مجله دانشگاه علوم پزشکی خراسان شمالی
۱۳۹۰؛ ۳ (ویژه نامه آمار زیستی و اپیدمیولوژی):۱۳۸- ۱۲۷

Journal of North Khorasan University of Medical Sciences
2011;3(Biostatistics and Epidemiology Supplement):127-138

کاربرد روش بیزی در برآورد پارامترهای مدل رگرسیون لوجستیک با مقادیر گمشده تصادفی در متغیر کمکی

الهه کاظمی[۱]، مسعود کریملو[۲]*، مهدی رهگذر[۲]، عنایت الله بخشی[۳]، ایمانه عسگری[۴]*

[۱] کارشناس ارشد آمار زیستی، دانشگاه علوم بهزیستی و توانبخشی، تهران، ایران
[۲] دانشیار گروه آمار زیستی، دانشگاه علوم بهزیستی و توانبخشی، تهران، ایران
[۳] استادیار گروه آمار زیستی، دانشگاه علوم بهزیستی و توانبخشی، تهران، ایران
[۴] دکتر دندانپزشک، دانشکده دندانپزشکی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران
***نویسنده مسئول:** تهران، اوین، بلوار دانشجو، خیابان کودکیار، دانشگاه علوم بهزیستی و توانبخشی.
پست الکترونیک: mkarimlo@yahoo.com

چکیده

زمینه و هدف: *رگرسیون لوجستیک مدلی عمومی برای تحلیل داده های پزشکی و اپیدمیولوژیکی می باشد و اخیراً محققین معدودی تحقیقات خود را به تحلیل مدل های رگرسیون لوجستیک با وجود مقادیر گمشده در متغیرهای کمکی معطوف داشته اند. در بسیاری از پژوهش ها محققین با مجموعه داده هایی مواجه هستند که دارای مقادیر گمشده است. گمشدگی تهدید عمده ای برای درستی نتایج حاصل از مجموعه داده ها محسوب می شوند و اجتناب از آن بسیار مشکل است.*

مواد و روش کار: *ساتن و کارول تابع درستنمایی ویژه ای را برای برآورد پارامترهای مدل رگرسیون لوجستیک وقتی که برخی متغیرهای کمکی با مقادیر گمشده از نوع مکانیسم گمشدگی تصادفی (MAR) باشند و سایر متغیرها به طور کامل مشاهده شده باشند، معرفی کرده اند. در این پژوهش از این تابع درستنمایی در تحلیل بیزی برای برآورد پارامترهای مدل رگرسیون لوجستیک استفاده شده است و نتایج به دست آمده با روش های جانهی چندگانه و واحد کامل مقایسه شده است.*

یافته ها: *روش های مذکور را بر روی داده های شبیه سازی شده و داده های دندانپزشکی اجرا کرده و نتایج مقایسه ها نشان داد که برآوردهای به دست آمده از روش بیزی دارای انحراف معیارکوچکتری نسبت به دو روش دیگر می باشند.*

نتیجه گیری: *پس از مقایسه نتایج حاصل از سه روش مذکور نتیجه گرفته شد که اگر مکانیسم گمشدگی تصادفی باشد، به کارگیری تحلیل بیزی با تکنیک زنجیرهای مارکوف مونت کارلویی (MCMC) منجر به برآوردهای دقیق و فاصله اطمینان کوتاه تری نسبت به روش جانهی چند گانه و روش واحد کامل می شود.*

واژه های کلیدی: *رگرسیون لوجستیک، گمشدگی تصادفی (MAR)، تحلیل بیزی، زنجیرهای مارکوف مونت کارلویی (MCMC)، جانهی چندگانه، DMFT*

مقدمه

مدل رگرسیون لوجستیک روشی تحلیلی است که به طور وسیعی در تحقیقات پزشکی و اپیدمیولوژیکی کاربرد دارد. با گسترش و تنوع این مدل ها تجزیه و تحلیل داده های حاصل از تحقیقات در علوم مختلف با به کار گیری این مدل ها روز به روز افزایش یافته و به لحاظ نظری لزوم تحقیق در زمینه های گوناگون این مدل ها را بیش از پیش فراهم نموده است.

هدف از تحلیل رگرسیون لوجستیک همانند مدل های رگرسیون معمولی دستیابی به مدلی مناسب و در عین حال ساده جهت بررسی ارتباط بین متغیر پاسخ (وابسته)[۱]

1. Response variable ، dependent variable

با یک یا مجموعه ای از متغیر های مستقل (کمکی)[1] است. با این ویژگی که در این گونه مدل ها متغیر پاسخ بر خلاف رگرسیون معمولی عموماً از نوع رسته ای دو یا چند حالتی می باشد.

در بسیاری از مطالعات با مجموعه داده هایی مواجه می شویم که بخشی از آنها گزارش نشده اند از قبیل خود داری از پاسخ، عدم تکمیل کامل پرسشنامه ها یا پرونده ها، ناقص بودن چارچوب مطالعه و غیره. در این صورت با داده های گمشده سروکار داریم که می تواند در متغیر پاسخ یا در متغیرهای کمکی بوجود آید. در این پژوهش گمشدگی در متغیرهای کمکی مورد نظر می باشد. مواجهه با داده گمشده مشکل بومی تحقیقات اجتماعی، پزشکی و اپیدمیولوژیکی است و در هنگام تحلیل، وجود این گونه موارد مشکلات عدیده ای را فراهم می سازد و عملاً تجزیه و تحلیل آماری را به سوی نتایج اریب سوق داده و نهایتاً دستیابی به یک نتیجه گیری مفید از داده های جمع آوری شده را با مشکل مواجه می سازد. مشکل دیگر برقراری فرض های تحلیل های آماری است که بر پایه داده های کامل بنا شده اند، و مقادیر گمشده منجر به پیچیده شدن آن ها می شوند. با وجود مشکلاتی که مقادیر گمشده به وجود می آورند، گمشدگی در داده ها بهتر از مقادیر اشتباه می باشد. ساده ترین روش برای تجزیه و تحلیل چنین داده هایی صرفه نظر کردن از موردهای دارای مقادیر گمشده و انجام آنالیز با داده های کامل می باشد (روش واحد کامل)، که این روش در عمل کارا نیست [۱]

به طور کلی سه روش در نحوه بررسی داده های گمشده مورد استفاده قرار می گیرد [۲] :

۱: روش های مبتنی بر واحد های کامل

۲: روش های مبتنی بر جانهی[2]

۳: روش های مبتنی بر مدل

مسئله گمشدگی در داده ها و مشکلاتی که در تجزیه و تحلیل داده ها به وجود می آورند نظر بسیاری از محققین را به خود جلب کرده است و محققین بسیاری در زمینه مسئله گمشدگی در داده های بقاء[3]، داده های طولی[4]، داده های کارآزمایی بالینی[5]، داده های پیمایشی و در سایر انواع داده ها به تحقیق و بررسی پرداخته اند [۱۴-۱۳،۳-۱].

