تأثیر نادیده گرفتن همبستگی در داده های طولی یا زوج شده بر استنباط پارامترها

مجتبی گنجعلی[۱] جواد قاسمیان[۲]

چکیده

در مواقعی که اندازه‌گیری از یک صفت خاص، روی داده‌های طولی[۳] (دیگل[۴] و دیگران، ۱۹۹۴) یا زوج شده[۵] (اگرستی[۶]، ۱۹۹۴) درمقاطع زمانی مختلف انجام می‌دهیم، بحث همبستگی بین داده‌ها مطرح می‌شود. دراین مقاله، با استفاده از داده‌های دودویی که درآن متغیر پاسخ، نشان دهندهٔ داشتن یا نداشتن آسم است، نشان می‌دهیم که درنظر گرفتن یا نگرفتن این همبستگی، چه تأثیری بر نتایج برآورد پارامترها و انحراف استاندارد برآوردکننده‌های پارامترهای منتسب با متغیرهای توصیفی، دارد.

واژه‌های کلیدی: پاسخهای دودویی، مدل پروبیت دو متغیره، متغیر پنهان، متغیر توصیفی مانا، آسم.

۱. مقدمه

داده‌های طولی در آزمایشهایی رخ می‌دهند که یک دنباله از اندازه‌گیریها، بر روی تعدادی آزمودنی در زمانهای مختلف به دست می‌آید. هرچند که واحدهای مختلف ازهم مستقلند، ولی پیشرفتهای اخیر در تحلیل این گونه داده‌ها، تأکید بر استفاده از مدلهایی دارند که وابستگی بین اندازه‌گیریهای متعلق به آزمودنی یکسان را درنظر می‌گیرند. در این مقاله نشان می‌دهیم که درنظر نگرفتن این وابستگی، هر چند روی برآورد پارامترها تأثیری ندارد ولی موجب بیش یا کم برآورد کردن خطای استاندارد برآوردکننده‌های پارامترهای مدل می‌شود.

دربخش بعد، داده‌های دودویی داشتن یا نداشتن آسم، [۶] و دربخش ۳ مدل پروبیت دو متغیره را معرفی کرده‌ایم. دربخش ۴ تابع درستنمایی بیان و در نهایت، دربخش ۵، تحلیل آماری داده‌های بخش ۲، با و بدون در نظر گرفتن همبستگی آورده شده است.

[۱] گروه آماری دانشگاه شهید بهشتی

[۲] دانشجوی کارشناسی ارشد، دانشگاه شهید بهشتی

[۳] *Longitudinal Data*

۲. داده‌های آسم

دراین مقاله ازداده‌های مربوط به آسم استفاده کرده‌ایم. این داده‌ها از ٦ شهرستان منطقهٔ هاروارد اهایو جمع آوری شده اند، به این ترتیب که ۷۰٦ پسر و ۷۱۳ دختر سفیدپوست را در ۹ سالگی و دوباره در ۱۳ سالگی از نظر ابتلا یا عدم ابتلا به آسم مورد مطالعه قرار داده اند [٦]. می‌خواهیم تأثیر زمان و جنس را بر احتمال داشتن آسم، از نظر آماری آزمون کنیم. از ۷۰٦ پسر ۱۴۹ نفر و از ۷۱۳ دختر ۱۲۳ نفر در ۱۳ سالگی، از دادن پاسخ مورد علاقه (داشتن یا نداشتن آسم) امتناع کردند. دریک تحلیل جداگانه ملاحظه شد که مکانیزم گم شدن دراین داده‌ها، کاملاً تصادفی است[٤]؛ بنابراین تحلیل داده‌های کامل (که در آن تنها آزمودنیهایی در نظر گرفته می‌شوند که هر دو پاسخ آنها مشاهده

[۵] *Paired Data*

[٦] *Agresti*

[٤] *Diggle*

شده است)، برای استنباط آماری کافی است.

جدول ۱ داده‌های کامل و بدون مقادیر گمشده برای متغیرهای پاسخ در این داده‌ها را نشان می دهد. همان طور که جدول ۱ نشان می دهد ۵۵۷ نفر پسر و ۵۹۰ نفر دختر به هر دو متغیر پاسخ (داشتن یا نداشتن آسم در ۹ و ۱۳ سالگی) جواب داده‌اند. با استفاده از این داده‌ها، نشان می‌دهیم که درنظر گرفتن همبستگی پاسخها در سن ۹ و ۱۳ سالگی، چه تأثیری بر استنباط پارامترها و خطای استاندارد برآوردکننده‌های پارامترها دارد.

