Statistical Research Center
J. Stat. Res. Iran 1 (2004) 85-100
مجلهٔ پژوهش‌های آماری ایران
دوره‌ی ۱، شماره‌ی ۱، بهار و تابستان ۱۳۸۳، صص ۸۵-۱۰۰

تحلیل یک مسئله با استفاده از دیدگاه‌های متفاوت

مجتبیٰ گنجعلی[†,*] و لیلا لطیفیان[‡]

[†] دانشگاه شهید بهشتی

[‡] بانک مرکزی جمهوری اسلامی ایران

چکیده. در این مقاله یک مسئله‌ی کاربردی که در آن، پاسخ مورد نظر، تعداد موفقیت‌ها در یک آزمایش به‌خصوص است مطرح می‌شود و برای مدل‌بندی از دیدگاه‌های متفاوت مورد بررسی قرار می‌گیرد. بررسی تأثیر مقدار پاسخ دورافتاده بر نتایج یک تحلیل رگرسیونی از اهمیت فراوان برخوردار است. به این دلیل با استفاده از روش‌های تشخیص، داده‌های دورافتاده شناسایی می‌شوند. نشان داده شده است که استفاده از روش تبدیل آرک‌سینوس ممکن است تشخیص دورافتاده بودن برخی داده‌ها را منحرف کند. در صورتی که حذف داده‌های دورافتاده در مدل‌بندی امکان‌پذیر نباشد، استفاده از روش‌های استوار پیشنهاد شده است. روش ماکسیمم درست‌نمایی که از آن در مدل‌های خطی تعمیم‌یافته و روش تبدیل، برای براورد پارامترها استفاده می‌شود، و روش شبه‌درست‌نمایی، استوار نیستند. روش دیگری که به نتایج استوار منجر می‌شود و به روش شبه‌درست‌نمایی استوار موسوم است، در این مقاله بازبینی و روشی که از نظر محاسباتی برای استفاده آسان‌تر است، برای محاسبه‌ی مقادیر معناداری در آن ارائه شده است. شیوه‌های مختلف مدل‌بندی نیز در مثال کاربردی مقایسه شده‌اند.

واژگان کلیدی. روش تبدیل؛ رگرسیون لوژستیک؛ مدل‌های خطی تعمیم‌یافته؛ مانده‌ی پی‌یرسونی؛ شبه‌درست‌نمایی استوار.

۱ مقدمه

تحلیل مسائل کاربردی آمار به تأمل و اندیشه‌ی عمیق نیاز دارد. گاهی در دید نخست، تحلیل یک مسئله‌ی آماری، ساده به نظر می‌رسد و روش‌های موجود یا بسیار سنّتی، به‌نظر به رهیافتی اساسی می‌انجامند. اما

* نویسنده‌ی عهده‌دار مکاتبات.

پس از تحلیل دقیق و در نظرگرفتن روش‌های مختلف، از جمله روش‌های سنّتی و مدرن است که می‌توان در مورد داده‌ها و آنچه آن‌ها به ما می‌آموزند اظهار نظر کرد و گاهی به اندیشه‌های جدید دست یافت.

هر چه در تحلیل توصیفی یک مجموعه داده‌ی آماری بیش‌تر بیندیشیم و سعی کنیم تا این مجموعه داده‌ها را با روش‌های آماری متعدد تحلیل کنیم، نه‌تنها استنباط بهتری می‌توانیم به دست آوریم، بلکه می‌توانیم اطلاع کسب کنیم که در چه موقعیت‌هایی چه روشی سودمندتر است؛ هر چند ممکن است هیچ روشی کامل نباشد. به همین دلیل است که شیوه‌های متعدد آماری در تحلیل‌های آماری رگرسیونی مورد استفاده قرار می‌گیرند.

به‌عنوان یک نمونه‌ی بارز کاربردی، حالتی را در نظر بگیرید که پاسخ مورد نظر، تعداد موفقیت‌ها در یک آزمایش به‌خصوص است و در صددیم تا تأثیر برخی متغیرهای تبیینی بر میانگین موفقیت‌ها را بررسی کنیم. یکی از روش‌های بسیار سنّتی برای تحلیل چنین داده‌هایی، استفاده از تبدیلی از نسبت موفقیت‌ها در نمونه، به‌گونه‌ای است که متغیرِ تبدیل‌یافته متغیری پیوسته و دارای توزیع نرمال با واریانسِ تثبیت‌شده باشد [۱۰]، [۲]. یکی دیگر از روش‌های تحلیل، استفاده از مدل‌های خطی تعمیم‌یافته است که از سوی نلدر و ودربرن [۱۵] معرفی شده است. البته برای پاسخ دوجمله‌ای، مدل رگرسیونی لوژستیک که حالت خاصی از مدل‌های خطی تعمیم‌یافته است، قبلاً توسط فینی [۸]، [۹] و برک‌سون [۳] بیان شده است. به هر حال، براوردگرهای ماکسیمم درست‌نمایی در مدل رگرسیونی لوژستیک نسبت به داده‌های دورافتاده نااستوارند و اخیراً رده‌ی براوردگرهای استوار، مورد تأکید قرار گرفته است. رهیافت شبه‌درست‌نمایی نیز توسط ودربرن [۱۷] معرفی شده است، که از معادلات براوردگر معرفی‌شده در آن نیز می‌توان برای براورد پارامترها استفاده کرد. به هر حال، براوردهای به دست آمده از این روش نیز نسبت به داده‌های دورافتاده نااستوارند.

باکس [۴] واژه‌ی «استواری» را برای نخستین بار ارائه کرد، هوبر [۱۲] نظریه‌ی استواری را مورد بررسی قرار داد، همپل [۱۱] آمار استوار را بر اساس تابع تأثیر بیان کرد و کانتونی و رونچتی [۵] استنباط استوار را برای مدل‌های خطی تعمیم‌یافته، بر اساس کیبش (deviance) استوار که تعمیم‌یافته‌ی شبه‌درست‌نمایی است، به دست آوردند. در این مقاله از این روش، روش تبدیل آرک‌سینوس و رگرسیون لوژستیک برای تحلیل داده‌هایی با پاسخ دوجمله‌ای استفاده می‌کنیم.