در تحلیل مقادیر گمشده دلایلی را که منجر به کامل نبودن داده ها می شوند را نیز باید مورد نظر قرار دهیم. بسیاری از روش هایی که برای مقابله با داده های گمشده ارائه شده اند وابسته به نوع مکانیسم گمشدگی می باشند، بنابراین نتایج این روش ها حساس به نوع مکانیسم می باشند [۵].

چهار نوع مکانیسم گمشدگی داریم. ۱.مکانیسم گمشدگی کاملاً تصادفی که مقادیر گمشده به وضعیت سایر متغیرها وابسته نمی باشند. ۲. مکانیسم گمشدگی تصادفی که مقادیر گمشده به وضعیت متغیرهایی که به طور کامل مشاهده شده اند، وابسته است.۳. مکانیسم گمشدگی غیر قابل اغماض که مقادیر گمشده به وضعیت متغیرهایی که گمشده شده اند وابسته می باشد. ۴. گمشدگی به علت ذات طرح نوع دیگری از مکانیسم گمشدگی داده است که مقادیر به علت این که آن ها را به طور طبیعی و معمول نمی توان اندازه گیری کرد گمشده اند، این یک نوع مکانیسم گمشدگی جدیدی است که شامل مشکلاتی در اندازه گیری است [۴،۱۷].

به دلیل مشکلاتی که در تشخیص نوع مکانیسم گمشدگی وجود دارد و برخی از آن ها آزمون پذیر نمی باشند، در بسیاری از تحقیقات روی داده های شبیه سازی شده کار کرده اند و یا با فرض پذیرش مکانیسم مورد نظر روش تحقیقات خود را روی داده های واقعی انجام داده اند. پس لزوم مطالعه و تحقیق در رابطه با مسئله گمشدگی و تشخیص صحیح مکانیسم ها به شدت حس می شود.

با توجه به این که در اکثر مجموعه داده ها با مشکل مقادیر گمشده مواجه هستیم و مقادیر گمشده منجر به نتایج اریب می شوند، لزوم تحقیق بیشتر در این زمینه کاملاً حس می شود.

1 .Independent (Covariate) variable
2 .Imputation

4 .Survival data
4.Longitudinal data
5.Clinical trial data

هدف این پژوهش بررسی و تحلیل بیزی با استفاده از تابع درستنمایی ساتن و کارول و مقایسه این روش با روش های واحد کامل[1] و جانهی چندگانه[2] می باشد. انتظار بر این است که با استفاده از روش بیزی برآوردهای دقیق تر و نتایج بهتری نسبت به روش های معمول به دست بیاوریم.

روش کار

الف: مدل رگرسیون لوجستیک باوجود مقادیر گمشده درمتغیر کمکی: در این بخش رگرسیون لوجستیک با پاسخ دو حالتی را هنگامی که یکی از متغیرهای کمکی دارای مقادیر گمشده است و سایر متغیرهای کمکی کاملاً مشاهده شده اند را مورد بحث و بررسی قرار می دهیم.

فرض می‌کنیم که برای فرد iام، متغیر y_i متغیری تصادفی باشد با پاسخ دوحالتی که نوعاً در مطالعات پزشکی و اپیدمیولوژیکی متغیر بیماری نامیده می‌شود. معمولاً yi=1 مشخص‌کنندهٔ افراد بیمار و yi=0 نشان‌دهندهٔ افراد سالم می‌باشد. همچنین فرض می‌کنیم که z بردار ستونی از متغیرهای کمکی باشد که مقادیر آن به طور کامل برای کلیه افراد مشاهده‌شده باشد و x بردار ستونی از متغیرهای کمکی است که مقادیر آن برای برخی از افراد به طور کامل مشاهده نشده به عبارت دیگر دارای مقادیر گمشده است. بردار(V(X,Z برداری از کلیهٔ متغیرهای کمکی باشد که می‌خواهیم وارد مدل نماییم. همچنین متغیر نشانگر دو وحالتی Δ_i را به صورت زیر تعریف می‌کنیم:

Δ_i=۱ اگر x_i مشاهده شده باشد ۱

اگر X_i گمشده باشد ۰

هدف از تحلیل رگرسیون لوجستیک برآورد پارامترهای مدل یعنی β_0 و بردار $\boldsymbol{\beta}$ جهت توصیف رابطهٔ بین متغیر پاسخ Y با مجموعه‌ای از متغیرهای کمکی بردار V(x,z) می‌باشد.

در حالتی که بردار x به طور کامل برای تمام افراد تحت مطالعه مشاهده نشده و بردار z به طور کامل مشاهده شده باشد، بخت و نسبت بخت ها در مدل رگرسیون لوجستیک از روابط زیر محاسبه می‌گردد:

(۱)

$$\widetilde{\theta}(z)=\frac{P(Y=1|Z=z)}{P(Y=0|Z=z)}$$

و

(۲)

$$\widetilde{\psi}(z,z')=\frac{\widetilde{\theta}(z)}{\widetilde{\theta}(z')}$$

که ´z مقداری متفاوت از z است. به علاوه تعریف می‌کنیم:

(۳)

$$\rho_0(x\,|z)=P(X=x\,|Y=0,Z=z)$$

(۴)

$$\rho_1(x\,|z)=P(X=x\,|Y=1,Z=z)$$

همان طور که ملاحظه می‌شود تابع احتمال $\rho_0(x\,|z)$ نشان‌دهندهٔ توزیع احتمال بردار تصادفی x در افراد سالم به شرط مشاهدات کامل z و $\rho_1(x\,|z)$ بیان‌کنندهٔ توزیع احتمال بردار تصادفی x در افراد بیمار به شرط مشاهدات کامل z است.

اولین نتیجه مقالات ساتن و کوپر [۷-۶]، نشان می دهد که:

(۵)

$$\widetilde{\theta}(z)=\sum_x \theta(x,z).\rho_0(x\,|z)$$

که در آن مجموع روی همهٔ مقادیر ممکن بردار x می‌باشد.

دومین نتیجهٔ مقاله مذکور عبارت است از:

(۶)

$$\rho_1(x|z)=\frac{\theta(x,z)\rho_0(x|z)}{\sum_x \theta(x,z)\rho_0(x|z)}$$

در روابط بالا اگر متغیر کمکی x پیوسته باشد به جای مجموع باید از انتگرال استفاده شود و عبارت های

1!.Complete Case
2 .Multiple Imputation

$\rho_0(x|z)$ و $\rho_1(x|z)$ نیز توابع چگالی احتمال درنظر گرفته شوند [۳].

