

معادله برآورد وزنی جهت تحلیل مقادیر گم‌شده متغیر مستقل در رگرسیون خطی و کاربرد آن در داده‌های بیماری پوکی استخوان

سیدمجتبی طباطبایی‌راد* (M.Sc)، حمید علوی مجد (Ph.D)، علی‌اکبر خادم‌معبودی (Ph.D)

دانشگاه علوم پزشکی شهید بهشتی، دانشکده پیراپزشکی، گروه آمار زیستی

چکیده

سابقه و هدف: در تحقیقات مختلف پزشکی گاهی اوقات مشاهدات، اندازه‌گیری نشده و یا مقدار آن‌ها در دسترس نمی‌باشد؛ که به آن‌ها داده‌های گم شده اطلاق می‌شود. در بسیاری از استنباط‌های آماری، چشم‌پوشی از این گونه داده‌ها می‌تواند باعث اربابی در نتایج گردد. لذا معرفی روش‌های آماری جهت تحلیل چنین داده‌هایی ضروری به نظر می‌رسد.

مواد و روش‌ها: یکی از روش‌های تعیین پارامترهای معادله رگرسیون، زمانی که مقادیر گم‌شده در متغیرهای مستقل وجود دارد، استفاده از روش معادله برآورد وزنی (WEE) می‌باشد. در این روش با استفاده از رگرسیون لجستیک، احتمال مشاهده شدن برای داده‌های گم‌شده محاسبه و عکس آن، به‌عنوان وزن در مشتق تابع درست‌نمایی یا به عبارت دیگر آماره امتیاز (Score statistics) وارد می‌شود و معادله آخر با استفاده از الگوریتم EM حل و پارامترهای معادله رگرسیون برآورد می‌گردد. نکته اصلی در این روش عدم نیاز به تعیین دقیق توزیع متغیر گم‌شده می‌باشد. روش فوق در یک مثال عملی با روش حداکثر درست‌نمایی (Maximum likelihood) مقایسه گردید.

یافته‌ها: نتایج نشان داد در صورتی که گم شده مقادیر به‌صورت تصادفی باشد، روش WEW در مقایسه با روش‌های معمول (روش Complete case و حداکثر درست‌نمایی) کارایی بالاتری دارد؛ همچنین با توجه به معادله رگرسیون به دست آمده، سن و برداشت رحم اثر کاهش‌دهنده بر روی تراکم معدنی استخوان دارند و افزایش شاخص توده بدنی تأثیر مثبت بر روی تراکم استخوان دارد.

نتیجه‌گیری: با توجه به مزایای روش معادله برآورد وزنی (عدم نیاز به تعیین توزیع متغیر گم‌شده و سادگی معادلات برآورد) چنان‌چه توزیع متغیر گم شدن به صورت تصادفی باشد و احتمال مشاهده شدن متغیر گم‌شده دقیق محاسبه گردد، استفاده از این روش توصیه می‌گردد.

واژه‌های کلیدی: رگرسیون خطی ساده و لجستیک، گم شدن تصادفی، معادله برآورد وزنی، الگوریتم EM، پوکی استخوان

مقدمه

رگرسیون خطی است. برای برآورد پارامترهای یک مدل رگرسیونی روش‌های مختلفی نظیر روش ماکزیم درست‌نمایی حداقل مربعات وزنی، روش‌های بیزی، روش‌های معادلات

گم شدن مقادیر داده‌های مستقل پدیده رایجی در تحلیل

* نویسنده مسئول. تلفن: ۰۹۱۲۲۵۰۷۹۱۶، نمابر: ۰۷۶۴-۴۴۲۱۶۳۵، E-mail: tabatabaei_rad@yahoo.com

مواد و روش‌ها

در یک مدل رگرسیون خطی با بردار $x_i = (x_{i1}, \dots, x_{ip})'$ که همیشه مشاهده شده‌اند و بردار $z_i = (z_{i1}, \dots, z_{im})'$ از m متغیر که در بعضی از نمونه‌ها مقادیر گم شده دارند، هدف برآورد ضرایب رگرسیون خطی $\beta' = [\beta_0, \beta'_x, \beta'_z]$ در معادله رگرسیون زیر می‌باشد.

