

مقایسه‌ای بین مدل‌های مورد استفاده در مطالعات مقطعی با پاسخهای دودویی

امید امدادی فر^۱ مجتبی گنجعلی^۲

چکیده

مدل خطی تعمیم یافته^۳ (GLM) برای تحلیل داده‌های دودویی^۴ در مطالعات مقطعی^۵ استفاده می‌شود. مدل لوژستیک، پروبیت و مدل لگ-لگ مکمل از جمله مدل‌های مورد استفاده در GLM ها هستند. روشی دیگر که عموماً استفاده می‌شود، روش شبه درست‌نمایی^۶ است که در آن نیازی به فرض درباره توزیع پاسخ‌ها نیست. هر چند این روش برآوردهای سازگاری برای ضرایب رگرسیونی مهیا می‌کند، در این مقاله با استفاده از شبیه‌سازی و در نظر گرفتن احتمال موفقیت به شرط مقادیر مشاهده شده برای متغیرهای کمکی، نشان داده شده است که روش شبه درست‌نمایی برای تعداد نمونه‌های کم در مقایسه با GLM ها جواب‌های نامناسبی می‌دهد. این روش با روش دیگری که آرانداز^۲ پیشنهاد کرده است، نیز مقایسه شده است. همچنین مدل‌های مختلف در یک مثال عملی بکار برده شده‌اند.

واژه‌های کلیدی: مدل پروبیت، مدل لوژستیک، مدل خطی تعمیم یافته، شبه درست‌نمایی، متغیر پنهان.

۱. مقدمه

همچنین برای برآورد ضرایب رگرسیونی از روشی به نام شبه درست‌نمایی [۶] استفاده می‌شود که فرضی درباره توزیع پاسخها ندارد و تنها رابطه‌ای بین میانگین پاسخها و متغیرهای کمکی و رابطه‌ای بین میانگین و واریانس پاسخها مشخص می‌کند. در مدل بیان شده در [۲] نیز فرضی درباره توزیع پاسخها نداریم ولی در این مدل نسبت به روش شبه درست‌نمایی یک پارامتر بیشتر برآورد می‌شود. در این مقاله روش آرانداز^۲ با روش شبه درست‌نمایی در نمونه کوچک در یک مثال شبیه‌سازی شده مقایسه می‌شوند.

در بخش دوم مدل‌های خطی تعمیم یافته و مدل‌هایی با استفاده از متغیرهای پنهان را معرفی می‌کنیم و خواهیم دید این دو نوع مدل‌بندی،

مطالعات مقطعی، مطالعاتی هستند که در مقطعی از زمان بر روی متغیرهای مورد نظر انجام می‌شوند و هدف در آنها مطالعه تأثیر چند متغیر کمکی بر روی متغیر پاسخ می‌باشد. هنگامی که متغیر پاسخ پیوسته است؛ دسته‌ای از مدل‌های خطی می‌تواند استفاده شود؛ اما هنگامی که متغیر پاسخ دودویی است، مدل‌های خطی مناسب نبوده و مدل‌های غیرخطی مورد استفاده قرار می‌گیرد [۱].

برای پاسخهای دودویی، مدل‌های خطی تعمیم یافته (GLM) ها و مدل‌های غیرخطی‌ای که از مفهوم متغیر پنهان استفاده می‌کنند، مورد استفاده قرار می‌گیرند.

^۴ Binary Data^۵ Cross-Sectional Studies^۶ Quasi Likelihood^۱ کارشناسی ارشد آمار^۲ عضو هیأت علمی دانشگاه شهید بهشتی^۳ Generalized Linear Model

مدلهای مشابهی را نتیجه می دهند. همچنین در این بخش مدل پیشنهادی آراندا اورداز [۲] رایان خواهیم کرد.

