی: تحلیل کواریانس، توزیع چوله نرمال، تابع درستنمایی داده‌های کامل، متغیر تبیینی
Mathematics Subject Classification [2010]: 62H25,62F12

۱ مقدمه

تعیین رابطه بین یک متغیر تصادفی پاسخ و مجموعه‌ای از متغیرهای تبیینی یا تیمارها همواره در مطالعات مختلف مورد توجه قرار گرفته است. برای این منظور، می‌توان از روش‌های مختلف تحلیل رگرسیونی و تحلیل واریانس ($ANOVA$) استفاده کرد. تحلیل کواریانس، تلفیقی از تحلیل رگرسیونی و تحلیل واریانس است، که در آن علاوه بر متغیرهای تبیینی، یک یا چند متغیر کیفی نیز در مدل وجود دارد. ایده‌ی اولیه تحلیل کواریانس توسط [۵] طرح شده است. [۴] مفاهیم و کاربردهای مختلف تحلیل کواریانس را مورد بحث قرار داده است. نظریه سنتی تحلیل کواریانس همانند تحلیل رگرسیونی و تحلیل واریانس، عمدتاً بر فرض نرمال بودن متغیر پاسخ بنا شده است. در مواردی که داده‌ها نرمال نباشند، از تبدیل‌هایی برای فراهم آوردن شرایط مورد نظر استفاده می‌شود. انتخاب تبدیل مناسب همواره مسأله‌ای چالش برانگیز است، زیرا در بسیاری موارد استفاده از تبدیل‌ها موجب پیچیده‌تر شدن مدل و تغییر ساختار تابعی آن می‌شود. راهکار دیگری که می‌توان در چنین شرایطی در پیش گرفت، برازش مدل تحلیل کواریانس تحت فرض توزیع‌های دیگری است که بتوانند به خوبی عدم تقارن مشاهدات را در مدل‌سازی لحاظ نماید. در شرایطی که مشاهدات متقارن نیستند، جایگزین نمودن یک توزیع چوله به جای توزیع نرمال به عنوان توزیع متغیر پاسخ، می‌تواند به بهبود قابل توجه در کارایی مدل تحلیل کواریانس منجر شود. یکی از توزیع‌هایی که لزوماً متقارن نیست اما خواصی مشابه توزیع نرمال دارد، به توزیع چوله‌نرمال معروف است. این توزیع دارای پارامتری برای مدل‌بندی چولگی است و توزیع نرمال را به عنوان یک عضو شامل می‌شود. اگر چه شکل اولیه توزیع چوله‌نرمال، توسط [۶] ارائه شد، اما تعریف رایج این خانواده برای نخستین بار توسط [۲] ارائه شد. در طول بیش از سه دهه گذشته توزیع‌های چوله متقارن و چوله نرمال متعددی توسط پژوهشگران مختلف معرفی و مورد مطالعه قرار گرفته است. در این مقاله، برای تحلیل کواریانس تحت فرض چوله نرمال بودن توزیع متغیر پاسخ مورد توجه قرار گرفته است. برای محاسبه براوردگرهای ماکسیمم درستنمایی پارامترهای مدل پیشنهادی یک الگوریتم EM توسعه داده شده است. به منظور ارزیابی براوردگرهای پیشنهادی از یک مطالعه شبیه‌سازی استفاده است.

