مجله دانشکده بهداشت و انستیتو تحقیقات بهداشتی
دوره ۴ شماره ۲ تابستان ۱۳۸۵ ، صفحات : ۳۳-۲۱

مقایسه روش بیزی(Bayesian) وکلاسیک دربرآورد پارامترهای مدل رگرسیون لجستیک با وجود مقادیر گمشده درمتغیرهای کمکی

مسعود کریملو: دانشجوی دکترای آمار زیستی، دانشکده بهداشت و انستیتو تحقیقات بهداشتی دانشگاه علوم پزشکی تهران و عضو هیئت علمی دانشگاه علوم بهزیستی و توانبخشی- نویسنده رابط : mkarimlo@uswr.ac.ir
دکترکاظم محمد: استاد گروه آمار زیستی و اپیدمیولوژی دانشکده بهداشت و انستیتو تحقیقات بهداشتی دانشگاه علوم پزشکی تهران
دکترمحمدرضا مشکانی: استاد گروه آمار دانشکده علوم ریاضی دانشگاه شهید بهشتی
غلامرضاجندقی: عضو هیات علمی دانشگاه تهران ، پردیس قم
دکترکرامت ا...نوری : استادیار دانشکده بهداشت و انستیتو تحقیقات بهداشتی دانشگاه علوم پزشکی تهران
دکترعین ا...پاشا : دانشیار دانشکده ریاضی دانشگاه تربیت معلم
دکترکمال اعظم: استادیار دانشکده بهداشت و انستیتو تحقیقات بهداشتی دانشگاه علوم پزشکی تهران
دریافت : ۸۴/۳/۱۶ پذیرش : ۸۴/۳/۳۰

چکیده :

زمینه و هدف: رگرسیون لجستیک ابزاری تحلیلی است که بطوروسیعی درتحقیقات پزشکی و اپیدمیولوژیک کاربرد دارد. دربسیاری از مطالعات بامجموعه داده هایی مواجه می شویم که بخشی از آنها گزارش نشده اندیا به عبارت دیگر گمشده می باشند. ساده ترین روش برای تجزیه وتحلیل چنین داده هایی صرفنظر کردن ازموردهای دارای مقادیر گمشده وادامه آنالیز باداده های کامل می باشد که این روش، درعمل کارآمد نیست.

روش کار: دراین مطالعه روشی برای آنالیز مدلهای رگرسیون لجستیک وقتی که مقادیر متغیر کمکی Z بطور کامل مشاهده شده ومقادیر متغیر کمکی X برای برخی از افراد تحت مطالعه گمشده باشند ارائه شده است. وقتی مقادیر X بطور تصادفی گمشده (MAR) باشند، مدل تابع درستنمایی ارائه شده برای کل داده های مشاهده شده موجود ، همانند زمانی که داده ها کامل وبدون گمشدگی اند ،عمل می کند.

نتایج: با بکاربردن این تابع درستنمایی ، برآورد پارامترها هم به روش کلاسیک ماکزیمم درستنمایی وهم به روش بیزی بااستفاده از تکنیک زنجیره های مارکوف مونت کارلویی (MCMC) انجام ومقادیر بدست آمده مقایسه گردیده اند. داده های این مطالعه مربوط به اطلاعات سمع ریه بدست آمده از طرح سلامت وبیماری درشهرتهران می باشد.

نتیجه گیری: برآوردهای حاصل از مدل ارائه شده دراین مطالعه وآنالیز آن به روش بیزی نسبت به سایر روشهای برآورد ، مقادیر نزدیکتری به برآوردهای مدل استاندارد برای داده های کامل داشتند.

واژگان کلیدی :رگرسیون لجستیک، گمشدن تصادفی (MAR) ، ماکزیمم درستنمایی ، زنجیره های مارکوف مونت کارلویی (MCMC) ، روش بیزی ، اختلالات تنفسی

مقدمه :

دردهه های اخیر مدلهای رگرسیون لجستیک نقش مهمی را در تحلیل داده های پزشکی و اپیدمیولوژیک و تحقیقات اجتماعی داشته اند. بدلیل غیر خطی بودن مدل رگرسیون لجستیک، تحلیل آن ازطریق ماکزیمم درستنمایی صورت می پذیرد ؛ چرا که درآن هیچگونه

محدودیتی برای متغیرهای مستقل درنظر گرفته نمی شود. در مطالعات کوهورت یا مطالعه مقطعی که در آنها جورکردن ملاک عمل نیست از روش ماکزیمم درستنمایی غیرشرطی اقدام به برآورد پارامترها می‌گردد (Kleinbaum D.G. , Klein M. ۲۰۰۲). به علاوه در اینگونه مطالعات با مواردی مواجه می شویم که در آنها بخشی از داده ها بدلایلی ازقبیل، خودداری از پاسخ، عدم تکمیل کامل پرسشنامه یا پرونده ، ناقص بودن چارچوب مطالعه و غیره گزارش نشده اند.