در بخش سوم روش شبه درستمایی را که توسط ودریرن [۶] پیشنهاد و توسط مک کولاخ [۵] بسط داده شده است را بیان می کنیم و در بخش چهارم با مثالی شبیه سازی شده نشان می دهیم که روش شبه درستمایی با وجود سازگاری برآوردها، در نمونه های کوچک نتایج مناسبی نخواهد داد. در بخش پنجم در یک مثال عملی روشهای مختلف مقایسه می شوند.

۲. مدل های غیر خطی با پاسخ دودویی

در GLM ها به جای آن که $E(Y|x)$ به صورت تابعی خطی از x ها $(x'\beta)$ در نظر گرفته شود، $h(E(Y|x))$ به صورت تابعی خطی از x در نظر گرفته می شود که h تابعی مشتق پذیر و یکنواست و به آن تابع پیوند گفته می شود.

در داده های دودویی می دانیم که

$$\mu_y = \Pr(Y=1|x) = E(Y|x)$$

در نتیجه

$$\mu_y = h^{-1}(x'\beta) \quad (1)$$

که برد تابع h^{-1} ، فاصله $[0,1]$ است. به عنوان مثالی برای h می توان به لوجیت^۱ که مدل لوزستیک را نتیجه می دهد، اشاره کرد. لوجیت $\Pr(Y=1)$ به صورت

$$\log \frac{\Pr(Y=1)}{1 - \Pr(Y=1)}$$

تعریف می شود. یعنی

$$\mu_y = \frac{\exp(x'\beta)}{1 + \exp(x'\beta)} \quad (2)$$

همچنین اگر h^{-1} را تابع توزیع نرمال استاندارد در نظر بگیریم، مدل پرویت را نتیجه می دهد یعنی:

$$\mu_y = \Phi(x'\beta) \quad (3)$$

که در آن $\phi(\cdot)$ تابع توزیع نرمال استاندارد است.

همچنین اگر h را به صورت زیر در نظر بگیریم، مدل لگ-لگ^۲ به دست خواهد آمد که مدلی نامتقارن می باشد

$$h(\mu_y) = \log\{-\log(\mu_y)\}$$

یعنی مدل لگ-لگ به صورت زیر خواهد بود:

$$\mu_y = \exp\{-\exp(x'\beta)\} \quad (4)$$

متغیر Y^* را پنهان گوئیم اگر Y^* مشاهده نشود و در فاصله $(-\infty, \infty)$ تغییر کند و Y^* ، متغیرهای مشاهده شده Y را به این صورت تولید کند که اگر Y^* مقادیر بزرگ را اختیار کرد آنگاه $Y=1$ و اگر Y^* مقادیر کوچک را گرفت $Y=0$ مشاهده خواهد شد. در فرم ریاضی متغیر Y^* مشاهدات دودویی Y را به صورت زیر تولید می کند:

$$y_i = \begin{cases} 1 & ; y_i^* > \tau \\ 0 & ; y_i^* \leq \tau \end{cases}$$

که در آن τ یک نقطه آستانه^۳ می باشد. برای قابل تشخیص بودن پارامترها فرض می کنیم که $\tau = 0$ [۴].

در مدل های غیر خطی با استفاده از مفهوم متغیر پنهان مدل خطی زیر را در نظر می گیرند:

$$y^* = x'\beta + \varepsilon$$

که در آن y^* متغیر پنهان است [۴]. مدل های مختلف با فرضهای مختلف درباره توزیع خطاها حاصل می شود. به عنوان مثال اگر ε از توزیع $N(0,1)$ در نظر گرفته شود، مدل پرویت و اگر از توزیع لوزستیک با میانه صفر و واریانس $\pi^2/3$ در نظر گرفته شود، مدل لوزستیک را نتیجه می دهد و اگر از توزیع وایبل انتخاب شود، مدل لگ-لگ (۴) نتیجه می شود. در این گونه مدل ها توزیع خطاها به طور کامل مشخص می شود و هیچ پارامتر مجهولی برای توزیع خطاها در نظر گرفته نمی شود و این به دلیل آن است که بتوان ضرایب رگرسیونی را برآورد کرد. در نتیجه مدل غیر خطی زیر به دست می آید:

$$\begin{aligned} \Pr(Y_i = 1 | x_i) &= \Pr(y_i^* > 0 | x_i) \\ &= \Pr(x_i'\beta + \varepsilon_i > 0 | x) \\ &= \Pr(\varepsilon_i > -x_i'\beta | x) \end{aligned}$$