برای ایجاد انگیزه، در بخش بعد، مثال کاربردی را تشریح می‌کنیم. در بخش ۳ روش‌های مختلف تحلیل را می‌آوریم. در بخش ۴ نتایجِ به دست آمده از روش‌های مختلف را مقایسه و در بخش ۵ نتیجه‌گیری می‌کنیم.

۲ مسئله‌ی کاربردی

در این بخش، داده‌هایی که از فلپس [۱۶] گرفته شده است، مورد ارزیابی قرار می‌گیرد. این داده‌ها به‌وسیله‌ی ویلیامز [۱۸] نیز مورد بررسی قرارگرفته است. مکولا و نلدر [۱۴] این داده‌ها را به کارگرفته‌اند تا روش‌هایی برای ارزیابی انحرافات زیاد از مدل، که به‌دلیل حضور مشاهده‌ی دورافتاده ایجاد می‌شود، مورد بررسی قرار گیرد. این داده‌ها نسبت گوجه‌فرنگی‌های آفت‌زده را در آزمایشی با سه بلوک و هشت مقدار مصرف مختلف حشره‌کش نشان می‌دهد. این داده‌ها در جدول ۱ نشان داده شده است. در این جدول، نسبت گوجه‌فرنگی‌های آفت‌زده و لگاریتم مقدار مصرف حشره‌کش، در همه‌ی بلوک‌ها داده شده است.

همچنان‌که در جدول ۱ دیده می‌شود، در بیش‌تر موارد با افزایش مقدار مصرف حشره‌کش، نسبت گوجه‌فرنگی‌های آفت‌زده در بلوک‌ها کاهش می‌یابد. در بلوک ۲ برای مقدار مصرف ۲٫۱۲ (که از این به بعد آن را مشاهده‌ی ۱۴ می‌نامیم) تناقضی غیر عادی رخ می‌دهد. نسبت گوجه‌فرنگی‌های آفت‌زده برای این مقدار مصرف، ۰٫۴۰۵ است که مقدار آن به‌نظر بالاست. این مشاهده را با این نگاه کاوشگرانه، می‌توان یک مشاهده‌ی دورافتاده تلقی کرد. بررسی آماریِ این که آیا این مشاهده واقعاً دورافتاده است یا خیر، و تأثیری که این مشاهده می‌تواند در نتایجِ یک مدل آماری داشته باشد، در بخش‌های بعد مورد بررسی قرار می‌گیرد.

جدول ۱. داده‌های مربوط به گوجه‌فرنگی‌های آفت‌زده

لگاریتم	بلوک ۱			بلوک ۲			بلوک ۳		
مقدار مصرف	Y	n	$\frac{Y}{n}$	Y	n	$\frac{Y}{n}$	Y	n	$\frac{Y}{n}$
۱٫۵۲۰	۱۰	۳۵	۰٫۲۸۶	۱۷	۳۸	۰٫۴۴۷	۱۰	۳۴	۰٫۲۹۴
۱٫۶۴۰	۱۶	۴۲	۰٫۳۸۱	۱۰	۴۰	۰٫۲۵۰	۱۰	۳۸	۰٫۲۶۳
۱٫۷۶۰	۸	۵۰	۰٫۱۶۰	۸	۳۳	۰٫۲۴۲	۵	۳۶	۰٫۱۳۹
۱٫۸۸۰	۶	۴۲	۰٫۱۴۳	۸	۳۹	۰٫۲۰۵	۳	۳۵	۰٫۰۸۶
۲٫۰۰۰	۹	۳۵	۰٫۲۵۷	۵	۴۷	۰٫۱۰۶	۲	۴۹	۰٫۰۴۱
۲٫۱۲۰	۹	۴۲	۰٫۲۱۴	۱۷	۴۲	۰٫۴۰۵	۱	۴۰	۰٫۰۲۵
۲٫۲۴۰	۱	۳۲	۰٫۰۳۱	۶	۳۵	۰٫۱۷۱	۳	۲۲	۰٫۱۳۶
۲٫۳۶۰	۲	۲۸	۰٫۰۷۱	۴	۳۵	۰٫۱۱۴	۲	۳۱	۰٫۰۶۵

۳ تحلیل‌های گوناگون

در این بخش، روش‌های تحلیل تبدیل آرک‌سینوس، رگرسیون لوژستیک، و شبه‌درست‌نمایی استوار را مورد بررسی قرار می‌دهیم.

۳٫۱ روش تبدیل آرک‌سینوس

یکی از روش‌هایی که از آن می‌توان برای تحلیل داده‌های جدول ۱ استفاده کرد، این است که نخست تبدیلی مناسب برای پاسخ یافت تا آن را به متغیری تبدیل کند که دارای توزیع نرمال است و سپس با استفاده از رگرسیون معمولی، تأثیر متغیرهای تبیینی بر پاسخِ تبدیل‌شده را یافت. این‌که آیا در این روش، پس از تحلیل، با بررسی مانده‌ها می‌توان داده‌های دورافتاده را تشخیص داد یا خیر، سؤالی است که در پیِ یافتن جوابی برای آن هستیم.

یکی از تبدیل‌هایی که از آن می‌توان برای نسبت‌ها استفاده کرد تبدیل آرک‌سینوس، معرفی‌شده توسط اَنسکامب [۲] است که سعی در تثبیت واریانس دارد و به‌صورت زیر تعریف می‌شود:

$$\widetilde{Y}_i = \sin^{-1}\sqrt{\frac{Y_i + \frac{3}{8}}{n_i + \frac{3}{4}}}. \qquad (1)$$