*سخنران

۲ مدل پیشنهادی

معمولاً رابطه بین میانگین متغیر پاسخ و مجموعه‌ای از تیمارها و متغیرهای تبیینی، از طریق ترکیب روش‌های مختلف تحلیل رگرسیونی و تحلیل واریانس مورد بررسی قرار می‌گیرد. فرض کنید مدل تحلیل واریانس و رابطه رگرسیونی به ترتیب از طریق ساختار ماتریس طرح $\boldsymbol{X}$ و ماتریس متغیرهای تبیینی $\boldsymbol{Z}$ مشخص می‌شوند. در این صورت، مدل تحلیل کواریانس را می‌توان به شکل $\boldsymbol{\mu} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} = \boldsymbol{W\theta}$ نوشت، که در آن $\boldsymbol{\mu}$ بردار میانگین مشاهدات پاسخ، $\boldsymbol{\beta} = (\beta_۰, \beta_۱, ..., \beta_{s-۱})'$ و $\boldsymbol{\gamma} = (\gamma_۱, ..., \gamma_q)$ به ترتیب نشان‌دهنده‌ی بردارهای تیمارها و ضرایب رگرسیونی و $\boldsymbol{W} = [\boldsymbol{X}|\boldsymbol{Z}]$ ماتریس مشاهدات را نشان می‌دهد. هدف اصلی، براورد ضرایب مدل تحلیل کواریانس است. $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})'$ بردار با استفاده از مشاهدات $\boldsymbol{y} = (y_{۱۱}, ..., y_{sr})'$ و $\boldsymbol{W}$ است. در این بخش تحلیل کواریانس تحت فرض چوله نرمال بودن توزیع متغیر پاسخ مورد توجه قرار گرفته است. متغیر تصادفی Y دارای توزیع چوله نرمال ساهو است ($Y \sim SSN(\mu, \sigma^۲, \lambda)$)، هرگاه تابع چگالی آن به‌صورت

$$f_Y(y) = ۲\phi(y|\mu, \sigma^۲ + \lambda^۲)\Phi\Big(\frac{\lambda}{\sigma}\frac{(y-\mu)}{(\sigma^۲+\lambda^۲)^{\frac{۱}{۲}}}\Big)$$

باشد، که درآن $\phi(.|\mu, \sigma^۲)$ و $\Phi(.|\mu, \sigma^۲)$ به ترتیب تابع چگالی و تابع توزیع تجمعی توزیع نرمال با میانگین μ و واریانس $\sigma^۲$ هستند [۷]. این توابع به ازای $(\mu, \sigma) = (۰, ۱)$ به ترتیب با نمادهای $\phi(\cdot)$ و $\Phi(\cdot)$ نشان داده می‌شوند. این توزیع به ازای $\lambda < ۰$ دارای چولگی منفی، به ازای $\lambda > ۰$ دارای چولگی مثبت و به ازای $\lambda = ۰$ متقارن بوده و به توزیع نرمال تبدیل می‌شود. مدل تحلیل کواریانس تحت فرض چوله نرمال بودن متغیر پاسخ را به‌صورت

$$Y_{ij}|\boldsymbol{w}_{ij} \sim SSN(\boldsymbol{w}_{ij}\boldsymbol{\theta} - \sqrt{\frac{۲}{\pi}}\lambda, \sigma^۲, \lambda) \quad i = ۱, ..., s;\ j = ۱, ..., r, \qquad (۱)$$

در نظر بگیرید. تابع درستنمایی متناظر با مدل (۱) به‌صورت

$$L(\boldsymbol{\theta}, \sigma^۲, \lambda|\boldsymbol{y}, \boldsymbol{W}) = \prod_{i=۱}^{s}\prod_{j=۱}^{r} ۲\phi(y_{ij}|\boldsymbol{w}_{ij}\boldsymbol{\theta} - \sqrt{\frac{۲}{\pi}}, \sigma^۲ + \lambda^۲)\Phi(\frac{\lambda}{\sigma}\frac{(y_{ij} - \boldsymbol{w}_{ij}\boldsymbol{\theta})}{(\sigma^۲+\lambda^۲)^{\frac{۱}{۲}}}),$$