Log-Log Model^۲
Threshold^۳

Link Function^۱
Logit^۴

$$S(\beta) = \sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \beta} \right)' v_i^{-1} \{y_i - \mu_i\} = 0 \quad (6)$$

معادله بالا، همان معادلات برآورد^۱ در GLM ها است، وقتی پیوند لججیت در نظر گرفته شود. همان طور که ملاحظه می‌شود، شبه درستنمایی نیاز به فرضی درباره توزیع پاسخها ندارد و همان طور که بیان شد، ودربرن [۶] و مک کولاخ [۵] ثابت کردند که برآورد ضرایب رگرسیونی در این روش، برآوردهای سازگاری ارائه می‌دهند. حال در مثال زیر با استفاده از شبیه‌سازی نشان داده می‌شود که در استفاده از شبه درستنمایی با حجم نمونه‌های کم، باید دقت بیشتری صورت گیرد. زیرا امکان دارد نتایج مناسب نباشد و همچنین نشان داده می‌شود در این حالت استفاده از روش آراند اورداز که در واقع یک پارامتر اضافه (λ) را برآورد می‌کند ممکن است ترجیح داده شود.

۴. مثال علمی

مقایسه بین مدل‌های مختلف برای پاسخهای دودویی در

مطالعات مقطعی در یک مطالعه شبیه سازی شده

در این مثال بین مدل‌های خطی تعمیم یافته و روش شبه درستنمایی که در بالا بیان شد، بر اساس داده‌های شبیه‌سازی شده، مقایسه‌ای انجام می‌دهیم. هدف نشان دادن این مطلب است که با وجود سازگاری، جوابهای شبه درستنمایی، جوابهای مناسبی برای نمونه‌های با حجم کم ارائه نمی‌دهند و باید دقت بیشتری صورت گیرد.

در این مثال متغیر پنهان y^* به صورت زیر تولید شده است:

$$y^* = 1 + x + e^*$$

که در آن x از توزیع $N(0,1)$ تولید شده است و $e^* = 1 - e$ که e از توزیع کی دو با ۱ درجه آزادی تولید شده است. در نتیجه y ها به صورت

$$y_i = \begin{cases} 1 & ; y^* > 0 \\ 0 & ; y^* \leq 0 \end{cases}$$

به دست می‌آیند. با استفاده از مدل زیر

$$\pi(x) = \Pr(Y = 1 | x) = F(\beta_0 + \beta_1 x)$$

که این احتمال را بر اساس توزیع مشخصی که برای خطاها در نظر گرفته می‌شود می‌توان حساب کرد. مثلاً برای مدل پرویت داریم

$$\begin{aligned} \Pr(Y_i = 1 | x_i) &= \Pr(\varepsilon_i \leq x_i' \beta | x_i) \\ &= \Phi(x_i' \beta) \end{aligned}$$

همان طور که ملاحظه می‌شود این مدل همان مدل (۲) است.

مدل لوژستیک نیز در این حالت همان مدل (۲) خواهد بود. به عنوان مثالی دیگر می‌توان ε_i را دارای توزیع χ^2 با U درجه آزادی در نظر گرفت که در این حالت بر غیر متقارن بودن توزیع ε ها تأکید می‌شود.