$\widetilde{Y}_i$ دارای واریانس مجانبی $\frac{1}{4n_i+2}$ است. از آن‌جا که واریانس مجانبی به اندازه‌ی نمونه وابسته است، می‌توان براورد پارامترها را برای بردار p-بعدی از متغیرهای تبیینی $(\mathbf{x}_i)$، به‌روش کم‌ترین توان‌های دوم وزنی (WLS) با مینیمم کردن $\sum_i w_i(\widetilde{Y}_i - \eta_i)^2$ که در آن $w_i = 4n_i + 2$ و $\eta_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i$ پیشگوی خطی است، به دست آورد [۷]. راه هم‌ارز دیگر آن است که از متغیر

$$Z_i = (4n_i + 2)^{\frac{1}{2}}\widetilde{Y}_i \qquad (2)$$

که انتظار می‌رود به‌طور مجانبی برای هر i دارای واریانس تثبیت‌شده‌ی ۱ باشد، استفاده کنیم. در این صورت می‌توان از روش LS یا ماکسیمم درست‌نمایی، براورد پارامترها را یافت. شکل ۱ نمودار پاسخ تبدیل‌شده‌ی Z_i را در مقابل لگاریتم مقدار مصرف برای همه‌ی بلوک‌ها نشان می‌دهد. همان‌طورکه شکل ۱ نشان می‌دهد، با ابزار چشمی، هر چند می‌توان روند کلّی کاهش پاسخ تبدیل‌شده‌ی Z_i با افزایش لگاریتم مقدار مصرف را مشاهده کرد، ولی خیلی سخت می‌توان در مورد تفاوت تأثیر بلوک‌های مختلف بر پاسخ تبدیل‌شده اظهار نظر کرد. مشاهده‌ی ۱۴ ممکن است با تأثیر زیاد، نتایج تحلیل رگرسیونی را تحت‌الشعاع خود قرار دهد. برای داده‌های بخش ۲ مدل زیر را در نظر می‌گیریم:

$$Z_i = \beta_0 + \beta_1 \log(dose_i) + \beta_2 block_{2i} + \beta_3 block_{1i} + \varepsilon_i, \qquad (3)$$

که در آن، $i = 1, 2, \ldots, 24$، Z_i پاسخ تبدیل‌شده برای iامین آزمایش، β_0، β_1، β_2 و β_3 پارامترهای مدل، و $block_{ji}$ برای $j = 1, 2$ متغیرهای نشانگر می‌باشند که اگر آزمایش iام در بلوک jام انجام گیرد مقدار ۱ و در غیر این صورت مقدار ۰ می‌گیرند. ε_i خطای اندازه‌گیری برای iامین آزمایش است که فرض

می‌شود دارای توزیعی نرمال با میانگین ۰ و واریانس تثبیت‌شده‌ی σ^2 است.

مانند هر مدل رگرسیونی خطی دیگر، برای بررسی نیکویی برازش مدل، مانده‌های پی‌یرسونی عبارت‌اند از:

$$r_i = \frac{z_i - \hat{\mu}}{\hat{\sigma}}, \qquad (۴)$$

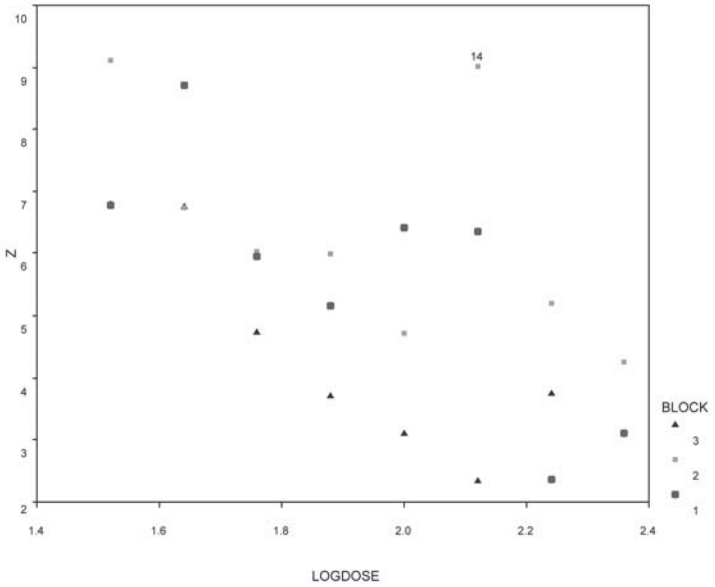
که در آن $\hat{\mu}_i = \widehat{E(Z_i|\mathbf{x}_i)}$. مانده‌ی پی‌یرسونی استاندارد شده عبارت است از:

$$\tilde{r}_i = \frac{r_i}{\sqrt{1 - \hat{h}_i}}, \qquad (۵)$$

که در آن $\hat{h}_i$ مؤلفه‌ی براورد شده‌ی iامِ ماتریس کلاه‌دار (معروف به تأثیر) می‌باشد. مانده‌ی کیبش [۱] عبارت است از:

$$\sqrt{d_i}\ \mathrm{sgn}(z_i - \hat{\mu}_i), \qquad (۶)$$

که در آن $\mathrm{sgn}(\cdot)$ تابع علامت است و $d_i = (z_i - \hat{\mu}_i)^2$.



شکل ۱. نمودار پاسخ تبدیل‌شده‌ی Z_i در مقابل لگاریتم مقدار مصرف برای همه‌ی بلوک‌ها

در رگرسیون خطی، نمودارهای (۱) مانده‌ها در مقابل مقادیر برازش‌یافته، (۲) نمودار Q-Q، (۳) جذر قدرمطلق مانده‌ی استاندارد شده در مقابل مقادیر برازش‌یافته، و (۴) نمودار فاصله‌ی کوک (h_iها) در مقابل شماره‌ی مشاهده، می‌توانند در تشخیص داده‌های دورافتاده و مؤثر مورد استفاده قرارگیرند [۷].