بطور کلی سه نوع مکانیزم گمشدگی وجود دارد: گمشدگی کاملا تصادفی MCAR (Missing Completely At Random)، گمشدگی تصادفی MAR (Missing At Random) ، و گمشدگی غیرقابل اغماض NI (Non Ignorable) (LittleR.J.A. and Rubin D.B. ۲۰۰۲). درمکانیزم گمشدن MCAR ،گمشدگی دریک متغیر به خود آن متغیر یا متغیرهای دیگر بستگی ندارد و لذا می توان آزمودنی با مقادیر گمشده را ازمطالعه حذف و آنالیزرا براساس حجم نمونه جدید اجرا نمود بدون اینکه دربرآورد ها اریبی ایجاد شود.

درمکانیزم گمشدگی MAR ،گمشدگی در متغیرکمکی (Covariate) بستگی به خود متغیر نداشته بلکه به متغیرهای دیگربستگی دارد. به عنوان مثال در بررسی ارتباط بین فشارخون و سیگارکشیدن ،گمشدگی تصادفی درمقادیر متغیر فشارخون بستگی به متغیر سیگارکشیدن دارد ولی به خود متغیر فشار خون بستگی ندارد ؛ به این معنی که درمقایسه با سیگاریها ، غیر سیگاریها بدلیل آگاهی و علاقمندی نسبت به وضعیت سلامت جسمی خود ، بیشتر به شرکت درمطالعه و پاسخگویی به سوالات تمایل دارند (Fleiss J.L. et al ۲۰۰۳).

با توجه به مثال فوق ملاحظه می شود که مشکل گمشدگی می تواند تجزیه و تحلیل های آماری رابه سوی استنباطهای اریب سوق دهد و نهایتا دستیابی به یک نتیجه گیری مفید از داده های جمع آوری شده را با مشکل مواجه سازد. گمشدگی می تواند در متغیرهای پاسخ یا متغیرهای کمکی رخ دهد. دراین مطالعه، گمشدن درمقادیر متغیرهای کمکی موردنظر بوده و مکانیزم آن ازنوع گمشدگی تصادفی (MAR) می باشد.

در نحوه برخوردبا داده های گمشده عمدتا سه روش مورد استفاده قرار می گیرد (Gao and Hui ۱۹۹۷) . ساده ترین روشی که درپیش فرض اغلب نرم افزارهای آماری لحاظ گردیده ، حذف آزمودنیهای با مقادیر گمشده است و انجام آنالیز براساس داده های کامل(Complete Cases) می باشد که عمدتا سبب ایجاد اریبی دربرآورد ها می گردد(Little R.J.A. and Rubin D.B. ۲۰۰۲). در روش دوم بوسیله «میانگین»، یا بوسیله «رگرسیون» و یا سایر روشهای متعارف، برآوردی برای داده گمشده بدست آورده ، جایگزین آن می کنیم و سپس با روشهای استاندارد اقدام به آنالیز می نمائیم. این روش درصورت بالا بودن تعداد آزمودنیهای گمشده ، دارای دو اشکال عمده می باشد : اولا شکل طبیعی توزیع متغیر دارای مقادیر گمشده تغییر می یابد و ثانیا ، میانگین، واریانس و خطای معیار برآورد(توابع نمونه ای) بدلیل اضافه شدن مقادیری یکسان تغییر خواهند یافت. روش دیگری که امروزه نظر محققان را بخود جلب نموده ، عبارت است از تعیین مدل احتمال برای متغیر دارای مقادیر گمشده که طرز عمل آن با مدلهای استاندارد یکی است و تنها شامل تغییراتی درتابع درستنمایی می باشد.

