# 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 داده های دودویی [٤] درمطالعات مقطعی [٥] استفاده می شود. مدل لوژستیک، پروبیت و مدل لگ-لگ مکمل از جملهٔ مدل های مورد استفاده در $GLM$ ها هستند. روشی دیگر که عموماً استفاده می شود، روش شبه درستنمایی [٦] است که در آن نیازی به فرض درباره توزیع پاسخ ها نیست. هر چند این روش برآوردهای سازگاری برای ضرایب رگرسیونی مهیا می کند، دراین مقاله با استفاده از شبیه سازی و در نظر گرفتن احتمال موفقیت به شرط مقادیر مشاهده شده برای متغیرهای کمکی، نشان داده شده است که روش شبه درستنمایی برای تعداد نمونه های کم در مقایسه با $GLM$ ها جواب های نامناسبی می دهد. این روش با روش دیگری که آراندا اورداز [۲] پیشنهاد کرده است، نیز مقایسه شده است. همچنین مدل های مختلف دریک مثال عملی بکار برده شده اند.

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


## ۱. مقدمه

مطالعات مقطعی، مطالعاتی هستند که در مقطعی از زمان بر روی متغیرهای مورد نظر انجام می شوند و هدف در آنها مطالعهٔ 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.

همچنین برای برآورد ضرایب رگرسیونی از روشی به نام شبه درستنمایی [٦] استفاده می‌شود که فرضی دربارهٔ 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 دربارهٔ توزیع پاسخها نداریم ولی دراین مدل نسبت به روش شبه درستنمایی یک پارامتر بیشتر برآورد می‌شود. دراین مقاله روش آراندا اورداز [۲] با روش شبه درستنمایی درنمونه کوچک دریک مثال شبیه سازی شده مقایسه می‌شوند.

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

---

[۱] کارشناسی ارشد آمار

[۲] عضو هیأت علمی دانشگاه شهید بهشتی

[۳] *Generalized Linear Model*

[٤] *Binary Data*

[٥] *Cross-Sectional Studies*

[٦] *Quasi Likelihood*

مدلهای مشابهی را نتیجه می‌دهند. همچنین در این بخش مدل پیشنهادی آراندا اورداز [۲] رابیان خواهیم کرد.

دربخش سوم روش شبه درستنمایی را که توسط ودربرن [۶] پیشنهاد و توسط مک‌کولاخ [۵] بسط داده شده است را بیان می‌کنیم و در بخش چهارم با مثالی شبیه سازی شده نشان می‌دهیم که روش شبه درستنمایی با وجود سازگاری برآوردها، در نمونه‌های کوچک نتایج مناسبی نخواهد داد. دربخش پنجم در یک مثال عملی روشهای مختلف مقایسه می‌شوند.

## ۲. مدلهای 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 می‌شود که $h$ تابعی مشتق‌پذیر و یکنواست و به آن تابع پیوند [۱] گفته می‌شود.

در داده‌های دودویی می‌دانیم که

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

در نتیجه

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

که برد تابع $h^{-1}$، فاصله $[0,1]$ است. به عنوان مثالی برای $h$ می‌توان به لوجیت [۲] که مدل لوژستیک را نتیجه می‌دهد، اشاره کرد. لوجیت $\Pr(Y=1)$ به صورت

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

تعریف می‌شود. یعنی

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

همچنین اگر $h^{-1}$ را تابع توزیع نرمال استاندارد در نظر بگیریم، مدل پروبیت را نتیجه می‌دهد یعنی:

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

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

همچنین اگر $h$ را به صورت زیر در نظر بگیریم، مدل لگ-لگ [۳] به دست خواهد آمد که مدلی نامتقارن می‌باشد

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

یعنی مدل لگ-لگ به صورت زیر خواهد بود:

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

متغیر $Y^*$ را پنهان گوییم اگر $Y^*$ مشاهده نشود و در فاصلهٔ $(-\infty,\infty)$ تغییر کند و $Y^*$، متغیرهای مشاهده شدهٔ $Y$ 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. درفرم ریاضی متغیر $Y^*$ مشاهدات دودویی $Y$ را به صورت زیر تولید می‌کند:

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

که در آن $\tau$ یک نقطه آستانه [۴] می‌باشد. برای قابل تشخیص بودن پارامترها فرض می‌کنیم که $\tau = 0$ [۴].