۳٫۲ رگرسیون لوژستیک

در این روش فرض می‌شود متغیر پاسخ (تعداد موفقیت‌ها) دارای توزیع دوجمله‌ای است و چون این توزیع از خانواده‌ی نمایی است می‌توان از مدل‌های خطی تعمیم‌یافته (GLM) [۱۴] برای یافتن تأثیر متغیرهای تبیینی بر میانگین تعداد موفقیت‌ها (تعداد آفت زدن‌های گوجه‌فرنگی، $\hat{Y}_i$، در مثال ما) کمک گرفت. در واقع اگر برای $\mu_i = \mu(\mathbf{x}_i) = E(Y_i|\mathbf{x}_i)$ فرض کنیم

$$\ln \frac{\mu(\mathbf{x}_i)}{n_i - \mu(\mathbf{x}_i)} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i, \tag{۷}$$

که در آن $\mathbf{x}$ بردار p-بعدی از متغیرهای تبیینی است و $\boldsymbol{\beta}$ بردار پارامترهای نامعلوم است، از یک GLM با پاسخ دوجمله‌ای و تابع ربط لوجیت استفاده کرده‌ایم. براورد پارامترهای این مدل به‌راحتی می‌تواند با استفاده از اکثر نرم‌افزارهای آماری همچون SPSS، S-Plus یا R به دست آید. تابع درست‌نمایی برای پاسخ دوجمله‌ای با N آزمودنی و احتمال موفقیت $\pi_i = \pi(\mathbf{x}_i)$ برای آزمودنی iام، عبارت است از:

$$L(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X}) = \prod_{i=1}^{N}\left[\binom{n_i}{y_i}\pi^{y_i}(\mathbf{x}_i)\{1-\pi(\mathbf{x}_i)\}^{n_i-y_i}\right],$$

که در آن $\mathbf{Y} = (Y_1,\ldots,Y_N)^{\mathrm{T}}$، و $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_N)^{\mathrm{T}}$ ماتریس طرح است و از آن‌جا

$$\begin{aligned}\ell(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X}) &= \ln\{L(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})\}\\ &= \sum_{i=1}^{N}\left[\ln\binom{n_i}{y_i} + y_i\ln\pi(\mathbf{x}_i) + (n_i-y_i)\ln\{1-\pi(\mathbf{x}_i)\}\right].\end{aligned} \tag{۸}$$

این تابع لُگ‌درست‌نمایی به‌روش WLS تکراری [۱۴، ص ۴۰] ماکسیمم می‌شود و براورد پارامترها به دست می‌آید.

مکولا و نِلدر [۱۴] برای بررسی نیکویی برازش، مدل کیبش را به‌صورت زیر تعریف کرده‌اند:

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = -2\{\ell(\hat{\boldsymbol{\mu}};\mathbf{y}) - \ell(\mathbf{y};\mathbf{y})\}, \tag{۹}$$

که در آن، $\ell(\hat{\boldsymbol{\mu}};\mathbf{y})$ ماکسیمم لُگ‌درست‌نمایی برای مدل را نشان می‌دهد و $\ell(\mathbf{y};\mathbf{y})$ ماکسیمم لُگ‌درست‌نمایی برای کلّی‌ترین مدل است که برای هر یک از مشاهدات، پارامتر مجزّایی دارد و براورد $\boldsymbol{\mu}$ برای آن، $\mathbf{y}$ است. این مدل، مدل اشباع‌شده نامیده می‌شود. $D(\mathbf{y};\hat{\boldsymbol{\mu}})$ دارای توزیع تقریبی خی‌دو با درجه‌ی آزادی $N-p$ است، که N تعداد مشاهدات و p تعداد پارامترهای مدل است. اگر Y_i دارای توزیع دوجمله‌ای با پارامترهای n_i و π_i باشد، کیبش عبارت است از:

$$D = ۲\sum_{i=۱}^{N}\left\{y_i \ln\frac{y_i}{n_i\hat{\pi}_i} + (n_i-y_i)\ln\frac{n_i-y_i}{n_i-n_i\hat{\pi}_i}\right\}.$$

برای مقایسه‌ی دو مدل آشیانی به‌ترتیب با p و q پارامتر $(q<p<N)$ از اختلاف بین کیبش آن‌ها می‌توان استفاده کرد. تغییر در کیبش به‌صورت زیر تعریف می‌شود:

$$\Delta D = ۲\{\ell(\mathbf{b}_۱;\mathbf{y}) - \ell(\mathbf{b}_۰;\mathbf{y})\}, \qquad (۱۰)$$

که در آن $\mathbf{b}_۰$ و $\mathbf{b}_۱$ به‌ترتیب براورد بردار پارامترها تحت مدل‌های با q و p پارامتر هستند. ΔD دارای توزیع تقریبی خی‌دو با $p-q$ درجه‌ی آزادی است.

مانده‌ی پی‌یرسونی iامین آزمودنی برای نیکویی برازش مدل، عبارت است از:

$$r_i = \frac{y_i-\hat{\mu}_i}{\{\widehat{\mathrm{var}}(y_i)\}^{\frac{۱}{۲}}}, \qquad (۱۱)$$

و مانده‌ی کیبش عبارت است از:

$$\sqrt{d_i}\ \mathrm{sgn}(y_i-\hat{\mu}_i), \qquad (۱۲)$$

که در آن،

$$d_i = y_i\ln\frac{y_i}{n_i\hat{\pi}_i} + (n_i-y_i)\ln\frac{n_i-y_i}{n_i-n_i\hat{\pi}_i}.$$

iامین مانده‌ی پی‌یرسونی استاندارد شده عبارت است از:

$$\tilde{r}_i = \frac{r_i}{\sqrt{۱-\hat{h}_i}}. \qquad (۱۳)$$

برای داده‌های ارائه شده در بخش ۲، مدل رگرسیون لوژستیک زیر را در نظر خواهیم گرفت:

$$\ln\frac{\mu(\mathbf{x}_i)}{n_i-\mu(\mathbf{x}_i)} = \beta_۰ + \beta_۱\log(dose_i) + \beta_۲ block_{۲i} + \beta_۳ block_{۱i}. \qquad (۱۴)$$