ب: تابع درستنمایی رگرسیون لوجستیک با مقادیر گمشده در متغیر کمکی

اگر این گمشدگی در متغیر x تنها به مقادیر متغیر پاسخ y بستگی داشته باشد و نه به خود x یا z، یا به عبارت دیگر گمشدن تصادفی MAR باشد، بنا به نظریهٔ لیتل[1] و روبین[2] [۸] تابع درستنمایی برای حجم نمونهٔ n تایی عبارتست از:

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}P(y_i,x_i|z_i)=\prod_{i=1}^{n}P(y_i|z_i).P(x_i|y_i,z_i) \quad (۷)$$

حال برای سادگی بدون اینکه از کلیت مسئله کاسته شود، فرض می‌کنیم که m (m<n) مقدار اول متغیر x مشاهده شده و n-m مقدار بعدی، گمشده باشند (شکل ۱) که در آنالیز داده‌های کامل معمولاً ازاین m مشاهدهٔ اول برای برازش مدل استفاده می‌شود. همانطور که در شکل ۱ ملاحظه می شود متغیر کمکی x دارای مقادیر گمشده ولی متغیر کمکی z و متغیر پاسخ y دارای داده های کامل هستند.

با توضیح فوق و با استفاده از تعریف متغیر نشانگر Δ می توان رابطه (۷) را به صورت زیر تغییر داد:

$$L(\beta)=\prod_{i=1}^{n}P(y_i\mid z_i)\ .\ \prod_{i=1}^{m}P(x_i|y_i,z_i)=\prod_{i=1}^{n}P(y_i|z_i)\cdot\left[P(x_i|y_i,z_i)\right]^{\Delta} \quad (۸)$$

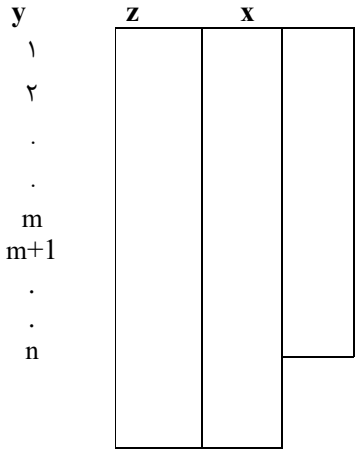
با توجه به دو حالتی بودن متغیر پاسخ y، عبارت داخل کروشه رابطه فوق، اگر $y_i=0$، باشد برابر است با:

$$P(x|y,z)=\mathrm{P}(x|y=0,z)=\rho_0(x|z)$$

که همان رابطهٔ (۳) است .

و اگر $yi=1$ باشد به رابطهٔ (۴) می‌رسیم:

$$P(x|y,z)=\mathrm{P}(x|y=1,z)=\rho_1(x|z)$$



شکل ۱: داده‌های مشاهده شده که در آن متغیر کمکی x ، n-m داده گمشده دارد.

با جایگذاری عبارات فوق در رابطهٔ (۸) داریم:

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}P(\mathrm{y_i}\mid \mathrm{z_i})\left[\left[\rho_0(\mathrm{x_i}\mid \mathrm{z_i})\right]^{1-y_i}\left[\rho_1(\mathrm{x_i}\mid \mathrm{z_i})\right]^{y_i}\right]^{\Delta_i}$$

از طرفی $P(yi|zi)$ عبارت درستنمایی برای داده‌های کامل است، لذا با استفاده از روابط (۱) فقط برای داده‌های متغیر کمکی z که به طور کامل مشاهده شده‌اند داریم:

$$P(y_i|z_i)=\frac{\left[\tilde{\theta}(z_i)\right]^{y_i}}{1+\tilde{\theta}(z_i)} \quad (۱۰)$$

با قرار دادن رابطهٔ فوق در رابطهٔ (۹) تابع درستنمایی بصورت زیر تغییر می‌کند:

1-!Little
2- Rubin!

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}\frac{\theta(z_i)^{y_i}}{1+\theta(z_i)}\times\rho_0(X_i|Z_i)^{\Delta i(1-Yi)}\rho_1(X_i|Z_i)^{\Delta i(Yi)} \quad (11)$$

با نگاهی اجمالی به تابع درستنمایی مذکور، متشکل از سه جمله مجزا، ملاحظه می‌شود که اگر $\Delta i=0$ باشد یعنی مقدار x برای نفر iام مشاهده نشده باشد فقط جمله اول در معادله باقی می‌ماند که درستنمایی برای داده‌های کامل بدون حضور متغیر کمکی x، یعنی رابطهٔ (۱۰) می‌باشد و در صورتی که $\Delta_i=1$ باشد یعنی مقدار x برای نفر i ام مشاهده شده است و لذا با دو حالت مواجه هستیم یا $y_i=0$ است که در این حالت جملهٔ سوم از مدل خارج می‌شود و درستنمایی با دو جملهٔ اول برقرار می‌گردد و یا $y_i=1$ است که در این صورت جملهٔ دوم حذف شده و درستنمایی با جملات اول و سوم برقرار خواهد شد.

برای ساده سازی تابع درستنمایی با استفاده از روابط (۵) و (۶) در رابطهٔ (۱۱) داریم:

$$L(\beta)=\prod_{i=1}^{n}\left\{\frac{\left[\sum_x\theta(x,z_i)\,\rho_0(x|z_i)\right]^{y_i}}{1+\sum_x\theta(x,z_i)\,\rho_0(x|z_i)}\cdot\rho_0(X_i|z_i)^{\Delta_i}\cdot\rho_0(X_i|z_i)^{-\Delta_i y_i}\rho_1(X_i|z_i)^{\Delta_i y_i}\right\}$$

$$=\prod_{i=1}^{n}\left\{\frac{\left[\sum_x\theta(x,z_i)\,\rho_0(x|z_i)\right]^{y_i}}{1+\sum_x\theta(x,z_i)\,\rho_0(x|z_i)}\cdot\rho_0(X_i|z_i)^{\Delta_i}\cdot\left[\frac{\rho_1(X_i|z_i)}{\rho_0(X_i|z_i)}\right]^{\Delta_i y_i}\right\},$$

$$=\prod_{i=1}^{n}\left\{\frac{\left[\sum_x\theta(x,z_i)\,\rho_0(x|z_i)\right]^{y_i}}{1+\sum_x\theta(x,z_i)\,\rho_0(x|z_i)}\cdot\rho_0(X_i|z_i)^{\Delta_i}\cdot\left[\frac{\rho_0(X_i|z_i)\theta(X_i,z_i)}{\sum_x\rho_0(x|z_i)\theta(x,z_i)}\right]^{\Delta_i y_i}\right\}$$

$$=\prod_{i=1}^{n}\left\{\frac{\left[\sum_x\theta(x,z_i)\,\rho_0(x|z_i)\right]^{y_i}}{1+\sum_x\theta(x,z_i)\,\rho_0(x|z_i)}\cdot\rho_0(X_i|z_i)^{\Delta_i}\cdot\left[\frac{\theta(X_i,z_i)}{\sum_x\theta(x,z_i)\rho_0(x|z_i)}\right]^{\Delta_i y_i}\right\}$$

آن گاه:

$$L(\beta)=\prod_{i=1}^{n}\left\{\frac{\theta(x_i,z_i)^{\Delta y_i}}{1+\sum_x\theta(x,z_i)\,\rho_0(x|z_i)}\cdot\left[\sum_x\theta(x,z_i)\rho_0(x|z_i)\right]^{y_i(1-\Delta)}\cdot\rho_0(X_i|z_i)^{\Delta}\right\} \quad (12)$$

رابطهٔ (۱۲) نشان دهنده تابع درستنمایی رگرسیون لوجستیک در حالت برداری برای زمانی است که در مدل، بردار متغیر کمکی x با مقادیر گمشده حضور داشته باشد. ملاحظه می‌شود که این رابطه تابعی از $\theta(.)$ و $\rho_0(.)$ است.