است. چون براوردگرهای ماکسیمم درستنمایی پارامترهای تحلیل کواریانس تحت فرض چوله نرمال بودن توزیع متغیر پاسخ فرم بسته ندارند، برای استفاده از الگوریتم EM لازم است مسأله در قالب داده‌های کامل فرمول‌بندی شود. برای این منظور، توزیع چوله نرمال را می‌توان به‌صورت آمیخته‌ای از توزیع‌های نرمال و نیم‌نرمال به‌شکل $T_{ij} \sim HN(۰, ۱)$ و $Y_{ij}|T_{ij} = t_{ij} \sim N(\boldsymbol{w}_{ij}\boldsymbol{\theta} + \lambda(t_{ij} - \sqrt{\frac{۲}{\pi}}), \sigma^۲)$, $i = ۱, ..., s;\ j = ۱, ..., r$ نوشت. بر این اساس اگر Y_{ij}ها داده‌های ناقص و T_{ij}ها داده‌های گم‌شده تلقی شوند، تابع چگالی توام (Y_{ij}, T_{ij}) را می‌توان به‌صورت

$$f_{(Y_{ij},T_{ij})}(y_{ij}, t_{ij}) = \frac{۱}{\sqrt{۲\pi\sigma^۲}}\exp\Big\{-\frac{۱}{۲\sigma^۲}(y_{ij} - \boldsymbol{w}_{ij}\boldsymbol{\theta} - \lambda(t_{ij} - \sqrt{\frac{۲}{\pi}}))^۲\Big\} \times \sqrt{\frac{۲}{\pi}}\exp\Big\{-\frac{t_{ij}^۲}{۲}\Big\},$$

نوشت. بنابراین تابع لگ‌درستنمایی داده‌های کامل به‌صورت

$$\begin{aligned}\ell_c(\boldsymbol{\theta}, \sigma^۲, \lambda|\boldsymbol{y}, \boldsymbol{W}, \boldsymbol{t}) &= C - \frac{sr}{۲}\log\sigma^۲ - \frac{۱}{۲\sigma^۲}\Big((\boldsymbol{y} - \boldsymbol{W\theta})'(\boldsymbol{y} - \boldsymbol{W\theta}) - ۲\lambda(\boldsymbol{y} - \boldsymbol{W\theta})'\boldsymbol{t} \\ &- \lambda^۲\mathbf{1}'_{sr}(۲\sqrt{\frac{۲}{\pi}}\boldsymbol{t} - \boldsymbol{t}^۲) + ۲\lambda\sqrt{\frac{۲}{\pi}}\mathbf{1}'_{sr}(\boldsymbol{y} - \boldsymbol{W\theta}) + sr\lambda^۲\frac{۲}{\pi}\Big),\end{aligned}$$

است، که در آن $\boldsymbol{t} = (t_{۱۱}, ..., t_{sr})'$، $\boldsymbol{t}^۲ = (t_{۱۱}^۲, ..., t_{sr}^۲)'$ و $\mathbf{1}_{sr}$ یک بردار ستونی sr بعدی از یک‌هاست. در مرحله E از الگوریتم EM، محاسبه امید ریاضی تابع لگ‌درستنمایی داده‌های کامل به شرط داده‌های گم‌شده ضروری است. با استفاده

از روابط مربوط به گشتاورهای توزیع نرمال بریده‌شده (به عنوان نمونه [۳] را ببینید)، روابط

$$\begin{aligned}\hat{t}_{ij} &= E(t_{ij}|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, y_{ij}) = \hat{\eta}_{ij} + \hat{\tau}\hat{\delta}_{ij},\\ \hat{t^{2}}_{ij} &= E(t_{ij}^{2}|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, y_{ij}) = \hat{\eta}_{ij}^{2} + \hat{\tau}^{2} + \hat{\tau}\hat{\delta}_{ij}\hat{\eta}_{ij},\end{aligned}$$