درمدلهای غیرخطی با استفاده از مفهوم متغیر پنهان مدل خطی زیر را در نظر می‌گیرند:

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

که در آن $y^*$ متغیر پنهان است [۴]. مدلهای مختلف با فرضهای مختلف دربارهٔ توزیع خطاها حاصل می‌شود. به عنوان مثال اگر $\varepsilon$ از توزیع $N(0,1)$ در نظر گرفته شود، مدل پروبیت و اگر از توزیع لوژستیک با میانگین صفر واریانس $\pi^2/3$ درنظر گرفته شود، مدل لوژستیک را نتیجه می‌دهد و اگر از توزیع وایبل انتخاب شود، مدل لگ-لگ (۴) نتیجه می‌شود. در این گونه مدلها توزیع خطاها به طور کامل مشخص می شود و هیچ پارامتر مجهولی برای توزیع خطاها در نظر گرفته نمی‌شود و این به دلیل آن است که بتوان ضرایب رگرسیونی را برآورد کرد. درنتیجه مدل غیرخطی زیر به دست می‌آید:

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

---

[۱] Link Function

[۲] *Logit*

[۳] *Log-Log Model*

[۴] *Threshold*

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

$$\Pr(Y_i = ۱ \mid x_i) = \Pr(\varepsilon_i \le x_i'\beta \mid x_i)$$
$$= \Phi(x_i'\beta)$$

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 $\varepsilon_i$ را دارای توزیع $\chi^۲$ با $\upsilon$ درجهٔ آزادی در نظر گرفت که در این حالت بر غیر متقارن بودن توزیع $\varepsilon$ها تأکید می‌شود.

آراندا ارداز [۲] تبدیلهای توانی از $\Pr(Y=۱)$ را درنظر 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:

$$h(\mu_y) = \log\left(\frac{(۱-\mu_y)^{-\lambda} - ۱}{\lambda}\right) \qquad (۵)$$

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 $\lambda = ۱$ به مدل لوژستیک و برای $\lambda \to ۰$ به مدل لگ- لگ مکمل [۵] [۱] تبدیل می‌شود که یک مدل نامتقارن است. بنابراین این دو مدل تنها به وسیلهٔ یک پارامتر $\lambda$ ممکن است مقایسه شوند.

## ۳. روش شبه درستنمایی

شبه درستنمایی ([۵ و ۶]) روشی برای برآورد پارامترهای رگرسیونی است که در آن به فرض دربارهٔ توزیع متغیر وابسته (فرضی که در $GLM$ ها ضروری است) نیازی نیست و تنها رابطهٔ بین میانگین متغیر پاسخ و متغیرهای کمکی و همچنین رابطهٔ بین واریانس و میانگین متغیر پاسخ نیاز است. در نتیجه در روش شبه درستنمایی تنها فرضهایی که اختیار می شود، فرضهای زیر هستند:

$$h(\mu_{y_i}) = x_i'\beta, \upsilon ar(Y_i) = \upsilon(\mu_{y_i})$$

که درآن $\upsilon(.)$ یک تابع معلوم و مشخص است.

دراین روش برآورد ضرایب رگرسیونی جواب معادله زیر است:

$$S(\beta) = \sum_{i=۱}^{n}\left(\frac{\partial \mu_i}{\partial \beta}\right)' \upsilon_i^{-۱}\{y_i - \mu_i\} = ۰ \qquad (۶)$$

معادلهٔ بالا، همان معادلات برآورد[۶] در $GLM$ ها است، وقتی پیوند لوجیت در نظر گرفته شود. همان طور که ملاحظه می‌شود، شبه درستنمایی نیاز به فرضی دربارهٔ 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 آراندا ارداز که در واقع یک پارامتر اضافه $(\lambda)$ را برآورد می کند ممکن است ترجیح داده شود.

## ۴. مثال علمی

### مقایسه بین مدلهای مختلف برای پاسخهای دودویی در مطالعات مقطعی در یک مطالعهٔ 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:

$$y^* = ۱ + x + e^*$$

که در آن $x$ از توزیع $N(۰,۱)$ تولید شده است و $e^* = ۱ - e$ که $e$ از توزیع کی دو با ۱ درجه آزادی تولید شده است. در نتیجه $y$ها به صورت