یکی از نکات کلیدی بحث حاضر، تعیین مدل مناسبی برای عبارت $\rho_0(x|z_i)$ در رابطه (۱۲) است، همان گونه که ذکر شد این عبارت بیان‌کنندهٔ توزیع متغیر دارای مقادیر گمشده در افراد سالم به شرط معلوم بودن مقادیر متغیر کمکی z می‌باشد. در صورتیکه متغیرهای کمکی x و z دارای مقادیری شمارا و محدود باشند برای توزیع احتمال ρ_0 مدلی از خانواده نمایی به فرم زیر توسط ساتن[1] و کارول[2] پیشنهاد شده است [۳] :

$$\rho_0(x|z)=\frac{e^{\gamma xz}}{\sum_{x'}e^{\gamma x'z}} \quad (13)$$

پ: تحلیل بیزی مدل رگرسیون لوجستیک با وجود مقادیر گمشده درمتغیر کمکی X : دراین بخش نحوه تحلیل مدل رگرسیون لوجستیک را با استفاده از تابع درستنمایی تعمیم یافته ساتن و کارول (۱۲)، به روش بیزی مورد بحث قرار می دهیم. برای سادگی دراین مرحله نیز فرض می کنیم که تنها دومتغیر کمکی x وz را داریم که متغیر x دارای مقادیر گمشده است. بنابراین دررابطه (۱۲) با در نظر گرفتن مدل اشباع رگرسیون لوجستیک، برآورد بردار پارامتر β به روش بیزی مورد نظر می باشد. حال با توجه به اینکه دراین مدل $\beta_i=\log(OR)$ و $0<OR<\infty$ بنابراین دامنه β_i از $-\infty$ تا $+\infty$ بوده ومی توان توزیع پیشین آن را نرمال به صورت $\beta_i\sim N(\mu_{\beta_i},\sigma^2_{\beta_i})$ درنظرگرفت. بنابراین درحالت کلی

1- Satten
2!- Carroll

با فرض آنکه $\pi(\boldsymbol{\beta})$ توزیع پیشین پارامتر $\boldsymbol{\beta}$ باشد، توزیع پیشین توام پارامترهای β_i نرمال چندمتغیره به صورت زیرخواهد بود.

(۱۴)

$$\pi(\boldsymbol{\beta}) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} e^{-(\boldsymbol{\beta}-\mu_\beta)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\mu_\beta)/2}$$

که درآن p عبارتست از تعداد پارامترها (دراینجا ۶ پارامتر) و μ_β و Σ^{-1} ابر پارامترهای توزیع پیشین $\boldsymbol{\beta}$ ودارای مقادیر معلومی هستند. بااستفاده ازتوزیع پیشین فوق و درستنمایی $l(\boldsymbol{\beta}\mid x,z)$ ازرابطه (۱۲)، توزیع توام مشاهدات وپارامترهای مدل برابراست با:

$$\pi\ (x,z,\boldsymbol{\beta}) = l(\boldsymbol{\beta}\mid x,z)\ \pi(\boldsymbol{\beta})$$

$$= \prod_{i=1}^{n} \frac{\left(e^{\beta_0+\beta_1 x_i+\beta_2 z_i+\beta_{12}x_i z_i}\right)^{\Delta_i y_i}}{1+\sum_x \left[e^{\beta_0+(\beta_1+\gamma_1)x+\beta_2 z_i+(\beta_{12}+\gamma_{12})xz_i} \cdot \frac{1}{\sum_x e^{\gamma_1 x+\gamma_{12}xz_i}} \right]}$$

$$\times \left\{ \sum_x e^{\beta_0+(\beta_1+\gamma_1)x+\beta_2 z_i+(\beta_{12}+\gamma_{12})xz_i} \cdot \frac{1}{\sum_x e^{\gamma_1 x+\gamma_{12}xz_i}} \times \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} e^{-(\boldsymbol{\beta}-\mu_\beta)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\mu_\beta)/2} \right\}^{y_i(1-\Delta_i)}$$

مشاهده می شود که توزیع پسین فوق بسیار پیچیده است لذا امکان محاسبه توزیع پسین پارامترهای مدل به روش تحلیلی وجود ندارد. بنابراین برای انجام تحلیل بیزی لازم است بااستفاده ازروش های زنجیرمارکوف مونت کارلوئی (MCMC) توزیع پسین پارامترها را تقریب بزنیم [۱]

توزیع پیشین هریک از پارامترهای β_i را می توان ناآگاهی بخش به صورت توزیع یکنواخت (۱۰۰و ۱۰۰-)U ویا توزیع نرمال با میانگین صفر وواریانس خیلی بزرگ (مثلا 10^6 $\mu_{\beta_i}=0, \sigma^2_{\beta_i}=10^6$) فرض کرد. درمرجع [۱]با هر دو توزیع پیشین تحلیل صورت پذیرفت،که مقادیر برآوردها با هردو یکسان بودند.

ج: جانهی چندگانه : ایده کلیدی جانهی چند گانه تولید چندین مقدار برای هر داده گمشده می باشد. تکرار جانهی ها این امکان را به ما می دهد که درجه حساسیت به جانهی را بتوان بررسی کرد و همچنین به موجب آن بتوان انحراف معیار معتبر را نیز محاسبه کرد. برای هر m تکرار جانهی، با ادغام داده های مشاهده شده و داده های جانهی شده، m مجموعه داده کامل را به دست می آوریم که برای هر کدام می توان برآورد های نقطه ای معمول و انحراف معیار را محاسبه کرد. سپس برآورد نقطه ای چند گانه به وسیله میانگین m برآورد نقطه ای معمول به دست می آید. جانهی مکرر[1] یک توصیف مناسب و شایسته برای این روش می باشد [۴].