در آمار کلاسیک ، تحلیل مدل های رگرسیون لجستیک مبتنی بر برآورد پارامترها ازطریق ماکزیمم نمودن تابع درستنمایی (Maximum Likelihood Estimation) MLE و محاسبه برآوردها به روش الگوریتم تکـــــراری (Expectation Maximization Algorithm)EM است. این روش علاوه برمشکلات محاسباتی ازمشکلات تکنیکی نیز برخوردار است ، به این ترتیب که ممکن است بجای ماکزیمم کلی تابع درستنمایی ، ماکزیمم موضعی بدست آید ویا اساسا همگرایی برآوردها حاصل نگردد. بعلاوه باتوجه به خواص مجانبی برآوردهای

ماکزیمم درستنمایی، در نمونه های کوچک با مشکلات استنباطی جدی روبرو هستیم (خیری و همکاران ۱۳۸۲).

در روش بیـــزی (Bayesian) اســـتنـــباط درباره پارامترهای مدل برمبنای توزیع پسین (Posteriordistribution) آنها صورت می پذیرد که تلفیقی است از داده های مشاهده شده و اطلاعات ناشی از مطالعات قبلی و یـــا تجارب شخصی که با عـــنوان توزیع پیشین(Prior distribution) شناخته می شود درصورت نامعلوم بودن توزیع پسین، می توان بابهره گیری ازروشهای شبیه سازی، زنجیره های مارکوف مونت کارلویی MCMC (Markov Chain Monte Carlo) برای هرحجم نمونه دلخواه ،استنباط دقیقی ازپارامترها بدست آورد.

در مطالعه حاضر ، از روش سوم وتعیین مدل احتمالی برای متغیر دارای مقادیر گمشده استفاده شده و برای این منظور ، تابع درستنمایی معرفی شده توسط ساتن و کارول (Satten G.A. and Carroll R.J. ۲۰۰۰) برای مدل رگرسیون لجستیک بکارگرفته شده است و نیز با هر دو تکنیک MLE وMCMC پارامترهای مورد نظر برآورد و با یکدیگر و با حالت برآوردهای بدون گمشدگی مقایسه شده اند.

آمارشناسان کلاسیک و بیزی تلاشهای گسترده ای را دربرخورد با مساله گمشدگی درمدلهای رگرسیون لجستیک انجام داده و مقالاتی نیز به چاپ رسانده اند که عموما درارتباط با مطالعات مورد- شاهدی می باشد. فاکس (Fuchs ۱۹۸۲) و لیتل و شلاختر (Little .J.A. and Schluchter M.D. ۱۹۸۵) با استفاده از الگوریتم EM ، به برآورد ماکزیمم درستنمایی باوجود مقادیر گمشده درمتغیرهای کمکی گسسته یا ترکیبی ازمتغیرهای گسسته و پیوسته دررگرسیون لجستیک اقدام نموده اند، که این روش عموما به تکرار نیاز داشته وگاهی به جهت محاسباتی پیچیده است . بلاک هورست و شلاختر (Blackhurst D.W. and Schluchter M.D. ۱۹۸۹) پیشنهاد داده اند که روش ماکزیمم درستنمایی بااستفاده از الگوریتم EM هنگامی که متغیر کمکی پیوسته دارای مقادیر گمشده ازتوزیع نرمال پیروی کند، به تکرار نیاز ندارد . به علاوه این دونویسنده روش مونت کارلو را هنگامیکه دومتغیر کمکی در کار باشد که یکی با مشاهدات کامل و دیگری دارای مشاهدات گمشده باشند، موردبررسی قرار دادند.آنها روش داده های کامل را باروش جانهی مقادیرگمشده و روش درستنمایی مقایسه نمودند و نتیجه گرفتند که روش درستنمایی ،زمانی که مدل به طور مناسبی معین شده باشد، بهتر عمل می نماید. ساتن وکوپر (Satten G.A.and Kupper L. ۱۹۹۳a,b) روش تحلیل رگرسیون لجستیک را وقتی که متغیر کمکی دارای مقادیر گمشده بود ، بسط دادند و از متغیرهای جانشین برای پیداکردن اطلاعی از اثر متغیرهای دارای مقادیر گمشده درمدل استفاده کردند. پیک و ساکو (Paik M.C. and Sacco R.L. ۲۰۰۰). ساتن و کارول (Satten G.A. and Carol S. ۲۰۰۰) در مقاله دیگری با تعیین توزیعی برای متغیر کمکی دارای مقادیر گمشده واعمال تغییراتی در توابع درستنمایی رگرسیون شرطی و غیر شرطی، برآورد پارامترها را بهبود بخشیدند. راتوز و همکاران (Rathouz P.J. et al. ۲۰۰۳) رده جدیدی از برآوردها را ارائه نمودند که براساس مدل بندی توزیع متغیرهای کمکی دارای داده های گمشده و مدل بندی روند گمشدن مقادیر متغیر کمکی پایه گذاری شده است.