۳. مدل مورد استفاده

برای بررسی تأثیر سن و زمان بر متغیر پاسخ (داشتن آسم)، مدل زیر را درنظر می گیریم :

$$y_{it}^{*} = a + bG_i + cI_{i\{age=13\}} + \varepsilon_{it} \qquad (1)$$

وقتی $t = 1,2$ ، که در آن برای آزمودنی iام، G جنس و $I_{\{age=13\}}$ به صورت زیر است:

$$I_{\{age=13\}} = \begin{cases} 1 & \text{اگر فرد ۱۳ ساله باشد،} \\ 0 & \text{اگر فرد ۱۳ ساله نباشد،} \end{cases}$$

و ε_{it} برای $t = 1,2$ خطای اندازه گیری متغیر y_{it}^{*} است. به متغیرهای y_{i1}^{*} و y_{i2}^{*} ، متغیرهای پنهان می گوییم. این متغیرها خود قابل مشاهده نیستند ولی باعث می‌شوند که پاسخهای گسسته قابل مشاهده باشند. برای مثال داشتن یا نداشتن آسم یک آزمودنی خاص (y_i) به علت وجود متغیر پیوسته دیگری (y_i^{*}) است که خود مشاهده نشده است ولی وقتی از آستانهٔ خاصی گذر کند، فرد دچار بیماری آسم می‌شود. دراین جا مقدار آستانه، صفر فرض شده است. در حالتی که مقدار آستانه مثلاً مقدار d باشد، می‌توان از متغیر $y_i^{*} - d$ به عنوان متغیر پنهان استفاده کرد. بنابراین ثابت d ، قابل برآورد کردن نیست [۵].

چون داده‌های مشاهده شده، در دو حالت وجود دارند، لذا برای $t = 1,2$ ، متغیرهای مشاهده شدهٔ y_{it} به صورت زیر تعریف می‌شوند:

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

y_{i1}^{*} مربوط به سن ۹ سالگی و y_{i2}^{*} مربوط به سن ۱۳ سالگی است. در ۱۳ سالگی متغیر $I_{\{age=13\}}$ ، وارد مدل شده است که پارامتر منتسب به آن (c) نشان دهندهٔ تأثیر زمان بر ابتلا یا عدم ابتلا به آسم است.

دراین مدل a ، b و c پارامترهای مدل می‌باشند که باید برآورد شوند. درحالت کلی خطاهای ε_{i1} و ε_{i2} وابسته‌اند و

$$\mathrm{var}(\varepsilon_{it}) = 1; t = 1,2$$

(در داده‌های دودویی، واریانس دلخواه برای خطاها، در معادلهٔ (۱) قابل برآورد نیست [۵]) و

$$\mathrm{cov}(\varepsilon_{i1}, \varepsilon_{i2}) = \rho .$$

در حالت کلی پارامتر ρ باید برآورد شود. اگر فرض کنیم $\rho = 0$ ، می‌توان نشان داد چه اتفاقی در برآورد پارامترها و خطاهای استاندارد برآوردکننده‌های پارامتر مدل، روی می‌دهد.

۴. تابع درستنمایی

برای نیل به هدف نهایی، دو حالت کلی را درنظر می گیریم. ابتدا دو متغیر Y_{i1} و Y_{i2} را مستقل فرض کرده و درمرحلهٔ بعد همبستگی بین این دو متغیر را در نظر می گیریم. تابع درستنمایی برای مدل کلی که همبستگی را در نظر می گیرد به صورت زیر است:

$$L(a,b,c,\rho \mid y_{i1}, y_{i2}, G) = \qquad (2)$$

$$\prod_{i=1}^{n} P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}) =$$

$$\prod_{i=1}^{n} p_{00}^{(1-y_{i1})(1-y_{i2})} p_{01}^{(1-y_{i1})y_{i2}} p_{10}^{y_{i1}(1-y_{i2})} p_{11}^{y_{i1}y_{i2}}$$

که در آن:

$$p_{00} = P(Y_1 = 0, Y_2 = 0) = \Phi_2(-a-bG, -a-bG-c, \rho),$$

$$p_{01} = P(Y_1 = 0, Y_2 = 1) = \Phi_2(-a-bG, a+bG+c, -\rho),$$

$$p_{10} = P(Y_1 = 1, Y_2 = 0) = \Phi_2(a+bG, -a-bG-c, -\rho),$$

$$p_{11} = P(Y_1 = 1, Y_2 = 1) = \Phi_2(a+bG, a+bG+c, \rho),$$

واضح است که

$$\sum_{i=0}^{1}\sum_{j=0}^{1} \rho_{ij} = 1 .$$

پارامتر مربوط به ρ همبستگی بین پاسخ در ۹ سالگی و ۱۳ سالگی است که باید برآورد شود. این پارامتر در زمان برآورد در فاصله ۱- تا ۱ محدود می‌شود. Φ_2 تابع توزیع نرمال دو متغیره را نشان می‌دهد، که به صورت زیر تعریف می شود:

$$\Phi_2(q_1,q_2,\rho)=\int_{-\infty}^{q_1}\int_{-\infty}^{q_2} f(x_1,x_2)dx_1dx_2$$