مشکل کاربردی (عملی) با مکانیسم گمشدگی تصادفی این است که هیچ راهی برای تایید این که احتمال گمشدگی داده ها تنها تابعی از متغیرهای مشاهده شده است وجود ندارد [۴،۱۷]. فلایس[2] و همکاران الگوریتمی را برای تشخیص گمشدگی تصادفی(MAR) معرفی کرده اند، این الگوریتم عمومی نبوده و زمانی کاربرد دارد که مجموعه داده ها به گونه ای باشد که بتوان از رگرسیون لو جستیک استفاده کرد و همچنین متغیری که دارای مقادیر گمشده می باشد متغیری رسته ای و دو حالتی باشد. شیوه کار به طور کامل در منبع [۴] شرح داده شده است.

یافته ها

در این بخش روش های معرفی شده در بخش قبل را روی داده های شبیه سازی شده و داده های دندانپزشکی اجرا کرده و به مقایسه نتایج به دست آمده می پردازیم.

ابتدا با استفاده از برنامه نوشته شده در نرم افزار R، به تولید اعداد تصادفی که دارای مقادیر گمشده در یک متغیر می باشند، پرداختیم. یک مجموعه داده ۵۰۰ تایی برای سه متغیر X، Y و Z که دارای ۴۰٪ گمشدگی در داده ها ی متغیر X می باشند تولید کردیم.

P(Y=0, Z=0)= 0.20 P(Y=0,Z=1)= 0.60
P(Y=1,Z=0)= 0.35 P(Y=1,Z=1)= 0.45

1 .Repeated Imputation
2!. Fleiss

این درصدهای در نظر گرفته شده انتخابی بوده و متفاوت بودن آن ها برای ایجاد مکانیسم گمشدگی تصادفی مد نظر بوده است و می توان درصدهای دیگری را که متفاوت از هم باشند و در مجموع ۴۰٪ گمشدگی کلی ایجاد کنند را نیز در نظر گرفت.

پس از ایجاد گمشدگی درمتغیر کمکی x به سه روش مطرح شده در بخش قبل، روش واحد کامل(CC)، روش جانهی چند گانه(MI) و روش بیز درستنمایی ساتن و کارول (SCMCMC) اقدام به برآورد پارامترهای مدل رگرسیون لوجستیک که در آن یک متغیر دارای مقادیر گمشده می باشد نمودیم. دو روش اول با استفاده از نرم افزار SPSS انجام شد. برای اجرای روش بیز درستنمایی ساتن و کارول با استفاده از دستور Zeros در قسمت Tricks برنامه ای در نرم افزار WinBUGS نوشته شد. برای این که مقایسه این روش ها امکان پذیر باشد با استفاده از روش ماکسیمم درستنمایی و درستنمایی ساتن و کارول برای داده های کامل نیز پارامترهای رگرسیون لوجستیک برآورد شد. برای این مجموعه داده برای طبقات مختلف داده های مشاهده شده درصد های متفاوتی که ۴۰٪ گمشدگی کلی را به دست دهند، در نظر گرفته و ۹ سری مجموعه داده دیگر نیز تولید شد و سپس تمامی روش های مذکور برای هر ۱۰ سری مجموعه داده اجرا گردید. پیش از برازش مدل رگرسیون لوجستیک روی داده های مذکور ابتدا با استفاده از الگوریتم معرفی شده در بخش قبل و برنامه نوشته شده آن در نرم افزار SAS وجود مکانیسم گمشدگی تصادفی را بررسی کرده و پس از تایید آن از مجموعه داده ها استفاده کرده ایم. میانگین نتایج حاصل ۱۰ سری مجموعه داده درجدول ۱ آمده است.در این جداول β_0 ضریب ثابت، β_1 ضریب متغیر X، β_2 ضریب متغیر Z و β_{12} ضریب اثر متقابل متغیرهای Xو Z می باشند. با مقایسه انواع برآوردهای پارامترها در جدول ۱ ملاحظه می شود که برآورد های MLE و SCMCMCکه مربوط به داده های کامل است به هم نزدیک بوده و اختلاف ناچیزی دارند. همچنین ملاحظه می شود که برآوردهای بیزی (SCMCMC) که با به کارگیری درستنمایی ساتن و کارول مربوط به داده های گمشده، به دست آمده اند نسبت به سایر برآوردها، واریانس کمتری داشته و لذا دقیق ترمی باشند. در ضمن برآوردهای بیزی دارای کوتاهترین فاصله اطمینان نیز می باشند. به طور کلی نتایج فوق نشان می دهد که درصورتی که باداده های گمشده با مکانیسم گمشدگی تصادفی مواجه باشیم با استفاده از تکنیک بیزی MCMC روی توزیع پسینی که ازترکیب تابع درستنمایی تعمیم یافته ساتن و کارول وتوزیع پیشین ناآگاهی بخش حاصل می شود، می توان به برآوردهای دقیق تر درمقایسه با حالت حذف آزمودنی ها ازمطالعه و نیز روش جانهی دست یافت. کریملو و همکاران [۱] در پژوهش خود نیز نتیجه گرفته بودند که روش بیز از روش ماکسیمم درستنمایی نتایج دقیق تری را منجر می شوند که مشابه نتایج گرفته شده از جداول ۱ می باشد.

سه روشی که در این تحقیق معرفی شد را روی داده های دندانپزشکی نیز پیاده کرده. از اطلاعات متراژ منزل (متغیر مستقل)، گروه سنی (متغیر مستقل) و DMF[1] (متغیر وابسته)، استفاده شده است. برای دانش آموزانی که DMF<=3 داشته اند کد ۰ و DMF>3 داشته اند کد ۱ در نظر گرفته شده است. داده های ۵۵۰ دانش آموز مورد تحلیل و بررسی قرار می گیرند که برای متغیر گروه سنی و DMF اطلاعات تمامی ۵۵۰ دانش آموز به طور کامل مشاهده شده است. متغیر متراژ منزل دارای مقادیر گمشده بودند. بنابراین در متغیر متراژ منزل ۴۳/۲۷٪ اطلاعات گمشده داریم. ساده ترین روش بررسی مکانیسم گمشدگی داده ها محاسبه نسبت گمشدگی در طبقات سایر متغیرها و مقایسه آن ها با یکدیگر می باشد.