$$y_i = \begin{cases} ۱ & ; y^* > ۰ \\ ۰ & ; y^* \le ۰ \end{cases}$$

به دست می‌آیند. با استفاده از مدل زیر

$$\pi(x) = \Pr(Y = ۱ \mid x) = F(\beta_۰ + \beta_۱ x)$$

---

[۵] *Compelemently Log- Log*

[۶] *Estimation Equations*

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

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

می‌تواند با استفاده از تابع توزیعهای کی دو، لوژستیک و نرمال برآورد شود. با توجه به این که تابع پیوندها یکی نیست، نمی‌توان بر اساس $\beta$ 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 $\Pr(Y=1\mid x)$ با تغییر فرضها برای خطاها تغییر نمی‌کند[٤] ، می‌توان عملکرد مدلهای مختلف را مقایسه کرد. برای تمام مدلها $\Pr(Y=1\mid x=2)$ را در نظر گرفته‌ایم. باید توجه داشت که برای مدل تولید شده داریم:

$$\pi(2)=F_{\chi(df=1)}(4)=0/954$$

مقایسهٔ این مدلها را با استفاده از نرم افزار R (نرم افزار آماری و گرافیکی که اولین بار توسط *Robert Gentleman* و *Ross Ihaka* 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 هرمدل انجام داده‌ایم. نتایج در جدول (۱) آمده است. همان طور که ملاحظه می‌شود شبه درستنمایی، لوژستیک و پروبیت، $\pi(x)$ را با تعداد نمونه‌های زیاد $(n=100)$ بیش برآورد می‌کنند $\hat{\pi}>0/954$ و با تعداد نمونهٔ کم $(n=30)$ برآوردی که برای روش شبه درستنمایی به دست می‌آید برآوردی غیرقابل انتظار است. دوباره مدل پروبیت 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 ۰/۹۵٤ برآورد شده است).

درنتیجه در روش شبه درستنمایی که فرضی برای توزیع پاسخها نداریم، امکان دارد برای نمونه‌های کوچک برآوردهای نامناسبی به دست آوریم. البته ما $\Pr(Y=1\mid x)$ را تنها به ازای $x=2$ که یک نقطهٔ انتهایی برای متغیر کمکی است، برای مدلهای مختلف مقایسه کرده‌ایم.

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

حال سوالی که مطرح می شود 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 $\lambda$ 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 پروبیت، لوژستیک و شبه درستنمایی، برآوردی نسبتا معقول ارائه می‌دهد. دراین حالت $\hat{\lambda}=1/0514$ 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 ـ شکل خواهیم داشت که همانند توزیعهای نرمال و لوژستیک متقارن نیست. برای برآورد مناسب باید دنبال توزیعهای نامتقارن باشیم. در این مثال مدلهای پروبیت، لوژستیک، شبه درستنمایی با پیوند پروبیت، لگ لگ مکمل و همچنین مدل پیشنهادی توسط آراندا اورداز را برازش 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 واقعا گمراه کننده‌ای می‌دهند. اگر پیوند لگ لگ مکمل را نیز برای شبه درستنمایی در نظر بگیریم، بازهم نتایج مشابه با پیوند پروبیت می‌دهد. دو مدل پروبیت و لوژستیک برای ۲ و۳ 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 برازش داده شده را نزدیک مقادیر مشاهده شده برآورد کرده‌اند. درمدل آراندا اورداز $\hat{\lambda}=-0/00172$ برآورد شده است که با مقدار تئوری آن که به صفر میل می‌کند، برابر است.

## ۶. نتیجه گیری

همان طور که درمثال بالا نشان داده شده استفاده از روش شبه درستنمایی با حجم نمونه‌های کم ممکن است جوابهای معقولی ندهد و 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 می کنند.

**جدول ۱: برآورد $\pi(۲)$ تحت مدلهای در نظر گرفته شده(برآورد $\Pr(Y=۱\,|\,x=۲)$)**

| N | ۳۰ | ۱۰۰ |
|---|---|---|
| روش شبه درستنمایی | ۱/۰۰۰ | ۰/۹۸۴ |
| مدل پروبیت | ۰/۹۹۸ | ۰/۹۸۷ |
| مدل لوژستیک | ۰/۹۹۶ | ۰/۹۸۱ |
| مدل با تابع پیوند معکوس تابع توزیع کی‌دو | ۰/۹۵۴ | ۰/۹۵۴ |
| مدل آراندا ارداز | 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 برحسب برآورد سوسکهای کشته شده**

| آراندا اورداز | لگ لگ- مکمل | شبه درستنمایی | لوژستیک | پروبیت | 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,* Biomertrika, 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