مجله علمی دانشگاه علوم پزشکی سمنان- جلد ۷، شماره ۳ و ٤، بهار و تابستان ۱۳۸۵

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 علوی‌مجد (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 اتلاق می‌شود. در بسیاری از استنباط‌های آماری، چشم‌پوشی از این گونه داده‌ها می‌تواند باعث اریبی 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 (Weighted estimating equation) 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 WEE 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، گم‌شدن تصادفی، معادله برآورد وزنی، الگوریتم EM، پوکی استخوان

مقدمه

گم شدن مقادیر داده‌های مستقل پدیده رایجی در تحلیل رگرسیون خطی است. برای برآورد پارامترهای یک مدل رگرسیونی روش‌های مختلفی نظیر روش ماکزیم درست‌نمایی حداقل مربعات وزنی، روش‌های بیزی، روش‌های معادلات

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

تاریخ دریافت: ۸۳/۷/۲۱؛ تاریخ پذیرش: ۸۴/۹/۸

تعمیم یافته و روش‌های نیمه‌پارامتری به کار می‌رود. رایج‌ترین تکنیکی که توسط تحلیل‌کننده‌ها به کار می‌رود، این است که به طور ساده نمونه‌هایی را که دارای مقادیر گم‌شده 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 گم‌شده از مطالعه خارج می‌شوند احتمال دارد نتایج برآوردها، اریب و شدیداً نامناسب باشد [۱].

در این مطالعه جهت افزایش کارایی و کاهش اریبی دربرآورد 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 (۱۹۹۴) [۲]، Zho و Lipsitz (۱۹۹۶) [۳] و هم‌چنین Ibrahim و Lipsitz (۱۹۹۶) [۴] به‌کار برده شد. در این مقالات در تحلیل رگرسیون زمانی که گم شدن مقادیر متغیرهای مستقل به‌صورت 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 به‌طور دقیق مشخص می‌گردد و مدل برای احتمالات مشاهده شدن Zها با استفاده از رگرسیون لجستیک برآورد می‌شود. بر این اساس نشان داده شد که این معادلات برآوردهایی ناریب و کاراتر از برآوردهای به‌دست آمده از روش حداکثر درست‌نمایی، ایجاد می‌کنند. در این مقاله روش برآورد وزنی در چارچوب Parzen و همکاران [۵] مورد استفاده قرار می‌گیرد.

مواد و روش‌ها

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

(۱)

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

حال متغیر نشان‌گر R_i به صورت زیر تعریف می‌شود:

$$R_i = \begin{cases} ۱ & \text{اگر } Z_i \text{ مشاهده شده باشد} \\ ۰ & \text{اگر } Z_i \text{ گم شده باشد} \end{cases}$$

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

چون مکانیسم گم شدن MAR درنظر گرفته شده است، پس R_i وابستگی به مقادیر z ندارد [۱]، در نتیجه

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

با استفاده از رگرسیون لجستیک احتمال مشاهده شدن Zها به صورت زیر محاسبه می‌شود:

$$\pi_i = \pi_i(\omega) = \frac{\exp(-\omega' m_i)}{1+\exp(-\omega' m_i)} \quad (۲)$$

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 به‌صورت زیر می‌باشد:

$$\sum_{i=1}^{n}\left\{r_i\left(\frac{d}{d_{(\beta,\alpha,\omega)}}\log[P(r_i,y_i,z_i|x_i,\alpha,\beta,\omega)]\right)+(1-r_i)\left(\frac{d}{d_{(\beta,\alpha,\omega)}}\log[P(r_i,y_i|x_i,\alpha,\beta,\omega)]\right)\right\}=0 \quad (۳)$$

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

$$S(\beta,\alpha,\omega)=\sum_{i=1}^{n}\left\{\frac{r_i}{\pi_i}\left(\frac{d}{d_{(\beta,\alpha)}}\log[P(y_i,z_i|\,x_i,\beta,\alpha)]\right)+(1-\frac{r_i}{\pi_i})\left(\frac{d}{d_{(\beta,\alpha)}}\log[P(y_i|\,x_i,\alpha,\beta)]\right)\right\}=0 \quad (۴)$$

سپس معادلهٔ فوق را با استفاده از الگوریتم EM [۶] یا روش نیوتن رافسون حل نموده و پارامترها به دست می‌آیند.

اگر π_i به درستی مشخص شود، می‌توان برآوردهای مناسب

و ناااریب 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 $P(z_i|x_i)$ می‌تواند 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 می‌شود حتی اگر z_iها گسسته (اسمی، رتبه‌ای و شمارشی) باشند.