آمار شناسان بیزی نیز در این زمینه مقالاتی ارائه نموده اند ؛ ازجمله : زلن و پارکر(Zclen M. and Parker R.A. ۱۹۸٦)، نورمینن و موتانن (Nurminen M. and Mutanen P. ۱۹۸۷) و اشبی وهمکاران (Ashby D. et al. ۱۹۹۳) ، روش بیزی را در مطالعات مورد- شاهدی هنگامی که متغیر عامل خطر دوحالتی بوده و اثر طبقات نیز ثابت فرض شوند، برای داده های کامل بررسی کرده اند. مولرو رویدر (Muller P. and Roeder K. ۱۹۹۷) و مولر و همکاران (Muller P. et al. ۱۹۹۹) مدل نیمه پارامتری را برای مطالعه مورد شاهدی جورنشده بامتغیرهای همراه پیوسته واحتمالا دارای مقادیر گمشده مورد بررسی قرار دادند.

سی من و ریچارد سون(Seaman S.R. and Richardson S. ۲۰۰۱) روش مولر وروی در را برای متغیرهای کمکی گسسته گسترش دادند و ارتباطی بین روشهای مولر و رویدر وزلن و پارکر هنگامی که خطای اندازه گیری وجود ندارد،برقرارکردند. سینها و همکاران (Sinha S. et al. ۲۰۰٤) در مطالعات مورد- شاهدی هنگامی که متغیر بیماری چند وضعیتی و متغیرهای کمکی دارای مقادیر گمشده اند، تابع درستنمایی رگرسیون لجستیک شرطی ساتن و کارول را تعیمم دادند و با روش MCMC، استنباط درمورد پارامترهای مدل را به انجام رساندند.

هیچ یک از مطالعات انجام شده ، تحلیل مدل رگرسیون لجستیک غیر شرطی مربوط به مطالعات مقطعی را وقتی داده های گمشده درمتغیرهای کمکی وجودداشته باشند ، به روش کلاسیک و بیزی را مورد مقایسه قرار نداده اند.

روش کار:

مدل وتعاریف : فرض کنید متغیر Y_i نشان دهنده متغیر پاسخ دووضعیتی متناظربا فرد i ام باشد، بطوریکه Y_i=۱ به معنی بیمار بودن فردi ام و Y_i=۰ به معنی سالم بودن فرد i ام باشد. به علاوه ، فرض می کنیم بردار متغیر کمکی **Z** دارای مقادیر بطور کامل مشاهده شده وبردار متغیر کمکی **X** احتمالا دربرخی مشاهدات دارای مقادیر گمشده است. بدون اینکه ازکلیت مساله کاسته شود ، مدلهای خود را براساس یک متغیر Z ویک متغیر X ارائه داده ، در ابتدا فرض می کنیم که هر دو متغیر با مشاهدات کامل و بدون گمشدگی باشند. دراین حالت احتمالات شرطی متغیربیماری به شرط متغیرهای کمکی ، با استفاده ازمدل لجستیک به صورت روابط ۱ الی ٤ تعریف می شوند، که درآن x' و z' سطوحی ازمتغیر X ومتفاوت ازسطوح x وz هستند. هدف از تحلیل رگرسیون لجستیک بدست آوردن برآوردی از پارامترهای مدل (دراینجا : β_0 , β_1 , β_2 , β_{12}) برای تبیین ارتباط بین متغیر پاسخ Y ومجموعه ای از متغیرهای کمکی X و Zمی باشد. درصورت کامل بودن مقادیر متغیرهای X و Z ازروشهای استاندارد جهت برآورد این پارامترها استفاده می شود. اما اگر فرض کنیم که برخی از مشاهدات متغیر X گمشده باشند دراین صورت داریم:

$$P\left(Y=1 \mid X=x, Z=z\right)=\frac{e^{\beta_0+\beta_1 x+\beta_2 z+\beta_{12}xz}}{1+e^{\beta_0+\beta_1 x+\beta_2 z+\beta_{12}xz}} \qquad (۱)$$

$$P\left(Y=0 \mid X=x, Z=z\right)=\frac{1}{1+e^{\beta_0+\beta_1 x+\beta_2 z+\beta_{12}xz}} \qquad (۲)$$