به‌دست می‌ایند، که در آن‌ها $\hat{\eta}_{ij} = \frac{\hat{\lambda}}{\hat{\sigma}^{2}+\hat{\lambda}^{2}}(y_{ij} - \boldsymbol{w}_{ij}\hat{\boldsymbol{\theta}} + \sqrt{\frac{2}{\pi}}\hat{\lambda})$، $\hat{\tau}^{2} = \frac{\hat{\sigma}^{2}}{\hat{\sigma}^{2}+\hat{\lambda}^{2}}$ و $\hat{\delta}_{ij} = \frac{\phi(\frac{\hat{\eta}_{ij}}{\hat{\tau}})}{\Phi(\frac{\hat{\eta}_{ij}}{\hat{\tau}})}$. از این رو، امید ریاضی تابع لگ‌درستنمایی داده‌های کامل به شرط داده‌های گم شده به‌صورت

$$\begin{aligned}E(\ell_c(\boldsymbol{\theta}, \sigma^{2}, \lambda|\boldsymbol{y}, \boldsymbol{W})) &= C - \frac{sr}{2}\log\sigma^{2} - \frac{1}{2\sigma^{2}}\Big((\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta})'(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta}) - 2\lambda(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta})'\hat{\boldsymbol{t}}\\ &- 2\lambda^{2}\sqrt{\frac{2}{\pi}}\mathbf{1}_{sr}'\hat{\boldsymbol{t}} + \lambda^{2}\mathbf{1}_{sr}'\hat{\boldsymbol{t}}^{2} + 2\lambda\sqrt{\frac{2}{\pi}}\mathbf{1}_{sr}'(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta}) + sr\lambda^{2}\frac{2}{\pi}\Big),\end{aligned}$$

است. در مرحله M الگوریتم، مقادیری از پارامتر که امید ریاضی شرطی تابع لگ‌درستنمایی داده‌های کامل را به شرط داده‌های گم شده ماکسیمم می‌سازند، به عنوان براوردگرهای ماکسیمم درستنمایی انتخاب می‌شوند. بر این اساس، در تکرار $(k+1)$ام الگوریتم، پارامترها را براساس مقادیر آن‌ها در تکرار tام به‌صورت زیر به‌دست می‌ایند:

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{(k+1)} &= (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\Big(\boldsymbol{y} + \hat{\lambda}^{(k)}(\sqrt{\frac{2}{\pi}}\mathbf{1}_{sr} - \hat{\boldsymbol{t}}^{(k)})\Big),\\ \hat{\sigma^{2}}^{(k+1)} &= \frac{1}{sr}\Big((\boldsymbol{y}-\boldsymbol{W}\hat{\boldsymbol{\theta}}^{(k)})'(\boldsymbol{y}-\boldsymbol{W}\hat{\boldsymbol{\theta}}^{(k)}) - 2\hat{\lambda}^{(k)}(\boldsymbol{y}-\boldsymbol{W}\hat{\boldsymbol{\theta}}^{(k)})'\hat{\boldsymbol{t}}^{(k)}\\ &- \hat{\lambda}^{2(k)}\mathbf{1}_{sr}'(2\sqrt{\frac{2}{\pi}}\hat{\boldsymbol{t}}^{(k)} - \hat{\boldsymbol{t}^{2}}^{(k)}) + 2\hat{\lambda}^{(k)}\sqrt{\frac{2}{\pi}}\mathbf{1}_{sr}'(\boldsymbol{y}-\boldsymbol{W}\hat{\boldsymbol{\theta}}^{(k)}) + sr\hat{\lambda}^{2(k)}\frac{2}{\pi}\Big),\\ \hat{\lambda}^{(k+1)} &= \Big(\sqrt{\frac{2}{\pi}}\mathbf{1}_{sr}'(\boldsymbol{y}-\boldsymbol{W}\hat{\boldsymbol{\theta}}^{(k)}) - (\boldsymbol{y}-\boldsymbol{W}\hat{\boldsymbol{\theta}}^{(k)})'\hat{\boldsymbol{t}}^{(k)}\Big)\Big(\mathbf{1}_{sr}'(2\sqrt{\frac{2}{\pi}}\hat{\boldsymbol{t}}^{(k)} - \hat{\boldsymbol{t}^{2}}^{(k)}) - sr\frac{2}{\pi}\Big)^{-1}\end{aligned}$$