(۱)

$$\mu_i = E[y_i | x_i, z_i] = \beta_0 + \beta'_x x'_i + \beta'_z z'_i$$

حال متغیر نشان‌گر R_i به صورت زیر تعریف می‌شود:

$$R_i = \begin{cases} 1 & \text{اگر } Z_i \text{ مشاهده شده باشد} \\ 0 & \text{اگر } Z_i \text{ گم شده باشد} \end{cases}$$

$$\pi_i = P(R_i = 1 | y_i, x_i, z_i)$$

چون مکانیسم گم شدن MAR در نظر گرفته شده است،

پس R_i وابستگی به مقادیر Z ندارد [۱]، در نتیجه

$$\pi_i = P(R_i = 1 | y_i, x_i)$$

با استفاده از رگرسیون لجستیک احتمال مشاهده شدن Z ‌ها

به صورت زیر محاسبه می‌شود:

$$\pi_i = \pi_i(\omega) = \frac{\exp(-\omega' m_i)}{1 + \exp(-\omega' m_i)} \quad (2)$$

که در آن $m_i = (y_i, x_i)'$ و ω برداری از

پارامترهاست (ضرایب رگرسیون لجستیک)؛ که به وسیله رگرسیون لجستیک معمولی از داده‌های مشاهده شده برآورد می‌شود.

معادله لگاریتم حداکثر درست‌نمایی داده‌ها به صورت زیر

می‌باشد:

$$\sum_{i=1}^n \left\{ r_i \left(\frac{d}{d_{(\beta, \alpha, \omega)}} \log [P(r_i, y_i, z_i | x_i, \alpha, \beta, \omega)] \right) + (1 - r_i) \left(\frac{d}{d_{(\beta, \alpha, \omega)}} \log [P(r_i, y_i | x_i, \alpha, \beta, \omega)] \right) \right\} = 0 \quad (3)$$

پس از برآورد π_i ، آن را در معادله (۳) وارد نموده که معادله برآورد وزنی (WEE) به فرم ذیل به دست می‌آید:

$$S(\beta, \alpha, \omega) = \sum_{i=1}^n \left\{ \frac{r_i}{\pi_i} \left(\frac{d}{d_{(\beta, \alpha)}} \log [P(y_i, z_i | x_i, \beta, \alpha)] \right) + \left(1 - \frac{r_i}{\pi_i} \right) \left(\frac{d}{d_{(\beta, \alpha)}} \log [P(y_i | x_i, \alpha, \beta)] \right) \right\} = 0 \quad (4)$$

اگر π_i به درستی مشخص شود، می‌توان برآوردهای مناسب

تعمیم یافته و روش‌های نیمه پارامتری به کار می‌رود. رایج‌ترین تکنیکی که توسط تحلیل‌کننده‌ها به کار می‌رود، این است که به طور ساده نمونه‌هایی را که دارای مقادیر گم شده هستند از مطالعه خارج کنند و آن‌گاه یک تحلیل رگرسیون را با داده‌های باقی مانده انجام دهند. در این روش به دلیل این‌که نمونه‌های با مقادیر گم شده از مطالعه خارج می‌شوند احتمال دارد نتایج برآوردها، اریب و شدیداً نامناسب باشد [۱].

در این مطالعه جهت افزایش کارایی و کاهش اریبی در برآورد پارامترهای رگرسیونی، یک معادله برآورد وزنی که بر اساس درست‌نمایی نرمال چند متغیره است، معرفی می‌شود. معادله برآورد وزنی برای اولین بار توسط Robins و همکاران [۲] (۱۹۹۴) و Lipsitz [۳] (۱۹۹۶) و همچنین Ibrahimi [۴] (۱۹۹۶) به کار برده شد. در این مقالات در تحلیل رگرسیون زمانی که گم شدن مقادیر متغیرهای مستقل به صورت MAR (Missing at random) یا MCAR (Missing completely at random) بود، استفاده شد. در این مقالات فرض می‌شود توزیع شرطی $(y|x, z)$ نرمال است و توزیع Z به طور دقیق مشخص می‌گردد و مدل برای احتمالات مشاهده شدن Z ‌ها با استفاده از رگرسیون لجستیک برآورد می‌شود. بر این اساس نشان داده شد که این معادلات برآوردهایی نااریب و کاراتر از برآوردهای به دست آمده از روش حداکثر درست‌نمایی، ایجاد می‌کنند. در این مقاله روش برآورد وزنی در چارچوب Parzen و همکاران [۵] مورد استفاده قرار می‌گیرد.