مثال کاربردی مربوط به داده‌های بیماری پوکی استخوان

برای نشان دادن کارایی روش WEE یک مطالعه مشاهده‌ای 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 می‌خواستیم میزان تراکم معدنی استخوان BMD (Bone mineral density) مهره‌های کمری در زنان ۷۰-۳۰ سال را پیش‌بینی کنیم. متغیرهای جمع‌آوری شده با توجه به مقالات Garnero و همکاران (۱۹۹۵) [۷] و همچنین Prior و Kirland (۲۰۰۱) [۸]، عبارت بودند از:

BMD. این متغیر میزان تراکم معدنی استخوان در مهره‌های کمر (در واحد گرم بر سانتی‌متر مربع) را نمایش می‌دهد، که به‌وسیلهٔ اشعه ایکس با دستگاه سنجش تراکم استخوان اندازه‌گیری 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 می‌باشد (متغیر مستقل).

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

Hysterectomy. متغیر دوحالته که نشان دهندهٔ برداشت یا عدم برداشت رحم می‌باشد (۰- برداشت رحم نداشته و ۱- برداشت رحم داشته)، نسبت افرادی که هیسترکتومی کرده‌اند برابر با ۳۹/۲٪ می‌باشد (متغیر مستقل)

در ابتدا معادله رگرسیون با استفاده از کل داده‌ها که هیچ مشاهده گم‌شده‌ای وجود نداشت (استفاده از اطلاعات کامل ۹۷ سطر) پیش‌بینی گردید.

جهت نشان دادن روش‌های برآورد، تعدادی (۳۳٪) از مقادیر متغیر هیسترکتومی را به‌صورت تصادفی (MAR) حذف نمودیم. با توجه به این‌که داده‌های گم‌شده مربوط به افرادی بود که سن آن‌ها 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) می‌باشد و اگر گم شدن به هیچ کدام از متغیرها وابستگی نداشته باشد، مکانیسم گم شدن به صورت کاملاً تصادفی (MCAR) در نظر گرفته می‌شود.

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 غیرمعنی‌دار بودند، در معادله لجستیک وارد نگردیدند. سپس جهت استفاده از روش WEE در Proc Iml نرم‌افزار SAS برنامه‌ای نوشته شد [۹،۱۰] و برآورد پارامترها و انحراف معیار آن‌ها از ۳ روش ذیل با استفاده از رگرسیون به‌دست آمد:

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

۲) روش ML. که روش مورد استفاده در Proc MI [۹] نرم‌افزار SAS می‌باشد.

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	انحراف‌معیار
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 انحراف‌معیار 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 انحراف‌معیار 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 حالت کل داده‌ها می‌باشد. در روش CC می‌بینیم که انحراف معیارهای بزرگتری نسبت به روش‌های دیگر در مورد β_0 و β_1 و β_z داریم. همان‌طور که مشاهده می‌شود برای این داده‌ها استفاده از روش WEE، ساده‌تر و کاراتر از دو روش دیگر می‌باشد؛ با توجه به این که نیاز به تعیین توزیع متغیر گم‌شده نیست. هم‌چنین در شبیه‌سازی‌های انجام شده توسط Parzen و همکاران (۲۰۰۲) [۵]، کاراتر بودن روش WEE در مقایسه با روش‌های دیگر مشخص شد.

معادله رگرسیون به‌دست آمده از داده‌های کامل به‌صورت ذیل می‌باشد:

BMD=1-(0.008Age)+(0.012BMI)-(0.5Hysterectomy)

همان‌طور که از معادله فوق دیده می‌شود، افزایش سن و برداشت رحم اثر کاهش‌دهنده بر روی تراکم معدنی استخوان دارند و افزایش شاخص توده بدنی تأثیر مثبت بر روی تراکم معدنی استخوان دارد.