P(Y=0, Z=0)= 0.46 P(Y=0,Z=1)= 0.36
P(Y=1,Z=0)= 0.43 P(Y=1,Z=1)= 0.51

1. Decayed،Missing،Filled

جدول۱: مقایسه نتایج به دست آمده برای برآورد پارامترها برای داده های کامل و داده های دارای مقادیر گمشده با استفاده از روش های واحد کامل،جانهی چند گانه و روش بیز درستنمایی ساتن و کارول

			β	SE	CI(95%)[1]	Dif[2]	OR	CI(95%)[3]	Dif[2]
β_0	برآورد پارامترها برای داده های کامل	MLE SPSS	-۰/۳۰۵	۰/۱۵۸۰	(-۰/۶۱۵,۰/۰۰۵)	۰/۶۱۹	۰/۷۳۷	(۰/۵۴۱,۱/۰۰۵)	۰/۴۶۴
		SCMCMC WinBugs	-۰/۳۱۸	۰/۱۵۶۰	(-۰/۶۲۴,-۰/۰۱۲)	۰/۶۱۲	۰/۷۲۸	(۰/۵۳۶,۰/۹۸۸)	۰/۴۵۲
	برآورد پارامترها با ۴۰٪ گمشدگی تصادفی در متغیر X	Complete Case SPSS	-۰/۲۶۵	۰/۲۰۴۸	(-۰/۶۶۷,۰/۱۳۶)	۰/۸۰۳	۰/۷۷۰	(۰/۵۱۳,۱/۱۴۶)	۰/۶۳۲
		MI SPSS	-۰/۲۲۳	۰/۱۸۱۸	(-۰/۵۸۰,۰/۱۳۳)	۰/۷۱۳	۰/۷۹۹	(۰/۵۶۰,۱/۱۴۲)	۰/۵۸۲
		SCMCMC WinBugs	-۰/۲۷۵	۰/۱۶۷۱	(-۰/۶۰۸,۰/۰۴۸)	۰/۶۵۵	۰/۷۵۵	(۰/۵۴۴,۱/۰۴۹)	۰/۵۰۴
β_1 (x)	برآورد پارامترها برای داده های کامل	MLE SPSS	۰/۹۱۲	۰/۲۷۶۰	(۰/۳۷۱,۱/۴۵۳)	۱/۰۸۲	۲/۴۸۹	(۱/۴۴۹,۴/۲۷۶)	۲/۸۲۷
		SCMCMC WinBugs	۰/۹۴۷	۰/۲۵۳۶	(۰/۴۴۹,۱/۴۴۴)	۰/۹۹۴	۲/۵۷۷	(۱/۵۶۷,۴/۲۳۶)	۲/۶۶۸
	برآورد پارامترها با ۴۰٪ گمشدگی تصادفی در متغیر X	Complete Case SPSS	۱/۱۳۸	۰/۴۰۵۲	(۰/۳۰۵,۱/۹۷۲)	۱/۶۶۷	۳/۱۲۱	(۱/۳۵۶,۷/۱۸۲)	۵/۸۲۶
		MI SPSS	۰/۷۸۵	۰/۳۹۱۰	(۰/۰۱۹,۱/۵۵۲)	۱/۵۳۳	۲/۱۹۳	(۱/۰۱۹,۴/۷۲۰)	۳/۷۰۰
		SCMCMC WinBugs	۰/۷۷۲	۰/۳۰۸۶	(۰/۱۷۷,۱/۳۸۷)	۱/۳۰۰	۲/۱۸۶	(۱/۱۹۴,۴/۰۰۲)	۲/۸۰۸
β_2 (z)	برآورد پارامترها برای داده های کامل	MLE SPSS	۰/۷۳۶	۰/۲۲۴۰	(۰/۲۹۷,۱/۱۷۵)	۰/۸۷۸	۲/۰۸۷	(۱/۳۴۶,۳/۲۳۸)	۱/۸۹۳
		SCMCMC WinBugs	۰/۷۴۹	۰/۲۱۱۶	(۰/۳۳۴,۱/۱۶۴)	۰/۸۲۹	۲/۱۱۵	(۱/۳۹۷,۳/۲۰۲)	۱/۸۰۵
	برآورد پارامترها با ۴۰٪ گمشدگی تصادفی در متغیر X	Complete Case SPSS	۰/۸۲۳	۰/۲۸۱۰	(۰/۲۷۳,۱/۳۷۴)	۱/۱۰۱	۲/۲۷۸	(۱/۳۱۴,۳/۹۵۱)	۲/۶۳۸
		MI SPSS	۰/۵۰۵	۰/۲۳۷۷	(۰/۰۳۹,۰/۹۷۱)	۰/۹۳۲	۱/۶۵۷	(۱/۰۴۰,۲/۶۴۰)	۱/۶۰۰
		SCMCMC WinBugs	۰/۵۵۷	۰/۲۲۴۵	(۰/۱۳۶,۱/۰۱۷)	۰/۸۸۰	۱/۷۸۰	(۱/۱۴۶,۲/۷۶۴)	۱/۶۱۸
β_{12} (xz)	برآورد پارامترها برای داده های کامل	MLE SPSS	-۱/۲۲۵	۰/۳۸۶۰	(-۱/۹۸۲,-۰/۴۶۸)	۱/۵۱۳	۰/۲۹۴	(۰/۱۳۸,۰/۶۲۶)	۰/۴۸۸
		SCMCMC WinBugs	-۱/۲۶۷	۰/۳۴۷۳	(-۱/۹۴۸,-۰/۵۸۶)	۱/۳۶۱	۰/۲۸۱	(۰/۱۴۳,۰/۵۵۶)	۰/۴۱۴
	برآورد پارامترها با ۴۰٪ گمشدگی تصادفی در متغیر X	Complete Case SPSS	-۱/۳۸۶	۰/۵۶۹۴	(-۲/۵۰۲,-۰/۲۷۰)	۲/۲۳۲	۰/۲۵۰	(۰/۰۸۲,۰/۷۴۶)	۰/۶۸۲
		MI SPSS	-۰/۸۲۹	۰/۵۰۴۰	(-۱/۸۱۷,۰/۱۵۹)	۱/۹۷۶	۰/۴۳۷	(۰/۱۶۳,۱/۱۷۲)	۱/۰۱۰
		SCMCMC WinBugs	-۱/۰۲۷	۰/۴۱۴۴	(-۱/۸۳۹,-۰/۲۱۵)	۱/۶۲۵	۰/۳۵۸	(۰/۱۵۹,۰/۸۰۷)	۰/۶۴۸

۱.فاصله اطمینان برای پارامتر

۲.طول فاصله اطمینان برای پارامتر

۳.طول فاصله اطمینان برای نسبت بخت ها

همان طور که ملاحظه می شود نسبت گمشدگی در طبقات با یکدیگر برابر نمی باشند و از ۳۶٪ تا ۵۱٪ تغییر می کنند.

نتیجه می گیریم که گمشدگی در طبقات سایر متغیرها به طور یکسان رخ نداده و لذا در این داده ها مکانیسم گمشدگی از نوع کاملاً تصادفی نمی باشند.

سپس با استفاده از روش های تحلیلی به بررسی دقیق تر وجود یا عدم وجود مکانیسم گمشدگی کاملاً تصادفی می پردازیم. گمشدگی کاملاً تصادفی تنها مکانیسم گمشدگی است که آزمون پذیر می باشد [۹]. ساده ترین روش برای ارزیابی گمشدگی کاملاً تصادفی استفاده از یک سری آزمون های t مستقل [۱۰] و آزمون گمشدگی کاملاً تصادفی لیتل [۱۱] می باشد. پس از انجام این دو آزمون با استفاده از نرم افزار SPSS 12 در هردو آزمون (P-value <۰/۰۵) بود و به طور کلی نتیجه گرفته شد که گمشدگی در طبقات سایر متغیرها به طور یکسان رخ نداده اند و دارای مکانیسم گمشدگی کاملاً تصادفی نمی باشند پس داده های مورد نظر دارای مکانیسم گمشدگی تصادفی یا غیر قابل اغماض می باشند که نیاز به بررسی های بیشتر داریم.