۳ مطالعه شبیه‌سازی

در این بخش، برای ارزیابی کارایی براوردگرهای ماکسیمم درستنمایی پیشنهادی، مطالعه‌ای شبیه‌سازی طراحی و اجرا شده است. برای این منظور، یک مدل تحلیل کواریانس با یک متغیر تبیینی و یک تیمار به‌صورت

$$\mu_{ij} = E(y_{ij}|\boldsymbol{X}, \boldsymbol{Z}) = \beta_{0} + \beta_{i} + \gamma(z_{ij} - \bar{z}_{i}) \quad i = 1, 2\ ;\ j = 1, \ldots, r, \qquad (2)$$

به ازای $\beta_{0} = 2$ ، $\beta_{1} = 5$ و $\gamma = 1$ در نظر گرفته شده است. متغیرهای پاسخ y_{ij}، $i = 1, 2$ و $j = 1, \cdots, r$، از توزیع $SSN(\boldsymbol{w}_{ij}\boldsymbol{\theta} - \sqrt{\frac{2}{\pi}}, \sigma^{2}, \lambda)$ شبیه‌سازی شده‌اند. به منظور ارزیابی قابلیت مدل پیشنهادی در مدل‌بندی مشاهدات دارای وضعیت‌های مختلف تقارن و چولگی، مقادیر مختلفی برای پارامتر چولگی در محدوده $\{-2, -1, 0, 1, 2\}$ لحاظ شده است. مقادیر ریشه میانگین توان دوم خطا $(RMSE)$ براوردگرهای ماکسیمم درستنمایی پارامترهای مدل تحلیل کواریانس تحت شرایط بالا محاسبه و در جدول (۱) نشان داده شده است. در تمامی حالات براوردگرهای حاصل از مدل نرمال نیز به عنوان براوردگرهای سنتی و برای نشان دادن عدم استواری این براوردگرها در مقابل انحراف از فرض نرمال بودن مشاهدات، به عنوان مدل رقیب مدنظر قرار گرفته است. به منظور لحاظ نمودن عدم قطعیت حاکم بر فرایند تولید نمونه‌های تصادفی و فراهم آوردن امکان تغییر احتمال به طریقه فراوانی نسبی، در تمام محاسبات تعداد تکرارها ۵۰۰۰ در نظر گرفته شده است. همانطور که انتظار می‌رود به ازای $\lambda = 0$ که مشاهدات متقارن بوده و توزیع‌های نرمال و چوله نرمال بر هم منطبق هستند،

جدول ۱: مقادیر ریشه میانگین توان دوم خطا براوردگرهای ماکسیمم درستنمایی ضرایب مدل تحلیل کواریانس چوله نرمال به همراه مقادیر متناظر برای توزیع نرمال به ازای اندازه نمونه‌های مختلف