آراند اورداز [۲] تبدیلهای توانی از $\Pr(Y = 1)$ را در نظر گرفت که تمام مدل‌های متقارن و نامتقارن را تنها در یک شکل نشان می‌دهد. او $h(\mu_y)$ را به صورت زیر در نظر گرفته است:

$$h(\mu_y) = \log\left(\frac{(1 - \mu_y)^{-\lambda} - 1}{\lambda}\right) \quad (5)$$

یعنی مدل $h(\mu_y) = x' \beta$ را برای تحقیق تأثیر متغیرهای تبیینی (x) بر متغیر وابسته y در نظر گرفته است. این مدل برای $\lambda = 1$ به مدل لوژستیک و برای $\lambda \rightarrow 0$ به مدل لگ-لگ مکمل^{*} [۱] تبدیل می‌شود که یک مدل نامتقارن است. بنابراین این دو مدل تنها به وسیله یک پارامتر λ ممکن است مقایسه شوند.

۳. روش شبه درستنمایی

شبه درستنمایی ([۵] و [۶]) روشی برای برآورد پارامترهای رگرسیونی است که در آن به فرض درباره توزیع متغیر وابسته (فرضی که در GLM ها ضروری است) نیازی نیست و تنها رابطه بین میانگین متغیر پاسخ و متغیرهای کمکی و همچنین رابطه بین واریانس و میانگین متغیر پاسخ نیاز است. در نتیجه در روش شبه درستنمایی تنها فرضیهایی که اختیار می‌شود، فرضیهای زیر هستند:

$$h(\mu_{y_i}) = x_i' \beta, \text{var}(Y_i) = v(\mu_{y_i})$$

که در آن $v(\cdot)$ یک تابع معلوم و مشخص است.

در این روش برآورد ضرایب رگرسیونی جواب معادله زیر است:

ممکن است مقادیری از x وجود داشته باشند که مدل‌های ذکر شده $\Pr(Y=1|x)$ را نزدیک به هم برآورد کنند.

حال سوالی که مطرح می‌شود این است که تابع توزیع مناسب را چگونه انتخاب کنیم. جواب این است که مدلی که توسط آراندا اورداز [۲] پیشنهاد شده است عمل کرده و ضرایب رگرسیونی و λ را بر اساس درستنمایی برآورد کنیم. نتایج برای داده‌های شبیه سازی شده برای مدل آراندا اورداز در جدول (۱) آمده است. همان گونه که مشاهده می‌شود، این مدل برای داده‌های با حجم نمونه کم نسبت به مدل‌های پرویت، لوژستیک و شبه درستنمایی، برآوردی نسبتاً معقول ارائه می‌دهد. در این حالت $\hat{\lambda} = 1/0.514$ برآورد شده است.

۵. اثر دزهای متفاوت گاز دی سولفید کربن بر روی سوسکها

داده‌های جدول ۲، که توسط اگرستی [۱] ارائه شده است، تعداد سوسکهای کشته شده بعد از ۵ ساعت قرار گرفتن آنها در معرض گاز دی سولفید کربن، در غلظتهای متفاوت را نشان می‌دهد. غلظت برحسب لگ دز بیان شده است. اگر پراکنش نسبت سوسکهای کشته شده را در مقابل لگ دزها رسم کنیم [۱]، مشاهده می‌شود که یک رابطه S -شکل خواهیم داشت که همانند توزیعهای نرمال و لوژستیک متقارن نیست. برای برآورد مناسب باید دنبال توزیعهای نامتقارن باشیم. در این مثال مدل‌های پرویت، لوژستیک، شبه درستنمایی با پیوند پرویت، لگ لگ مکمل و همچنین مدل پیشنهادی توسط آراندا اورداز را برآش داده ایم که نتایج آنها در جدول ۳ آمده است. در این جدول تعداد سوسکهای برآورد شده تحت مدل‌های مختلف را نشان می‌دهد. همان گونه که ملاحظه می‌شود مدل شبه درستنمایی نتایج واقعاً گمراه کننده‌ای می‌دهند. اگر پیوند لگ لگ مکمل را نیز برای شبه درستنمایی در نظر بگیریم، بازهم نتایج مشابه با پیوند پرویت می‌دهد. دو مدل پرویت و لوژستیک برای ۲ و ۳ مقدار اول لگ دزها، نتایج را کم برآورد می‌کنند، مدل لگ لگ مکمل که توسط اگرستی [۱] برای این داده‌ها انتخاب شده است و مدل آراندا اورداز مقادیر برآش داده شده را نزدیک مقادیر مشاهده شده برآورد کرده‌اند. در مدل آراندا اورداز $\hat{\lambda} = -0.00172$ برآورد شده است که با مقدار تئوری آن که به صفر میل می‌کند، برابر است.