بنابراین با تعاریف فوق بخت (Odds) ونسبت بخت بیماری (Odds Ratio) به ترتیب عبارت خواهند بوداز:

$$\theta\left(x,z\right)=\frac{P\left(Y=1 \mid X=x,Z=z\right)}{P\left(Y=0 \mid X=x,Z=z\right)}=e^{\beta_0+\beta_1 x+\beta_2 z+\beta_{12}xz} \qquad (۳)$$

$$\psi\left(x,z,x',z'\right)=\frac{\theta(x,z)}{\theta(x',z')} \qquad (٤)$$

$$\tilde{\theta}(Z)=\frac{P\left(Y=1 \mid Z=z\right)}{P\left(Y=0 \mid Z=z\right)} \qquad (٥)$$

بعلاوه می توان تعریف کرد :

$$\pi\left(x \mid z\right)=P\left(X=x \mid Y=0, Z=z\right) \qquad (٦)$$

$$\rho\left(x \mid z\right)=P\left(X=x \mid Y=1, Z=z\right) \qquad (۷)$$

توابع π و ρ به ترتیب نشان دهنده توزیع احتمال مقادیر متغیر X در افراد سالم و بیمار می باشند. نتیجه مطالعه ساتن وکوپر با استفاده از قضیه بیز عبارت است از (Satten G.A. and Kupper L. ۱۹۹۳ a,b):

$$\tilde{\theta}(z)=\sum_{x}\theta(x,z).\pi(x\mid z) \qquad (۸)$$

که درآن، مجموع روی همه مقادیر ممکن متغیر X صورت می پذیرد. دومین نتیجه مقاله مذکور عبارتست از:

$$\rho(x\mid z)=\frac{\pi(x\mid z)\,\theta(x,\,z)}{\sum_{x}\pi(x\mid z)\,\theta(x,z)} \qquad (۹)$$

اگر Xمتغیر پیوسته باشددراینصورت به جای مجموعها باید از انتگرالها استفاده شود وعبارتهای $\pi(x\mid z)$ و $\rho(x\mid z)$ نیز توابع چگالی احتمال درنظر گرفته شوند.

تابع درستنمایی، توزیع پیشین وتوزیع پسین:
می توان نتایج بخش قبل را برای تعریف تابع درستنمایی جهت برآورد پارامترهای مورد نظر بوسیله داده های مشاهده شده، بکار برد. تابع درستنمایی رگرسیون لجستیک استاندارد درحالتیکه مقادیرX وZ بطورکامل مشاهده شده باشند، عبارتست از:

$$L(\beta)=\prod_{i=1}^{n}\frac{[\theta(X_i,Z_i)]^{y_i}}{1+\theta(X_i,Z_i)} \qquad (۱۰)$$

درصورتی که متغیرکمکی X دارای مقادیر گمشده باشد ، متغیر نشانگر Δ_i را به این صورت تعریف می کنیم : Δ_i=۱ اگر x_i مشاهده باشد و Δ_i=۰ اگر x_i مشاهده نشده باشد . بااین تعریف ، می توان تابع درستنمایی داده های کامل را بصورت زیر بازنویسی کرد:

$$P(Y,X,\Delta\mid Z)=P(Y\mid Z)\;P(\Delta\mid Y,Z)\;P(X\mid Y,Z,\Delta) \qquad (۱۱)$$