		مدل تحلیل کواریانس						
		چوله نرمال				نرمال		
λ	اندازه نمونه	λ	β_0	β_1	γ	β_0	β_1	γ
-۲	۱۰۰	۴/۴۵۲۰	۰/۹۱۲۶	۴/۷۳۳۴	۱۲/۲۶۰۳	۱/۰۱۸۶	۵/۰۰۸۴	۱۲/۶۷۷۱
	۲۰۰	۳/۹۰۹۰	۰/۹۲۰۶	۴/۸۴۸۵	۱۰/۱۲۴۹	۱/۰۰۷۷	۵/۰۰۲۹	۹/۶۶۵۶
-۱	۱۰۰	۳/۷۴۶۴	۰/۶۹۹۵	۴/۷۳۷۰	۸/۲۷۰۹	۱/۰۰۶۸	۵/۰۰۶۸	۸/۶۴۵۳
	۲۰۰	۳/۴۰۹۷	۰/۷۱۷۰	۴/۷۹۸۲	۵/۷۷۴۹	۱/۰۰۱۷	۵/۰۰۴۰	۶/۱۴۴۲
۰	۱۰۰	$< ۱ \times ۱۰^{-۱۳}$	۱/۰۰۳۶	۵/۰۰۱۳	۵/۴۰۹۳	۱/۰۰۳۶	۵/۰۰۱۳	۵/۴۰۹۳
	۲۰۰	$< ۱ \times ۱۰^{-۱۳}$	۱/۰۰۱۹	۴/۹۹۸۱	۳/۶۸۲۶	۱/۰۰۱۹	۴/۹۹۸۱	۳/۶۸۲۶
۱	۱۰۰	۰/۷۷۸۸	۰/۶۷۱۱	۴/۹۵۹۵	۷/۴۶۲۵	۱/۰۰۶۷	۵/۰۰۳۲	۷/۸۱۴۳
	۲۰۰	۰/۹۵۹۹	۰/۹۹۸۹	۴/۹۹۸۴	۶/۱۸۷۵	۱/۰۰۳۷	۵/۰۰۰۲	۶/۲۸۷۷
۲	۱۰۰	۳/۳۶۴۹	۰/۵۳۰۸	۴/۱۷۵۰	۹/۶۹۹۴	۱/۰۱۳۱	۵/۰۰۹۰	۱۲/۹۵۱۱
	۲۰۰	۳/۸۳۸۹	۰/۴۴۰۱	۳/۸۸۴۵	۷/۶۱۰۹	۱/۰۱۰۰	۵/۰۰۱۲	۱۰/۲۶۷۶

به وضوح تفاوتی بین مدل‌های تحلیل کواریانس تحت فرض نرمال و چوله نرمال وجود ندارد. به ازای مقادیر مثبت و منفی پارامتر چولگی λ، که به ترتیب متناظر با مشاهدات چوله به راست و چپ هستند، به وضوح مدل پیشنهادی از کارایی بیشتری برخوردار است.

۴ بحث و نتیجه‌گیری

به‌صورت سنتی یکی از فرض‌های اولیه در تحلیل کواریانس، نرمال بودن توزیع متغیر پاسخ است. این در حالی است که در بسیاری از کاربردها توزیع مشاهدات نامتقارن است. در چنین شرایطی استفاده از توزیع نرمال نتایج گمراه کننده‌ای به دنبال خواهد داشت. استفاده از توزیع چوله نرمال برای مدل‌بندی تغییرات متغیر پاسخ، کارایی مدل را به‌صورت قابل ملاحظه‌ای افزایش می‌دهد.

مراجع

[1] D.Aigner, C. Lovell, P. Schmidt, *Formulation and Estimation of Stochastic Frontier Production Function Model*, Journal of Econometrics, 12 (1977), PP. 21–37.

[2] A. Azzalini, *A Class of Distributions which Includes the Normal ones*, Scandinanavi Journal Statistics, 12 (1985), PP. 171–178.

[3] R. Barr., E. Donald, and sherril, *Mean and Variance of Truncated Normal Distribution*, The American Statistician, 53 (1999), PP. 357–361.

[4] W. Cochran, *Analysis of Covariance: Its Nature and Uses*, Biometrics, 13 (1957), PP. 261–281.

[5] R. A. Fisher, *Statistical Methods for Research Workers*, 4th ed., Oliver and Boyed, Edinburgh, 1932.

[6] C. Roberts, *A Correlation Model Useful in The Study of Twins*, Journal of American Statistical Association, 61 (1966), PP. 1184–1190.

[7] S. Sahu, D. Dey, and M. Branco, *A New Class of Multivariate Distributions with Applications to Bayesian Regression Models*, Canadian Journal of Statistics, 29 (2003), PP. 129–150.

پست الکترونیکی: a.fallah@sci.ikiu.ac.ir
پست الکترونیکی: goodarzizahra67@yahoo.com