۳٫۳ روش شبه‌درست‌نمایی استوار

برای بررسی استواری براوردگرها از معیار تابع تأثیر استفاده می‌شود. در روش اِم‌براوردگر، به‌جای مینیمم کردن مجموع توان دوم مانده‌ها، مجموع $\sum_{i=1}^{n}\rho(r_i)$ برای یافتن براوردها مینیمم می‌شود، که در آن ρ تابعی متقارن است. تابع تأثیر یک اِم‌براوردگر، متناسب با تابع امتیازش است، که تابع امتیاز، مشتق $\sum_{i=1}^{n}\rho(r_i)$ نسبت به ضرایب رگرسیونی است (همپل [۱۱] را ببینید). بنا بر این، براوردگرهای ماکسیمم درست‌نمایی در GLM که نوعی اِم‌براوردگرند، استوار نیستند؛ یعنی، انحراف‌های بزرگِ متغیر پاسخ از میانگین‌اش یا نقاط مؤثر روی محور رگرسیونی، می‌توانند تأثیر بزرگی روی براوردگر داشته باشند. در نتیجه، با استفاده از یک تابع امتیازکراندار می‌توان براوردگر استوار به دست آورد. کانتونی و رونچتی [۵] با توجه به براوردگرهای شبه‌درست‌نمایی، براوردهای شبه‌درست‌نمایی استوار را به دست آورده‌اند. در GLM، براوردگرهای شبه‌درست‌نمایی پارامترهای مدل، از حل معادلات براوردگر زیر به دست می‌آیند:

$$\sum_{i=1}^{n}\frac{y_i-\mu_i}{\mathrm{V}(Y_i)}\mu_i'=\mathbf{0}, \qquad (۱۵)$$

که در آن، $\mu_i=E(Y_i|\mathbf{x}_i)$، $\mu_i'=\frac{\partial\mu_i}{\partial\beta}$، و $\mathrm{V}(Y_i)=\mathrm{var}(Y_i)$. توجه کنید که مانند هر GLM دیگر، برای تابع ربط یکنوای $h(\cdot)$ داریم $h(\mu_i)=\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i$. در این روش تابع ربط، ارتباط واریانس پاسخ با میانگین آن و استقلال پاسخ‌ها را دانسته فرض می‌کنند؛ امّا فرضی در مورد توزیع پاسخ اعمال نمی‌شود.

در روشی که توسط کانتونی و رونچتی [۵] پیشنهاد شده است، براوردگرها از حل معادلات براوردگر زیر به دست می‌آیند:

$$\sum_{i=1}^{n}\boldsymbol{\Psi}(y_i,\mu_i)=\mathbf{0}, \qquad (۱۶)$$

که $\boldsymbol{\Psi}(y_i,\mu_i)=\nu(y_i,\mu_i)w(\mathbf{x}_i)\mu_i'-a(\boldsymbol{\beta})$، و

$$a(\boldsymbol{\beta})=\frac{1}{n}\sum_{j=1}^{n}E\{\nu(Y_j,\mu_j)\}w(\mathbf{x}_j)\mu_j'$$

جمله‌ای تصحیح‌کننده است که ما را از سازگاری فیشر مطمئن می‌کند. یا به‌طور معادل، عبارت زیر نسبت به β مینیمم می‌شود:

$$Q_M(\mathbf{y},\boldsymbol{\mu})=\sum_{i=1}^{n}Q_M(y_i,\mu_i),$$

که:

$$Q_M(y_i,\mu_i)=\int_{\tilde{s}}^{\mu_i}\nu(y_i,t)w(\mathbf{x}_i)\,dt-\frac{1}{n}\sum_{j=1}^{n}\int_{\tilde{t}}^{\mu_j}E\{\nu(Y_j,t)\}w(\mathbf{x}_i)\,dt,$$

و $\tilde{s}$ و $\tilde{t}$ به‌ترتیب به‌گونه‌ای انتخاب می‌شوند که $\nu(y_i,\tilde{s})=0$ و $E\{\nu(Y_i,\tilde{t})\}=0$. در این معادلات، اثر داده‌ی دورافتاده‌ی پاسخ با انتخاب تابع کرانداری برای $\nu(y_i,\mu_i)$ کنترل می‌شود. همچنین اثر مشاهدات مؤثر در متغیرهای تبیینی با انتخاب مناسب $w(\mathbf{x}_i)$ کنترل می‌شود [۵].

برای داده‌های مثال بخش ۲، با وجود یک مشاهده‌ی دورافتاده در متغیر پاسخ و نداشتن نقطه‌ی مؤثر در متغیرهای تبیینی، در این معادله فرض می‌کنیم:

$$\nu(y_i,\mu_i)=\frac{\Psi_c(r_i)}{\sqrt{\mathrm{V}(\mu_i)}} \qquad \text{و} \qquad w(x_i)=1$$

که در آن، $r_i=\frac{y_i-\mu_i}{\sqrt{\mathrm{V}(\mu_i)}}$ مانده‌های پی‌یرسونی و $\Psi_c(r_i)$ تابع هوبر است که برابر است با:

$$\Psi_c(r_i)=\begin{cases} r_i, & |r_i|\leqslant c \\ c\times \mathrm{sgn}(r_i), & |r_i|>c \end{cases}.$$

بنا بر این، براوردها از حل معادلات براوردگر زیر به دست می‌آیند:

$$\sum_{i=1}^{n}\left\{\frac{\Psi_c(r_i)}{\sqrt{\mathrm{V}(\mu_i)}}\mu_i'-a(\boldsymbol{\beta})\right\}=\mathbf{0}, \qquad (۱۷)$$

که در آن،

$$a(\boldsymbol{\beta})=\frac{1}{n}\sum_{j=1}^{n}E\left\{\frac{\Psi_c(r_j)}{\sqrt{\mathrm{V}(\mu_j)}}\right\}\mu_j'$$

برای مدل‌های دوجمله‌ای و پواسون دارای فرمی بسته است و نیازی به انتگرال‌گیری عددی ندارد [۵]. ثابت تنظیم‌کننده‌ی هوبر، c، برای حفظ کارایی تقریبی، برابر با $\frac{1}{2}$ در نظر گرفته می‌شود. با این شرایط، براوردگرِ به دست آمده، براوردگر شبه‌درست‌نمایی هوبر نامیده می‌شود. براوردها با حل معادله‌ی (۱۷) با استفاده از روش نیوتون-رافسون یا امتیازبندی فیشر به دست می‌آیند.