تحت شرط گمشدن تصادفی MAR می توان فرض کرد که ، $P(X\mid Y,\,Z,\,\Delta)=P(X\mid Y,Z)$ (Little and R.J.A. and Rubin D.B. ۲۰۰۲)، همچنین فرض می کنیم که احتمال گمشدگی $P(\Delta\mid Y,Z)$ بستگی به بردار پارامترها ، $\boldsymbol{\beta}$ ندارد.[۹] بنابراین باحذف این عبارت ازدرستنمایی فوق ،درستنمایی غیر شرطی اصلاح شده برای داده های مشاهده شده عبارت خواهد بوداز:

$$L(\beta)=\prod_{i-1}^{n}\tilde{\theta}(Z_i)^{Y_i}\left[1+\tilde{\theta}(Z_i)\right]^{-1}\times \qquad (۱۲)$$
$$\pi(X_i\mid Z_i)^{\Delta_i(1-Y_i)}\;\rho_1(X_i\mid Z_i)^{\Delta_i Y_i}$$

درارابطه فوق توزیع احتمال $\pi(x\mid z)$ نامعلوم می باشد، درصورتیکه X وZ مقادیر شمارا و محدودی را اختیار کنند توزیع مناسبی که می توان برایπ (۰) مفروض دانست عبارتست از (Satten G.A. andCarol R.J. ۲۰۰۰) :

$$\pi(x\mid z)=\frac{e^{\gamma xz}}{\sum_{x'}e^{\gamma x'z}}=\frac{e^{\gamma_0+\gamma_1 x+\gamma_2 z+\gamma_{12}xz}}{\sum_{x'}e^{\gamma_0+\gamma_1 x'+\gamma_2 z+\gamma_{12}x'z}}=\frac{e^{\gamma_1 x+\gamma_{12}x\,z}}{\sum_{x'}e^{\gamma_1 x'+\gamma_{12}x'z}} \qquad (۱۳)$$

بابکارگیری روابط (۳) تا (۹) در رابطه (۱۳) وبازنویسی مجدد تابع درستنمایی (۱۲) داریم :

$$l(\beta|X,Z)=\prod_{i=1}\frac{\left(e^{\beta_0+\beta_1x_i+\beta_2z_i+\beta_{12}x_iz_i}\right)^{\Delta_i y_i}}{1+\sum_x\left[e^{\beta_0+(\beta_1+\gamma_1)x+\beta_2 z_i+(\beta_{12}+\gamma_{12})xz_i}\cdot\frac{1}{\sum_x e^{\gamma_1x+\gamma_{12}xz_i}}\right]} \qquad (۱٤)$$

$$\times\left\{\sum_x e^{\beta_0+(\beta_1+\gamma_1)x+\beta_2 z_i+(\beta_{12}+\gamma_{12})x z_i}\cdot\frac{1}{\sum_x e^{\gamma_1x+\gamma_{12}xz_i}}\right\}^{y_i(1-\Delta_i)}\times\left(\frac{e^{\gamma_1x_i+\gamma_{12}x_iz_i}}{\sum_x e^{\gamma_1x+\gamma_{12}xz_i}}\right)^{\Delta_i}$$

که تابعی ازپارامترهای $\beta_0, \beta_1, \beta_2, \beta_{12}, \gamma_1, \gamma_{12}$ می باشد. با لگاریتم گیری ازتابع فوق ومشتق گیری ازآن نسبت به تک تک پارامترهاومساوی صفرقراردادن آنها دستگاه معادلات عددی حاصل می گردد که بدلیل غیر خطی بودن معادلات بابکارگیری روشهای عددی اقدام به محاسبه MLE می نمائیم. برآوردهای حاصل از بکارگیری درستنمایی مذکور بسیار کارآتر از روش استاندارد است که درآن از سایر اطلاعات آزمودنیهای دارای مقادیر گمشده صرف نظرمی کنیم.

برای انجام تحلیل بیزی (Bayesian) باروش MCMC ، لازم است توابع چگالی شرطی پارامترهای مدل یعنی $\boldsymbol{\beta}=(\beta_0, \beta_1, \beta_2, \beta_{12}, \gamma_1, \gamma_{12})$ تعیین شود وسپس بطورمتوالی ازآنها نمونه گیری به عمل آید. دراین مطالعه، توزیع پیشین (Prior distribution) پارامترهای مدل را ناآگاهی بخش باتوزیع $\beta_i \sim N(\mu=0, \sigma^2=10^6)$ درنظرگرفته و توزیع احتمال مشاهدات راتابع درستنمایی (۱٤) ، $L(\beta|x,z)$ قرار دادیم. برای استنباط بیزی لازم است توزیع توام مشاهدات وپارامترها را بصورت زیر محاسبه نمائیم.