که در آن $F(0)$ می‌تواند هر یک از توابع توزیع لوژستیک نرمال و کی دو با ۱ درجه آزادی باشد، استفاده کرده‌ایم. همچنین از روش شبه درستنمایی با تابع پیوند پرویت $(\pi(x) = \Phi(\beta_0 + \beta_1 x))$ ، برآوردهای β_0 و β_1 و در نتیجه $\pi(x)$ ها را به دست آوردیم. می‌دانیم که:

$$\pi(x) = \Pr(y^* > 0 | x)$$

می‌تواند با استفاده از تابع توزیعهای کی دو، لوژستیک و نرمال برآورد شود. با توجه به این که تابع پیوندهای کی نیست، نمی‌توان بر اساس β ها عملکرد آنها را مقایسه کرد و نتیجه گرفت کدام بهتر عمل کرده‌اند. اما بر اساس برآوردهایی که برای $\pi(x)$ به دست می‌آوریم و با توجه به این که در مدل‌های با پاسخ دودویی $\Pr(Y=1|x)$ با تغییر فرضها برای خطاها تغییر نمی‌کند [۴]، می‌توان عملکرد مدل‌های مختلف را مقایسه کرد. برای تمام مدلها $\Pr(Y=1|x=2)$ را در نظر گرفته‌ایم. باید توجه داشت که برای مدل تولید شده داریم:

$$\pi(2) = F_{\chi(df=1)}(4) = 0.954$$

مقایسه این مدلها را با استفاده از نرم افزار R (نرم افزار آماری و گرافیکی که اولین بار توسط *Ross Ihaka* و *Robert Gentleman* نوشته شده و در سایت WWW.ci.tuwien.ac.at/hornik/R/ قابل دسترس است)، با حجم نمونه‌ای ۱۰۰ و ۳۰ تایی و با ۵۰۰ تکرار، برای هر مدل انجام داده‌ایم. نتایج در جدول (۱) آمده است. همان طور که ملاحظه می‌شود شبه درستنمایی، لوژستیک و پرویت، $\pi(x)$ را با تعداد نمونه‌های زیاد ($n=100$) بیش برآورد می‌کنند $0.954 > \hat{\pi}$ و با تعداد نمونه کم ($n=30$) برآوردی که برای روش شبه درستنمایی به دست می‌آید برآوردی غیرقابل انتظار است. دوباره مدل پرویت و مدل لوژستیک با تعداد نمونه‌های کم احتمال را بیش برآورد می‌کنند و نسبت به نمونه‌های بزرگ برآوردهای قابل قبولی ارائه نمی‌دهند. همان گونه که ملاحظه می‌شود و انتظار می‌رفت مدل غیرخطی با استفاده از مفهوم متغیر پنهان با تابع پیوند کی دو با تعداد نمونه‌های زیاد و کم، جواب قابل قبولی می‌دهد (احتمال نزدیک ۰/۹۵۴ برآورد شده است). در نتیجه در روش شبه درستنمایی که فرضی برای توزیع پاسخها نداریم، امکان دارد برای نمونه‌های کوچک برآوردهای نامناسبی به دست آوریم. البته ما $\Pr(Y=1|x)$ را تنها به ازای $x=2$ که یک نقطه انتهایی برای متغیر کمکی است، برای مدل‌های مختلف مقایسه کرده‌ایم.