برای ارزیابی نیکویی برازش، مدل شبه‌کیبش به‌صورت زیر تعریف می‌شود:

$$D_{QM}(\mathbf{y},\boldsymbol{\mu})=-2Q_M(\mathbf{y},\boldsymbol{\mu})=-2\sum_{i=1}^{n}Q_M(y_i,\mu_i). \qquad (۱۸)$$

برای انتخاب بین دو مدل آشیانی، از اختلاف شبه‌کیبش دو مدل استفاده می‌شود، که عبارت است از:

$$\Lambda_{QM}=\{D_{QM}(\mathbf{y},\boldsymbol{\mu}^{\circ})-D_{QM}(\mathbf{y},\hat{\boldsymbol{\mu}})\}, \qquad (۱۹)$$

که در آن، $\boldsymbol{\mu}^{\circ}$ و $\hat{\boldsymbol{\mu}}$ به‌ترتیب با محاسبه‌ی براورد پیشگوی خطی تحت فرض صفر (دارای $p-q$ پارامتر) و براورد پیشگوی خطی تحت فرض مقابل (دارای p پارامتر) به دست می‌آیند.

برای به دست آوردن مقدار احتمال، لازم است که توزیع تقریبی آماره‌ی آزمون محاسبه شود. کانتونی و رونچتی [۵] نشان داده‌اند که توزیع تقریبی Λ_{QM} مانند توزیع $\sum_{i=1}^{q} d_i N_i^{2}$ است، که N_iها متغیرهای نرمال استاندارد مستقل، d_iها ویژه‌مقدارهای مثبت ماتریس $Q(\boldsymbol{\Psi},F)C(\boldsymbol{\Psi},F)$، و q بُعد پارامتر تحت فرض صفر است، که

$$Q(\boldsymbol{\Psi},F) = E\{\boldsymbol{\Psi}(Y,\mu)\boldsymbol{\Psi}^{\mathrm{T}}(Y,\mu)\},$$
$$C(\boldsymbol{\Psi},F) = M^{-1}(\boldsymbol{\Psi},F) - \widetilde{M}^{+}(\boldsymbol{\Psi},F),$$

که در آن،

$$M(\boldsymbol{\Psi},F) = -E\left\{\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\Psi}(Y,\mu)\right\},$$

و اگر ماتریس $M(\boldsymbol{\Psi},F)$ به‌صورت زیر افراز شود:

$$M(\boldsymbol{\Psi},F) = \begin{pmatrix} M_{11}(\boldsymbol{\Psi},F) & M_{12}(\boldsymbol{\Psi},F) \\ M_{21}(\boldsymbol{\Psi},F) & M_{22}(\boldsymbol{\Psi},F) \end{pmatrix},$$

که در آن، $M_{11}(\boldsymbol{\Psi},F)$ از مرتبه‌ی $(p-q)\times(p-q)$، $M_{12}(\boldsymbol{\Psi},F)$ از مرتبه‌ی $(p-q)\times q$، $M_{21}(\boldsymbol{\Psi},F)$ از مرتبه‌ی $q\times(p-q)$، و $M_{22}(\boldsymbol{\Psi},F)$ از مرتبه‌ی $q\times q$ است، آن‌گاه $\widetilde{M}^{+}(\boldsymbol{\Psi},F)$ ماتریسی است که $\widetilde{M}_{11}^{+}(\boldsymbol{\Psi},F) = M_{11}^{-1}(\boldsymbol{\Psi},F)$ و $\widetilde{M}_{12}^{+}(\boldsymbol{\Psi},F) = \widetilde{M}_{21}^{+}(\boldsymbol{\Psi},F) = \widetilde{M}_{22}^{+}(\boldsymbol{\Psi},F) = \mathbf{0}$.

کانتونی و رونچتی [۵]، از روش ایمهاف [۱۳] برای محاسبه‌ی مقادیر بحرانی توزیع Λ_{QM} استفاده کرده‌اند، که نیاز به اجرای الگوریتم‌های عددیِ طولانی دارد. ما برای محاسبه‌ی مقدار احتمال، از روش گشتاوری با تقریب ساترت‌وِیت ([۶] را برای یک مثالِ داده‌شده ملاحظه کنید) که محاسبات ساده‌تری را در بر دارد استفاده می‌کنیم و توزیع تجمعی ترکیب خطی از متغیرهای تصادفی با توزیع خی‌دو را با توزیع خی‌دویی که به درجه‌ی آزادی آن تقسیم شده است و این درجه‌ی آزادی برابر است با

$$\nu = \frac{\left(\sum_{i=1}^{q} d_i N_i^{2}\right)^{2}}{\sum_{i=1}^{q} d_i^{2} N_i^{4}}, \qquad (۲۰)$$

تقریب می‌زنیم.

۴ نتایج تحلیل‌ها

جدول ۲ نتایج تحلیل با استفاده از روش‌های گوناگون را نشان می‌دهد.

جدول ۲. نتایج تحلیل با روش‌های مختلف

پارامتر	براورد ML در مدل لوژستیک	براورد ML در روش تبدیل	براورد شبه‌درست‌نمایی استوار
β_0	۱٫۴۸۰ (۰٫۶۵۶)	۱۳٫۱۰۷ (۱٫۹۵۹)	۱٫۹۳۹ (۰٫۷۰۰)
β_1	−۱٫۸۱۷ (۰٫۳۴۳)	−۴٫۵۴۵ (۰٫۹۸۱)	−۲٫۰۴۹ (۰٫۳۷۰)
β_2	۰٫۸۴۳ (۰٫۲۲۶)	۲٫۰۹۲ (۰٫۶۶۰)	۰٫۶۸۵ (۰٫۲۴۰)
β_3	۰٫۵۴۲ (۰٫۲۳۲)	۱٫۳۱۳ (۰٫۶۶۰)	۰٫۴۵۰ (۰٫۲۴۰)
مانده‌ی کیبش	۳۹٫۹۷۶	۳۴٫۸۸۶	۳۲٫۷۴۰
درجه‌ی آزادی کیبش	۲۰	۲۰	۲۰