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	۱ ± ۰/۰۹۵ (p = ۰/۰۰۰)	−۰/۰۰۸ ± ۰/۰۰۱ (p = ۰/۰۰۰)	۰/۰۱۲۴ ± ۰/۰۰۳ (p = ۰/۰۰۰)	−۰/۰۰۴۷ ± ۰/۰۰۳ (p = ۰/۱۰۶)	−۰/۵ ±۰/۰۲۶ (p = ۰/۰۵۹)
CC	۰/۷۱۲ ± ۰/۱۵ (p = ۰/۰۰۰)	−۰/۰۰۵ ± ۰/۰۰۲ (p = ۰/۰۰۸)	۰/۰۱۵۵ ± ۰/۰۰۳ (p = ۰/۰۰۰)	−۰/۰۰۳۱ ± ۰/۰۰۳ (p = ۰/۳۰۱)	−۰/۰۳۴۱ ± ۰/۰۳۱ (p = ۰/۱۹۷)
ML	۰/۹۹ ± ۰/۰۹۷ . (p = ۰/۰۰۰)	−۰/۰۰۸ ± ۰/۰۰۱ (p = ۰/۰۰۰)	۰/۰۱۲۵ ± ۰/۰۰۳ (p = ۰/۰۰۰)	−۰/۰۰۴۷ ± ۰/۰۰۳ (p = ۰/۱۰۶)	−۰/۰۴۴ ± ۰/۰۲۹ (p = ۰/۰۷۷)
WEE	۰/۹۹ ± ۰/۰۹۷ . (p = ۰/۰۰۰)	−۰/۰۰۸ ± ۰/۰۰۱ (p = ۰/۰۰۰)	۰/۰۱۲۴ ± ۰/۰۰۳ (p = ۰/۰۰۰)	−۰/۰۰۴۷ ± ۰/۰۰۳ (p = ۰/۱۰۶)	−۰/۰۴۷ ± ۰/۰۲۶ (p = ۰/۰۶۳)

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 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،

سپس عکس این احتمال، به عنوان وزن در مشتق تابع درست‌نمایی یا به عبارت دیگر آماره نمره وارد و معادله ایجاد شده که معادله برآورد وزنی نام دارد با استفاده از الگوریتم EM حل شد. آن‌گاه با استفاده از برآوردهای به‌دست آمده، پارامترهای مدل رگرسیونی برآورد گردید. نکته اصلی در این روش، عدم نیاز به تعیین دقیق توزیع متغیر گم‌شده می‌باشد و می‌توان آن‌را 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 هم‌چنین این معادلات برآوردی، برخلاف روش‌های دیگر نیازی به مونت‌کارلو یا انتگرال عددی ندارند. زمانی که کوواریت‌های گم‌شده توزیع غیرنرمال دارند، WEE بحث شده، در مقایسه با روش‌های ارائه شده توسط Robinz و همکاران (۱۹۹۴) کارایی بیش‌تری دارد [۲]. بر اساس نتایج به‌دست آمده از مطالعات شبیه‌سازی شده توسط Parzen و همکاران (۲۰۰۲) و به خاطر سادگی معادلات برآورد شده، احساس می‌شود که می‌توان از روش WEE در برخورد با هر نوع متغیرگم‌شده (پیوسته و گسسته) استفاده کرد و کارایی آن نیز زیاد خواهد بود. با توجه به مطالعات انجام شده چنین نتیجه‌گیری می‌شود که WEE نسبت به ML کاراتر است، زیرا با استفاده از آن می‌توان برآوردهای ثابتی را زمانی که $(z|x)$ P نامشخص و π_i مشخص است، به‌دست آورد. نتایج فوق در مطالعه بررسی عوامل مؤثر بر پوکی استخوان که حدود ۳۳٪ گم‌شده در یکی از متغیرها داشت به‌کار برده شد و پارامترهای معادله رگرسیون برآورد گردید، که معادله به‌دست آمده نشان داد، افزایش سن و برداشت رحم اثر کاهش‌دهنده و افزایش شاخص توده بدنی تأثیر مثبت بر روی تراکم استخوان دارد. هم‌چنین نتایج با داده‌های کامل مقایسه شد و به نظر می‌رسد که انحراف معیارهای حاصل از روش WEE کوچک‌تر از دو روش دیگر (ML و CC) می‌باشد و هم‌چنین روش WEE شرایط ساده‌تری (عدم نیاز به تعیین توزیع متغیر گم‌شده) نسبت به دو روش دیگر دارد. با این وجود برای رسیدن به نتایج قطعی انجام یک مطالعه شبیه‌سازی 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 Stastist Asoc, 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, Lipshultz S. A weighted estimating equation for linear regression with missing covariate data. Stat Med, 2002; 21(16):2421-36.

[6] Dempster AP, Larid NM, Rubin DB. Maximum likelihood from incomplete data via the EM alogorithm. 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.

[۹] طباطبایی‌راد سیدمجتبی. معادله برآورد وزنی (WEE) در رگرسیون خطی با داده‌های گم‌شده. پایان‌نامه کارشناسی ارشد، تهران: دانشگاه علوم پزشکی شهیدبهشتی، ۱۳۸۳.

[10] SAS Institute. What's new in Sas software: Release 8.1 and release 8.2. SAS Publishing; CD-Rom edition, 2001.