سپس معادله فوق را با استفاده از الگوریتم EM [۶] یا

روش نیوتن رافسون حل نموده و پارامترها به دست می‌آیند.

و ناریب از β را با روش WEE به دست آورد [۳، ۵]. در حالی که برای استفاده از روش ML باید توزیع $P(z_i | x_i)$ دقیقاً تعیین شود، اما برای روش WEE چنین نیست. به طور خلاصه برای این که WEE برآوردهای مناسبی ارائه کند، باید مدل برای π_i و μ_i به درستی تعیین شود، اما $P(z_i | x_i)$ می تواند غیردقیق باشد. برای روش ML، $P(y_i | z_i, x_i, \beta)$ و $P(z_i | x_i)$ باید به درستی تعیین شوند، اما مدل برای π_i نیاز نیست. در واقع جهت استفاده از WEE، $P(z_i | x_i)$ نرمال، فرض می شود حتی اگر z_i ها گسسته (اسمی، رتبه ای و شمارشی) باشند.

مثال کاربردی مربوط به داده های بیماری پوکی استخوان. برای نشان دادن کارایی روش WEE یک مطالعه مشاهده ای مقطعی در مورد ۹۷ نفر از زنان مراجعه کننده به مرکز سنجش تراکم استخوان بیمارستان میلاد در تابستان سال ۸۲ که به طور تصادفی ساده انتخاب شدند، انجام گرفت. در این بررسی می خواستیم میزان تراکم معدنی استخوان BMD (Bone mineral density) مهره های کمری در زنان ۳۰-۷۰ سال را پیش بینی کنیم. متغیرهای جمع آوری شده با توجه به مقالات Garnero و همکاران (۱۹۹۵) [۷] و هم چنین Prior و Kirland (۲۰۰۱) [۸]، عبارت بودند از:

BMD. این متغیر میزان تراکم معدنی استخوان در مهره های کمر (در واحد گرم بر سانتی متر مربع) را نمایش می دهد، که به وسیله اشعه ایکس با دستگاه سنجش تراکم استخوان اندازه گیری شده است (متغیر وابسته).

سن. سن افراد مورد مطالعه بر حسب سال (متغیر مستقل). شاخص توده بدنی BMI (Body mass index). شاخص توده بدنی در افراد که حاصل تقسیم وزن به کیلوگرم بر مجذور قد به متر می باشد (متغیر مستقل).

LD-HD. مدت زمان مصرف قرص های ضدبارداری بر حسب سال (متغیر مستقل).

Hysterectomy. متغیر دوحالته که نشان دهنده

برداشت یا عدم برداشت رحم می باشد (۰- برداشت رحم نداشته و ۱- برداشت رحم داشته)، نسبت افرادی که هیستریکتومی کرده اند برابر با ۳۹/۲٪ می باشد (متغیر مستقل) در ابتدا معادله رگرسیون با استفاده از کل داده ها که هیچ مشاهده گم شده ای وجود نداشت (استفاده از اطلاعات کامل ۹۷ سطر) پیش بینی گردید.

جهت نشان دادن روش های برآورد، تعدادی (۳۳٪) از مقادیر متغیر هیستریکتومی را به صورت تصادفی (MAR) حذف نمودیم. با توجه به این که داده های گم شده مربوط به افرادی بود که سن آن ها زیر ۵۳ سال بود، مکانیسم گم شدن به صورت تصادفی (MAR) در نظر گرفته شد. لازم به ذکر است که چنانچه گم شدن مقادیر یک متغیر مستقل به متغیر مستقل دیگری وابسته باشد، گم شدن به صورت تصادفی (MAR) می باشد و اگر گم شدن به هیچ کدام از متغیرها وابستگی نداشته باشد، مکانیسم گم شدن به صورت کاملاً تصادفی (MCAR) در نظر گرفته می شود.

در ابتدا با استفاده از رگرسیون لجستیک، توزیع π_i به صورت زیر تعیین شد:

$$\text{Logit}(\pi_i) = -69.43 - (0.71 \text{ BMD}) + (1.5 \text{ Age})$$

با توجه به این که ضرایب BMI و LD-HD غیرمعنی دار بودند، در معادله لجستیک وارد نگردیدند. سپس جهت استفاده از روش WEE در Proc Iml نرم افزار SAS برنامه ای نوشته شد [۹، ۱۰] و برآورد پارامترها و انحراف معیار آن ها از ۳ روش ذیل با استفاده از رگرسیون به دست آمد:

۱) روش CC (Complete case). که در آن، سطرهای متناظر با داده های گم شده حذف گردید و با ۶۵ سطر موجود، تحلیل انجام گرفت.