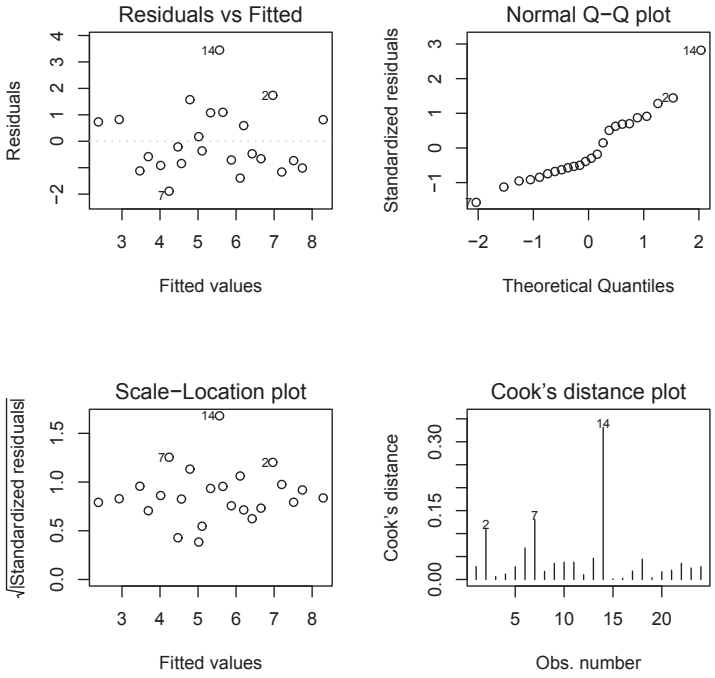
همان‌طور که جدول ۲ نشان می‌دهد، مانده‌ی کیبش در همه‌ی مدل‌ها نسبت به درجه‌ی آزادی آن زیاد است، ولی در مدل شبه‌درست‌نمایی استوار، این مقدار، کم‌ترین است. کیبش زیاد به دو دلیل ممکن است رخ دهد: (۱) بیش‌پراکندگی و (۲) وجود نقطه‌ی دورافتاده [۱۴]. در روش تبدیل، $\widehat{\sigma} = ۱٫۳۲۱$ و $R^2 = ۰٫۶۱۴$ به دست می‌آید. این براورد زیاد σ ممکن است به دلیل وجود داده‌ی دورافتاده باشد. شکل‌های ۲ و ۳ چهار نمودار بحث‌شده در زیربخش ۳٫۱ را به‌ترتیب برای مدل با استفاده از روش تبدیل آرک‌سینوس و مدل رگرسیون لوژستیک نشان می‌دهند.

شکل ۲ نشان می‌دهد که مشاهده‌ی ۱۴، هم دارای مانده‌ای زیاد است و هم نسبت به سایر مشاهدات، دارای تأثیری زیاد است. به هر صورت، از آن‌جا که مانده بر انحراف معیار بیش‌براورد شده تقسیم می‌شود تا مانده‌ی پی‌یرسونی به دست آید، ممکن است این روش، توانایی تشخیص داده‌ی دورافتاده را نداشته باشد و داده‌ی دورافتاده‌ای به‌غلط داده‌ای معمولی تلقی شود. شکل ۳ نشان می‌دهد که مشاهده‌ی ۱۴ یک داده‌ی دورافتاده است. البته در این حالت، رگرسیون لوژستیک و بررسی‌های تشخیصی آن، تأکید بیش‌تری بر دورافتاده بودن این مشاهده دارد.

حال که با استفاده از رگرسیون لوژستیک از دورافتاده بودن مشاهده‌ی ۱۴ مطمئن شده‌ایم، دو راه برای

مواجهه با این مشاهده وجود دارد: یا این مشاهده باید از تحلیل خارج شود و بدون آن، تحلیل دوباره انجام شود، که این با مشورت با محقق اصلی باید انجام گیرد، یا این‌که از روش‌های استواری که نسبت به داده‌ی دورافتاده حساس نیستند استفاده شود. اجازه دهید تا در حالت اوّل، تغییر در نتایج را جویا شویم. پس از حذف مشاهده‌ی ۱۴، در مدل با استفاده از روش تبدیل، $R^2 = ۰/۷۲۷$ به دست می‌آید و مانده‌ی کیبش به مقدار ۲۱/۰۰ تقلیل می‌یابد. $\widehat{\sigma} = ۱/۰۵۱$ به دست می‌آید و برازش پس از بررسی نمودارهای تشخیصی، برازش نیکویی را نتیجه می‌دهد. در این روش، همه‌ی متغیرهای تبیینی، معنادارند.

پس از حذف مشاهده‌ی ۱۴، در رگرسیون لوژستیک، مانده‌ی کیبش به مقدار قابل ملاحظه‌ای کاهش یافته، برابر با ۲۵/۲۸۹ می‌شود. در این روش نیز همه‌ی متغیرها معنادارند و نمودارهای تشخیصی، برازش مناسبی را نشان می‌دهند. بنا بر این، در صورت حذف داده‌ی دورافتاده، هر یک از دو روش تبدیل و رگرسیون لوژستیک می‌تواند مورد استفاده قرار گیرد. پس از حذف مشاهده‌ی ۱۴، در روش استوار نیز مانده‌ی کیبش کاهش می‌یابد و به ۲۴/۶۳ می‌رسد و همه‌ی متغیرهای تبیینی معنادار می‌شوند. البته در این روش، متغیر مربوط به بلوک ۱ به‌طور ضعیف معنادار است (پی‌مقدار = ۰/۰۸۰).