مده برای برآورد پارامترها برای داده ها ی گمشده برای گروه سنی، متراژ و DMF

		β	SE	CI(95%)[1]	Dif[2]
β_0	Complete Case	-۰/۴۲۴	۰/۲۱۴۳	(-۰/۸۴۳,-۰/۰۰۵)	۰/۸۳۹
	SPSS MI	-۰/۴۱۰	۰/۱۸۱۱	(-۰/۷۶۵,-۰/۰۵۵)	۰/۷۱۰
	SPSS SCMCMC				

با استفاده از الگوریتم تشخیص مکانیسم گمشدگی تصادفی که در بخش قبل معرفی شد به بررسی وجود مکانیسم گمشدگی تصادفی در این داده ها پرداختیم که با توجه به نتایج به دست آمده مشاهده می شود که برآورد پارامتر α_x (۰/۰۰۰۱۹-) تقریباً برابر صفر است و انحراف استاندارد (۰/۰۰۰۸۲) بسیار کوچکی دارد، که نشان دهنده این است که گمشدگی داده ها از نوع مکانیسم گمشدگی تصادفی می باشند. بنابراین از متغیر متراژ به عنوان متغیری که دارای مقادیر گمشده است در مدل استفاده می کنیم.

همانطور که در جدول ۲ ملاحظه می شود واریانس های برآورد بیزی نیز کمتر از سایر برآوردهای به دست آمده از روش واحد کامل و روش جانهی می باشند. و فواصل اطمینان برآورد های بیزی SCMCMC در مقایسه با فواصل اطمینان برآورد های به دست آمده از روش های واحد کامل و جانهی چندگانه باریک تر است. که مشابه نتایج به دست آمده از جدول ۱ می باشد.

بحث

داده گمشده یک مشکل عمومی در مجموعه داده ها می باشد. محققین چندین روش (واحد کامل، جانهی تکی، جانهی چندگانه، ماکسیمم درستنمایی و روش بیزی) را برای مقابله با مشکل گمشدگی داده معرفی کرده اند. هدف این پژوهش معرفی و بکار گیری روش بیزی برآورد پارامترهای مدل رگرسیون لوجستیک با تکنیک زنجیرهای مارکوف مونت کارلویی (MCMC)، با استفاده از تابع درستنمایی ویژه معرفی شده توسط ساتن و کارول با وجود داده های گمشده، از نوع مکانیسم گمشدگی تصادفی، و مقایسه آن با روش واحد کامل و روش جانهی چند گانه بوده است.

در بخش قبل با استفاده از داده های شبیه سازی شده و داده های دندانپزشکی سه روش واحد کامل، جانهی چندگانه و روش بیزی را مقایسه کرده و پس از بررسی جداول ۱ و ۲ نتایج زیر حاصل شده است.

مقادیر برآوردها با استفاده از روش واحد کامل به شدت اریب بوده، بنابراین درعمل باحذف آزمودنی های دارای مقادیر گمشده درحالت گمشدگی تصادفی، صدمات جبران ناپذیری به تفسیر نتایج مطالعه وارد می نمائیم. استامی[1] و همکاران [۱۳] (۲۰۰۹) در مطالعه خود با استفاده از داده های شبیه سازی شده اثر در نظر نگرفتن داده های گمشده در مطالعات پی گیری کننده را که منجر به اریبی نتایج می شوند را نشان داده اند.

واریانس برآوردها با استفاده از روش جانهی چندگانه نسبت به روش واحد کامل کوچکتر می باشند. فواصل اطمینان برآورد های به دست آمده از روش جانهی چندگانه که مربوط به مجموعه داده های با مقادیر گمشده است در مقایسه با فواصل اطمینان برآورد های به دست آمده از روش واحد کامل باریک تر است.

مقادیر برآوردها با استفاده از روش بیزی با تکنیک زنجیرهای مارکوف مونت کارلویی (MCMC)، با استفاده از تابع درستنمایی ویژه معرفی شده توسط ساتن و کارول نسبت به روش های واحد کامل و جانهی دارای اریبی کمتری می باشد ضمن این که مقدار واریانس برآوردها نیز بسیار کوچکتر از واریانس های این دو روش می باشد. فواصل اطمینان برآورد های SCMCMC که مربوط به مجموعه داده های با مقادیر گمشده است در مقایسه با فواصل اطمینان برآورد های به دست آمده از روش های واحد کامل و جانهی چندگانه باریک تر است.

ساتن و کارول [۳] (۲۰۰۰) نیز به روش کلاسیک نشان دادند که برآوردهای مدل رگرسیون لوجستیک درمطالعات مورد- شاهدی با بکارگیری تابع درستنمایی تعمیم یا فته بهبود می یا بد. روش بیزی نسبت به سایر روش های برآورد، مقادیر نزدیکتری به داده های کامل داشتند که این مسئله توسط سینها[2] و همکاران [۱۴] (۲۰۰۴) نیزدر مورد مطالعات مورد- شاهدی تایید شده است.

برندل[3] [۱۵] (۲۰۰۴) دو روش جانهی بیزی مختلف را با چهار روش عمومی جانهی مقایسه کرده است. دو روش بیزی، روش بیز تجربی[4] و روش الگوریتم حداکثر انتظار[5] می باشند. چهار روش عمومی جانهی، روش های LOCF،

1- Stamey!
2. Sinha!
3.!Brandel
4 . Empirical Bayes method
5 . Expectation Maximization algorithm

بدترین حالت[۱]، بهترین حالت[۲] و روش مقدار میانگین[۳] می باشند. پس از مقایسه، برندل نتیجه گیری کرده است که روش های بیزی نسبت به روش های کلاسیک جانهی منجر به نتایج بهتری می شوند.

کریملو[۴] و همکاران [۱] (۲۰۰۶) در پژوهش خود نیز نتیجه گرفته بودند که روش بیز از روش ماکسیمم درستنمایی و واحد کامل نتایج دقیق تری را منجر می شوند که مشابه نتایج گرفته شده در این پژوهش می باشد.