۶. نتیجه گیری

بهتر است از مدل آراندا اورداز استفاده شود با وجود آن که یک پارامتر بیشتر برآورد می‌کند. همچنین ملاحظه می‌شود مدل‌های GLM با حجم نمونه‌های کم از روش شبه درستنمایی بهتر عمل می‌کنند.

همان طور که در مثال بالا نشان داده شده استفاده از روش شبه درستنمایی با حجم نمونه‌های کم ممکن است جوابهای معقولی ندهد و

جدول ۱: برآورد $\pi(2)$ تحت مدل‌های در نظر گرفته شده (برآورد $(Pr(Y=1|x=2))$)

N	۳۰	۱۰۰
روش شبه درستنمایی	۱/۰۰۰	۰/۹۸۴
مدل پرویت	۰/۹۹۸	۰/۹۸۷
مدل لوژستیک	۰/۹۹۶	۰/۹۸۱
مدل با تابع پیوند معکوس تابع توزیع کی دو	۰/۹۵۴	۰/۹۵۴
مدل آراندا اورداز	۰/۹۲۶	۰/۹۳۹

جدول ۲: داده‌های اثر دزهای متفاوت دی سولفید کربن بر روی سوسکها

لگ دز	۱/۸۸۴	۱/۸۶۱	۱/۸۳۷	۱/۸۱۱	۱/۷۸۴	۱/۷۵۵	۱/۷۲۴	۱/۶۹۱
تعداد سوسکها	۶۰	۶۲	۵۹	۶۳	۵۶	۶۲	۶۰	۵۹
تعداد کشته شده‌ها	۶۰	۶۱	۵۳	۵۲	۲۸	۱۸	۱۳	۶

جدول ۳: نتایج برازش مدل‌ها بر حسب برآورد سوسکهای کشته شده

آراندا اورداز	لگ لک - مکمل	شبه درستنمایی	لوژستیک	پرویت	لگ دز
۵/۶۵۹	۵/۶۵۴	۰	۳/۵۰۳	۳/۴۰۷	۱/۶۹۱
۱۱/۲۷۸	۱۱/۲۸۲	۰	۹/۸۲۰	۱۰/۶۸۶	۱/۷۲۴
۲۰/۹۴۱	۲۰/۹۴۲	۰	۲۲/۴۲۱	۲۳/۴۳۷	۱/۷۵۵
۳۰/۳۳۲	۳۰/۳۳۸	۵۶	۳۳/۸۷۵	۳۳/۷۸۱	۱/۷۸۴
۴۷/۶۷۱	۴۷/۶۷۸	۶۳	۵۰/۰۴۷	۴۹/۵۵۶	۱/۸۱۱
۵۴/۱۸۶	۵۴/۱۸۵	۵۹	۵۳/۳۳۹	۵۳/۳۶۸	۱/۸۳۷
۶۱/۱۲۰	۶۱/۱۱۶	۶۲	۵۹/۲۳۹	۵۹/۶۸۱	۱/۸۶۱
۵۹/۹۴۹	۵۹/۹۴۸	۶۰	۵۸/۷۵۵	۵۹/۲۳۹	۱/۸۸۴

مراجع

- [1] Agresti, A., 1990, *Categorical Data Analysis*, New York: John Wiley.
- [2] Aranda-Ordaz, F.J., 1981, *On Two Families of Transformations to Additivity for Binary Response Data*, *Biometrika*, 88, 357- 363.
- [3] Diggle, p.J., Liang, K.Y. and Zeger, S.I., 1996, *Analysis of Longitudinal Data*, Oxford: Oxford University Press.
- [4] Long, S.J., 1997, *Regression Models for Categorical Dependent Variables*, London: SAGE.
- [5] McCullagh, P., 1983, *Quasi Likelihood Functions*, *The Annals of Statistics*, 11, 59-67.
- [6] Wedderburn, R.W.M., 1974, *Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss - Newton Method*, *Biometrika*, 61, 439- 447.

زندگی‌ام را روی ریختن تاسی گذرانده‌ام و خطر مرگ را پذیرا هستم.

ویلیام شکسپیر