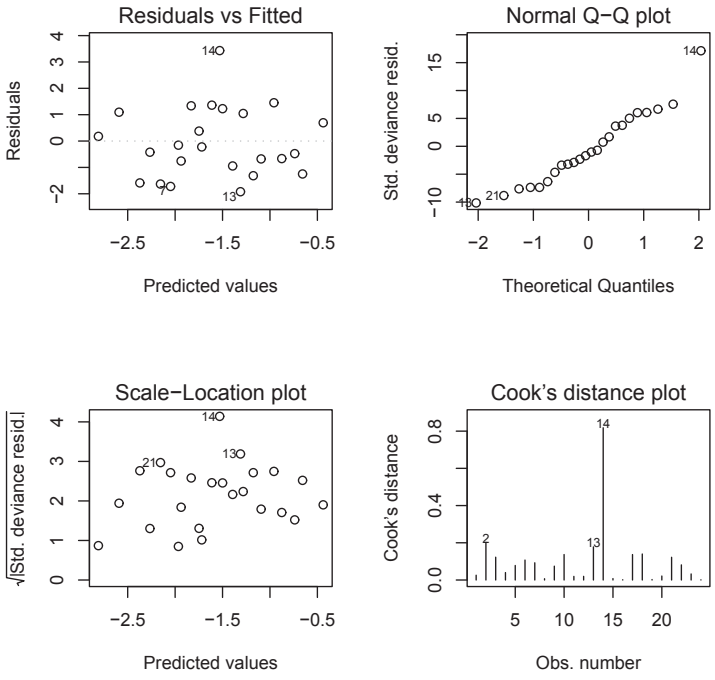


شکل ۲. نمودارهای تشخیصی برای روش تبدیل آرک‌سینوس

در حالتی که داده‌ی دورافتاده در تحلیل باقی بماند، از آن‌جا که در روش استوار، به این داده وزن کم‌تری داده می‌شود، نتایج تحلیل، کم‌تر تحت تأثیر این داده قرار می‌گیرد. وزنِ داده‌شده به آزمودنی iام، $\frac{\Psi_c(r_i)}{r_i}$ است که برای مشاهده‌ی ۱۴، برابر با ۰٫۲۶ و برای بقیه‌ی مشاهدات بین ۰٫۷ تا ۱ است.

۵ نتیجه‌گیری

روش‌های تشخیصیِ دورافتاده بودن مشاهدات در تبدیل آرک‌سینوس، کم‌تر از روش‌های تشخیصیِ رگرسیون لوژستیک، توانایی نشان دادند. این ممکن است به دلیل آن باشد که در نظرگرفتن واریانس مجانبی برای برخی از مشاهدات در بعضی موارد امکان‌پذیر نیست. در هر صورت، وجود مشاهده‌ی دورافتاده باعث می‌شود که روش تبدیل آرک‌سینوس، واریانس را بیش‌براورد کند و این امر نشان‌دهنده‌ی این است که این روش در تشخیص دورافتاده بودن برخی مشاهده‌ها توانایی کم‌تری نشان می‌دهد. اگر پس از تشخیص دورافتاده



شکل ۳. نمودارهای تشخیصی برای روش رگرسیون لوژستیک

بودن داده‌ای، بتوان این داده را از تحلیل خارج کرد، می‌توان با برازش دوباره و بررسی‌های تشخیصی، از نیکویی برازش اطمینان حاصل کرد. در غیر این صورت، استفاده از روش‌های استوارکه نتایج آن‌ها زیاد تحت تأثیر داده‌ی دورافتاده قرار نمی‌گیرد، ترجیح داده می‌شود.

مرجع‌ها

[1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. Wiley, New York.

[2] Anscombe, F.J. (1948). The transformation of Poisson, binomial and negative binomial data. *Biometrika* **35**, 246-254.

[3] Berkson, J. (1957). Tables for the maximum likelihood estimate of the logistic function. *Biometrics* **13**, 28-34.

[4] Box, G.E.P. (1953). Non-normality and tests on variances. *Biometrika* **40**, 318-335.

[5] Cantoni, E.; Ronchetti, E. (2001). Robust inference for generalized linear models. *J. Amer. Statist. Assoc.* **96**, 1022-1030.

[6] Casella, G.; Berger, R.L. (1990). *Statistical Inference.* Brooks/Cole, Pacific Grove, CA.

[7] Draper, N.R.; Smith, H. (1981). *Applied Regression Analysis,* 2nd ed. Wiley, New York.

[8] Finney, D.J. (1952). *Probit Analysis*, 2nd ed. Cambridge University Press, UK.

[9] Finney, D.J. (1964). *Statistical Method in Biological Assay,* 2nd ed. Griffin, London.

[10] Fisher, R.A.; Yates, F. (1938). *Statistical Tables for Biological, Agricultural and Medical Research.* Oliver and Boyd, Edinburgh.

[11] Hampel, F.R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69**, 383-393.

[12] Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73-101.

[13] Imhof, J.P. (1961). Computing the distribution of quadratic forms in normal variables *Biometrika* **48**, 419-426.

[14] McCullagh, P.; Nelder, J.A. (1989). *Generalized Linear Models,* 2nd ed. Chapman and Hall, London.

[15] Nelder, J.A.; Wedderburn, R.W.M. (1972). Generalized linear models. *J. Roy. Statist. Soc.* **A 135**, 370-384.

[16] Phellps, K. (1982). Use of the complementary log-log function to describe dose response relationships in insecticide evaluation field trails. In *Proceedings of the International Conference on Generalized Linear Models*, R. Gilchrist, ed. Springer, New York.

[17] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.

[18] Williams, D.A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Appl. Statist.* **36**, 181-191.

دریافت: ۱۰ مرداد ۱۳۸۳
آخرین اصلاح: ۱۱ آذر ۱۳۸۳
انتشار: ۲۱ دی ۱۳۸۳

مجتبی گنجعلی
گروه آمار، دانشکده‌ی علوم ریاضی،
دانشگاه شهید بهشتی،
بولوار دانشگاه، اوین،
تهران، ایران.
پیام‌نگار: *m-ganjali@cc.sbu.ac.ir*

Analysis of a Problem Using Various Visions

M. Ganjali[†] and L. Latifian[‡]

[†] Shahid Beheshti University
[‡] Central Bank of the Islamic Republic of Iran

Abstract. In this paper an applied problem, where the response of interest is the number of success in a specific experiment, is considered and by various visions is studied. The effects of outlier values of response on results of a regression analysis are so important to be studied. For this reason, using diagnostic methods, outlier response values are recognized. It is shown that use of arc-sine transformation may be misleading in recognizing response outliers. If deleting of outliers is not possible, use of robust modeling approach is suggested. Method of maximum likelihood for estimating parameters in generalized linear model, transformation method and also pseudo-likelihood method are not robust. A method, which is called robust pseudo-likelihood and leads to robust results, is reviewed and a simpler method of computing P-values for model selection is presented. Various approaches for modeling are also compared in the applied example.

Keywords. transformation method; logistic regression; generalized linear models; Pearson residuals; robust pseudo-likelihood.