۲) روش ML. که روش مورد استفاده در Proc MI [۹] نرم افزار SAS می باشد.

۳) روش WEE. که در آن از معادله برآورد وزنی معرفی شده، جهت برآورد پارامترها استفاده شد.

نتایج

میانگین و انحراف معیار متغیرهای اندازه‌گیری شده در جدول ۱ آورده شده است.

جدول ۱. میانگین و انحراف معیار متغیرهای اندازه‌گیری شده

متغیر	میانگین	انحراف معیار
BMD	۰/۸۸۲ gr/cm ²	۰/۱۶۱
Age	۵۴/۲ سال	۱۱
BMI	۲۸/۲ kg/m ²	۴/۵
LD-HD	۲/۹ سال	۲/۸

ضرایب برآورد شده معادله رگرسیون و انحراف معیار آن‌ها بر اساس ۴ روش در جدول ۲ نمایش داده شده است. همچنین مقدار P متناظر با هر پارامتر در داخل پرانتز آورده شده است. با استفاده از مقادیر جدول ۲ دیده می‌شود که برآوردهای پارامترها و انحراف معیار آن‌ها برای دو رهیافت WEE و ML خیلی مشابهند و فقط در برآورد β_z مشاهده می‌شود که نتایج به‌دست آمده از روش WEE نزدیک‌تر به

جدول ۲. برآورد ضرایب مدل به همراه انحراف معیار و P-Value آن‌ها با روش‌های مختلف

روش	β_0	β_1 (AGE)	β_2 (BMI)	β_3 (LD-HD)	β_z (Hister)
کل داده‌ها	$1 \pm 0/095$ (p = 0/000)	$-0/008 \pm 0/001$ (p = 0/000)	$0/0124 \pm 0/003$ (p = 0/000)	$-0/0047 \pm 0/003$ (p = 0/106)	$-0/5 \pm 0/026$ (p = 0/059)
CC	$0/712 \pm 0/15$ (p = 0/000)	$-0/005 \pm 0/002$ (p = 0/008)	$0/0155 \pm 0/003$ (p = 0/000)	$-0/0031 \pm 0/003$ (p = 0/301)	$-0/0341 \pm 0/031$ (p = 0/197)
ML	$0/99 \pm 0/097$ (p = 0/000)	$-0/008 \pm 0/001$ (p = 0/000)	$0/0125 \pm 0/003$ (p = 0/000)	$-0/0047 \pm 0/003$ (p = 0/106)	$-0/044 \pm 0/029$ (p = 0/077)
WEE	$0/99 \pm 0/097$ (p = 0/000)	$-0/008 \pm 0/001$ (p = 0/000)	$0/0124 \pm 0/003$ (p = 0/000)	$-0/0047 \pm 0/003$ (p = 0/106)	$-0/047 \pm 0/026$ (p = 0/063)

مورد بحث و بررسی قرار گرفت و در نهایت در سال ۲۰۰۲ توسط Parzen و همکاران روش فوق برای استفاده در رگرسیون خطی با داده‌های مستقل گم‌شده بسط داده شد و شرایط ساده‌تری جهت استفاده از این معادلات بیان شد [۵]. به طور خلاصه در معادلات برآورد وزنی ابتدا از داده‌های کامل استفاده شد و احتمال این‌که متغیر گم‌شده به طور کامل مشاهده شود، توسط رگرسیون لجستیک محاسبه گردید،

بحث و نتیجه‌گیری

در تحقیق صورت گرفته هدف ما معرفی روش معادله برآورد وزنی و مقایسه آن با روش حداکثر درست‌نمایی و همچنین کاربرد روش فوق در تحلیل داده‌ها با استفاده از نرم‌افزارهای آماری می‌باشد. این معادلات برای اولین بار توسط Robinz و همکاران (۱۹۹۴) معرفی شد [۲] و سپس توسط Zhao و همکاران (۱۹۹۶) [۳] در حالت‌های مختلف