جوست[۵] و همکاران [۱۶] (۲۰۰۷) دو روش جانهی چندگانه TW-DA[۶] بیزی و W-E[۷] را برای داده های گمشده در پرسشنامه ها مقایسه کرده اند و نتیجه گرفته اند که روش بیزی بهتر است و منجر به نتایج ناریب در تحلیل واریانس دو طرفه[۸]، تحلیل مولفه اصلی[۹] و آلفای کرونباخ[۱۰] می شود

نتیجه گیری

بنابراین درصورت مواجه شدن باداده های گمشده اولین قدم بازنگری ومشاهده مجدد واحدهای مورد مطالعه وتکمیل مقادیر گمشده است. درمرحله بعد می بایست با محاسبه احتمالات گمشدگی درسطوح مختلف متغیرها نسبت به تشخیص گمشدگی تصادفی MAR اطمینان حاصل نمود. درصورت مثبت بودن تشخیص، برای انجام تحلیل رگرسیون لو جستیک، بر اساس یافته های این تحقیق، استفاده ازمدل های تعمیم یافته ساتن و کارول به همراه تحلیل بیزی با تکنیک زنجیرهای مارکوف مونت کارلویی (MCMC)، روی توزیع پسینی که ازترکیب تابع درستنمایی تعمیم یافته ساتن و کارول وتوزیع پیشین ناآگاهی بخش حاصل می شود، به لحاظ اجتناب از نتیجه گیری های نادرست، توصیه می گردد. و می توان به برآوردهای دقیق تر درمقایسه با حالت حذف آزمودنی ها ازمطالعه و نیز روش جانهی دست یافت.

تشکر و قدردانی

با تشکر از سرکار خانم دکتر ایمانه عسگری که داده های مورد استفاده در این پژوهش را در اختیار اینجانب قرار داده اند.

1 . worst case
2 . best case
3 . mean value method
4!. Karimlou
5!. Joost
6.Two Way Data Augmentation
7 .Two Way Estimation
8.Two Way ANOVA
9. Principal Component Analysis
10. Cronbach’s alpha

References

1. Karimlou M, Jandaghi G.R, Mohammad K, Wolfe R, Azam K, A Comparison of Parameter Estimates in Standard Logistic Regressin Using WinBUGS MCMC and MLE Methods in R for Different Sample Size,Far East J,Theo,stat 2006; 19: 281-292[Persian]
2. Gao S, Hui S.L, Logistic Regression Models With Missing Covariate Values For Complex Survey Data, Statistics In Medicine 1997; 16: 2419-2428.
3. Satten G.A, Carroll R.J, Conditional and Unconditional Categorical Regression Models with Missing Covariates, Biometrics 2000 ; 56:384-388.
4. Fleiss J.L, Levin B, Paik M.C, Statistical Methods For Rates And Proportions , 3rd ed, John Wiley & Sons 2002; ISBN 0-471-52629-0.
5. Shi X, Zhu H, Ibrahim J.G, Local Influence for Generalized Linear Models with Missing Covariates, Biometrics 2009 ;65:1164-1174.
6. Satten G.A, Kupper L, Inferences about Exposure–Disease Associations Using Probability of Exposure Information, statists 1993; 88:200-208.
7. Satten G.A, Kupper L, Conditional Regression Analysis Of The Exposure-Disease Odds Ratio Using Known Probability Of Exposure Values, Biometrics 1993; 44:429-440.
8. Little R.J.A, Rubin D.B, Statistical analysis with Missing data, 2nd ed, New York: John Wiley & sons 2002, ISBN 978-0471183860.
9. Enders C.K, Applied Missing Data Analysis, New York and London: Guilford Press 2010 ISBN 978-1-60623-639-0.
10. Dixon W.J, BMDP statistical software, Los Angeles: University of California Press 1988 ISBN 0-520-06473-9.
11. Little R.J.A, A test of missing completely at random for multivariate data with missing values, JASA 1988; 83, 1198–1202.
12. SPSS Missing Value Analysis™ 17. SPSS Inc 2007. Printed in the United States of America. Available from: URL: http://www.spss.com (Accessed: 21 June 2010).
13. Stamey JD, Bekele B.N, Powers S, Bayesian Modeling of Follow-up Studies with Missing Data.Elsevier Inc: Annals of Epidemiology 2009; 16: 416-422
14. Sinha S, Mukherjee B, Ghosh M, Bayesian Semiparametric Modeling For Matched Case-Control Studies With Multiple Disease States , Biometrics 2004; 60: 41-49.
15. Brandel J, Empirical Bayes methods for missing data Analysis. June 2004. Available from: URL: http://www2.math.uu.se/research/pub/Brandel.pdf (Accessed: 19 February 2011).
16. Joost R.V, L. Van der Ark V, Sijtsma K, Vermunt JK, Two-way imputation: A Bayesian method for estimating missing scores in tests and questionnaires and an accurate approximation, Elsevier: Computational Statistics & Data Analysis 2007; 51:4013 – 4027.
17. Marwala T, Computational Intelligence for Missing Data Imputation, Estimation and Management: Knowledge Optimization Techniques, South Africa: University of Witwatersrand IGI Global 2009 ISBN 978-1-60566-336-4.

Application of Bayesian Method in Parameters Estimation of Logistic Regression Model with Missing at Random Covariate

Kazemi E[1], Karimlo M[2], Rahgozar M[2], Bakhshi E[3], Asgari E[4]*

[1]M.Sc of Biostatistics, University Of Social Welfare and Rehabilitation!
[2]Associated Professor, Department of Biostatistics, University Of Social Welfare and Rehabilitation
[3]Assistant Professor, Department of Biostatistics, University Of Social Welfare and Rehabilitation
[4]Dentist, School Of Dentistry, Shahid Beheshti University of Medical Sciences

***Corresponding Author**: Evin, Daneshjo Boulevard, Kodakyar Street, University Of Social Welfare and Rehabilitation
Email: mkarimlo@yahoo.com*

Abstract

Background & Objectives: *Logistic Regression is a general model for medical and epidemiological data analysis. Recently few researchers have directed their studies to analysis of Logistic Regression with missing value at covariate variable. While the missing is a major threat in results authenticity of data set, in many studies the researchers face data with missing value and it is difficult to avoid such a case in studies.*

Material & Methods: *Satten and Carroll, in the case of completely observed value of covariate variable and some covariate variable with missing at random mechanism (MAR), introduced a special likelihood function for parameters estimation of Logistic Regression model. In this research the above- mentioned likelihood function has been used in Bayesian analysis for parameters estimation of Logistic Regression model and the conclusions are compared with the Multiple Imputation method and Complete Case method.*

Results: *The above-mentioned methods were applied on both simulation data and dentistry data and concluded that The parameters estimation from SCMCMC method had less variance in comparison with parameters estimation from Multiple Imputation and Complete Case methods.*

Conclusion: *After comparison of the three mentioned methods results it had been concluded that if the mechanism is of missing at random the application of Bayesian analysis with MCMC causes to more accurate estimation and shorter Confidence Intervals than the Multiple Imputation method and Complete Case.*

Key words: *Logistic Regression, Missing at Random (MAR), Bayesian Analysis, Markov Chain Monte Carlo, Multiple Imputation, DMFT.*