که درآن $f(x_1,x_2)$ تابع چگالی نرمال دو متغیرهٔ استاندارد شده است. از آنجا که تابع *ms* در *S-Plus* مجموعه‌ای از توابع غیرخطی را برای برآورد پارامترها با استفاده از روش بهینه سازی شبه- نیوتن[۲] کمینه می‌کند، برای برآورد پارامترها، منهای لگاریتم تابع درستنمایی را با استفاده از تابع ms کمینه کرده‌ایم.

۵. نتایج استفاده از دو مدل

برای آزمون معنی‌داری پارامترها، دو حالت $\rho=0$ و $\rho\neq 0$ را درنظر گرفته، نتایج را در جدول (۲) خلاصه می‌کنیم. همان طور که برآورد پارامترها در مدل کامل (عدم استقلال) نشان می‌دهد، داده‌ها شواهد کافی بر تأثیر زمان در ابتلا به آسم دارند (۰/۰۰۱ = p مقدار).

همچنین در این مدل داده‌ها نشان می‌دهند که جنس در ابتلا به آسم تأثیر دارد ولی این تأثیر چندان قوی نیست (۰/۰۴۳ = p مقدار).

مقایسه بین مدل استقلال و عدم استقلال حاکی از اهمیت برآورد ρ است. هرچند برآورد پارامترها دراین دو مدل به هم نزدیک‌اند ولی به طور کلی خطای استاندارد برآوردها دردو مدل متفاوت است. خطای استاندارد پارامترهایی که متغیر توصیفی متناظر با آنها با زمان، تغییر نمی‌کنند (متغیر توصیفی مانا)، مانند b دراین مثال، تمایل به کم برآورد شدن دارند (کم برآورد شدن p مقدار را نتیجه می‌دهد) و خطای استاندارد پارامترهایی که متغیر توصیفی متناظر با آنها با زمان تغییر می‌کند (متغیر توصیفی نامانا)، مانند c دراین مثال، تمایل به بیش برآورد شدن دارند(زیاد برآورد شدن p مقدار را نتیجه می‌دهد).

مانده‌هایی که با $\rho=0$ به دست می آیند، فرض استقلال بین دو پاسخ را مدنظر قرار می‌دهند. اگر $\rho\neq 0$ مانده‌ها نیز باید به گونه‌ای تصحیح شوند که همبستگی بین پاسخها را در نظر گیرید. بررسی مانده‌ها به عنوان کار بعدی پیشنهاد می‌شود.

جدول ۱: داده‌های مربوط به آسم

وضعیت آسم			۱۳ سالگی: دارد	۱۳ سالگی: ندارد	جمع کل
پسرها	۹ سالگی	دارد	۲۲	٦	۲۸
		ندارد	۱۵	۵۱٤	۵۲۹
		جمع کل	**۳۷**	**۵۲۰**	**۵۵۷**
دخترها	۹ سالگی	دارد	۱۳	۳	۱٦
		ندارد	۱۳	۵٦۱	۵۷٤
		جمع کل	**۲٦**	**۵٦٤**	**۵۹۰**

جدول ۲: برآورد پارامترها و p مقدار

پارامترها	p مقدار	مدل استقلال	p مقدار	مدل عدم استقلال
a	-	-۱/۹۰۰	-	-۱/۸۹۰
b	۰/۰۰۹	۰/۲۴۰	۰/۰۴۳	۰/۲۳۰
c	۰/۰۵۷۰	۰/۱۷۰	۰/۰۰۱	۰/۱۷۰
d	-	-	۰/۰۰۰	۰/۹۳۰
ρ				

مراجع:

[1] Agresti,A., 1990, *Categorical Data Analysis*. New York, John Wiley.

[2] Chambers, J.M. and Hastie, T.J., 1992, *Statistical Models in S*, Chapter 10: Nonlinear models, Pacific Groves, CA: Wadsworth Brooks /Cole.

[3] Diggel, P.J., Liang, K. and Zeger, S., 1996, *Analysis of Longitudinal Data*, Oxford Science publication.

[4] Little, R.J. and Rubin, D., 1987, *Statistical Analysis with Missing Data*, New York, Wiley.

[5] Long , S.J., 1997, *Regression Models for Categorical and Limited Dependent Variables*, London, SAGE.

[6] Ronitzky, A. and Wypij, D., 1994, *A Note on the Bias of Estimastion with Missing Data*, Biometrics 147, 87-99.

علم آمار تحقق اندیشهٔ انسان در گشودن درهای بررسی غیر ملموس حقایق است.