گم شده در یکی از متغیرها داشت به کار برده شد و پارامترهای معادله رگرسیون برآورد گردید، که معادله به دست آمده نشان داد، افزایش سن و برداشت رحم اثر کاهش دهنده و افزایش شاخص توده بدنی تأثیر مثبت بر روی تراکم استخوان دارد. هم چنین نتایج با داده های کامل مقایسه شد و به نظر می رسد که انحراف معیارهای حاصل از روش WEE کوچک تر از دو روش دیگر (ML و CC) می باشد و هم چنین روش WEE شرایط ساده تری (عدم نیاز به تعیین توزیع متغیر گم شده) نسبت به دو روش دیگر دارد. با این وجود برای رسیدن به نتایج قطعی انجام یک مطالعه شبیه سازی شده ضرورت دارد.

منابع

- [1] Little JA, Rubin B. (Editors). Statistical analysis with missing data. 2nd ed. New York: John Wiley & Sons, 2002.
- [2] Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Statist Assoc, 1994; 89:846-66.
- [3] Zhao LP, Lipsitz S, Lew D. Regression analysis with missing covariate data using estimating equations. Biometrics, 1996; 52(4):1165-82.
- [4] Lipsitz SR, Ibrahim JG. A conditional model for incomplete covariates in parametric regression models. Biometrika, 1996; 83(4):916-22.
- [5] Parzen M, Lipsitz SR, Ibrahim JG, Lipschultz S. A weighted estimating equation for linear regression with missing covariate data. Stat Med, 2002; 21(16):2421-36.
- [6] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Royal Statist Soc (Methodological), 1977; 39:1-22.
- [7] Garnero P, Sornay-Rendu E, Delmas PD. Decreased bone turnover in oral contraceptive users. Bone, 1995; 16(5):499-503.
- [8] Prior JC, Kirkland SA, Joseph L, Kreiger N, Murray TM, Hanley DA, et al. Oral contraceptive use and bone mineral density in premenopausal women: cross-sectional, population-based data from the Canadian Multicentre Osteoporosis Study. CMAJ, 2001; 165(8):1023-9.
- [9] طباطبایی راد سیدمجتبی. معادله برآورد وزنی (WEE) در رگرسیون خطی با داده های گم شده. پایان نامه کارشناسی ارشد، تهران: دانشگاه علوم پزشکی شهید بهشتی، ۱۳۸۳.
- [10] SAS Institute. What's new in Sas software: Release 8.1 and release 8.2. SAS Publishing: CD-Rom edition, 2001.

سپس عکس این احتمال، به عنوان وزن در مشتق تابع درست نمایی یا به عبارت دیگر آماره نمره وارد و معادله ایجاد شده که معادله برآورد وزنی نام دارد با استفاده از الگوریتم EM حل شد. آن گاه با استفاده از برآوردهای به دست آمده، پارامترهای مدل رگرسیونی برآورد گردید. نکته اصلی در این روش، عدم نیاز به تعیین دقیق توزیع متغیر گم شده می باشد و می توان آن را نرمال فرض نمود.

در این مطالعه از یک معادله برآورد وزنی که توسط Parzen و همکاران (۲۰۰۲) معرفی شده بود استفاده گردید [۵]. نتایج مطالعه نشان داد که WEE بیان شده، خواصی مشابه معادله برآورد ML برای یک توزیع نرمال چند متغیره دارد و هم چنین این معادلات برآوردی، برخلاف روش های دیگر نیازی به مونت کارلو یا انتگرال عددی ندارند. زمانی که کوواریت های گم شده توزیع غیر نرمال دارند، WEE بحث شده، در مقایسه با روش های ارائه شده توسط Robinz و همکاران (۱۹۹۴) کارایی بیشتری دارد [۲]. بر اساس نتایج به دست آمده از مطالعات شبیه سازی شده توسط Parzen و همکاران (۲۰۰۲) و به خاطر سادگی معادلات برآورد شده، احساس می شود که می توان از روش WEE در برخورد با هر نوع متغیر گم شده (پیوسته و گسسته) استفاده کرد و کارایی آن نیز زیاد خواهد بود. با توجه به مطالعات انجام شده چنین نتیجه گیری می شود که WEE نسبت به ML کارا تر است، زیرا با استفاده از آن می توان برآوردهای ثابتی را زمانی که $(z|x)$ P نامشخص و π_i مشخص است، به دست آورد. نتایج فوق در مطالعه بررسی عوامل مؤثر بر پوکی استخوان که حدود ۳۳٪