$$\Pi(X,Z,\beta)=L(\beta|X,Z)\,\Pi(\beta) \qquad (۱٥)$$

با فرض معلوم بودن مشاهدات (X,Z) می توان به کمک قضیه بیز توزیع β به شرط (X,Z) رابدست آورد (Gilks et l ۱۹۹۷) :

$$\Pi(X,Z,\beta)=L(\beta|X,Z)\,\Pi(\beta) \qquad (۱٦)$$

که به آن توزیع پسین(Posterior distribution) می گوئیم. توزیع پسین عقیده مارا درباره پارامترهای مجهول پس از مشاهده داده ها ، نشان می دهد. دراستنباط بیزی تنها وتنها توزیع پسین پایه استنباط درباره پارامترهاست.

دراین مطالعه باضرب نمودن چگالی توزیع نرمال هرپارامتردر درستنمایی (۱٤) شکل توزیع پسین حاصل ، نامعلوم بود وعملا ناچار به استفاده از الگوریتم متروپولیس(Metropolis)، با معرفی توزیع پیشنهادی (Proposal distribution) نرمال شدیم . بنابراین درهرمرحله ازشبیه سازی MCMC به ترتیب ازتوابع چگالی شرطی کامل هرپارامتر $\beta_0, \beta_1, \beta_2, \beta_{12}, \gamma_{12}$ نمونه گیری به عمل آمد وبراساس آن استنباط درباره پارامتر مذکور صورت پذیرفت.

نتایج:

مقایسه برآورد پارامترها به روش MCMC و MLE: در بخش قبل، متدولوژی جامعی از دو روش بیزیMCMC و ماکزیمم درستنمایی برای برآورد پارامترهای مدل رگرسیون لجستیک غیر شرطی بیان گردید. حال کاربرد آن بر روی داده های یک مطالعه مقطعی و سپس مقایسه برآوردهای ناشی از آن نشان داده می شود. دراین بخش باارائه مثالی ساختگی ،بحث را آغاز می کنیم ودربخش بعد با یک نمونه واقعی ازطرح سلامت وبیماری مطلب را ادامه خواهیم داد.جدول زیر راکه درآن x وz متغیرهای کمکی وd متغیر وابسته یا پاسخ است در نظر بگیرید:

x	z	d	n
۰	۰	۲۸۰	٦٦۰
۰	۱	٤۰۰	٦٦۰
۱	۰	۲۲۰	۳٤۰
۱	۱	۱۸۰	۳٤۰

به عنوان مثال سطر اول این جدول نشان دهنده این است که ازکل ٦٦۰ نفر موجود در طبقه , x=۰ z=۰تعداد ۲۸۰ نفر بیمار می باشند.به این ترتیب ، یک نمونه ۲۰۰۰ تایی خواهیم داشت که ازجهاتی با داده ها ی نوعی مطالعه حقیقی مشابه است که در آن توزیع متغیر z به صورت ۵۰٪-۵۰٪ وتوزیع متغیر x به صورت ٦٦٪- ۳٤٪ و کلیه مقادیر نسبت بخت (OR) بین ۲ تا۳ می باشند.

برای آنالیزداده های فوق ازسه بسته نرم افزاری متداول S-Plus ۲۰۰۰ ، R نسخه ۱/۹ و WinBUGS نسخه ٤/۱ استفاده گردید. WinBUGS نرم افزار توانمندی است که درسالهای اخیر توسط اشپیگل هالتر وهمکاران(Spiegelhalter D. et al. ۲۰۰۳) جهت اجرای استنباط های بیزی باتکنیک MCMC طراحی و به صورت آزاد دراختیار محققین و بخصوص آمارشناسان بیزی قرارگرفته است.

در مجموع با پنج روش پارامترهای $\beta_{12}, \beta_2, \beta_1, \beta_0$ با استفاده از مجموعه داده های فوق ونرم افزارهای مذکور بر آورد گردیدند. که نتایج آن درجدول ۱ درج شده است. ضمنا از ذکر مقادیر برآوردشده برای پارامترهای γ_1, γ_{12} بدلیل اینکه اصولا نقش کاربردی درتفسیر نتایج ندارند خودداری شده است.

از چپ به راست مقادیر ستونهای یک تا شش این جدول به شرح زیر می باشند:

ستون یک - شامل نام پارامترها است.

ستون دو- شامل برآوردهای ماکزیمم درستنمایی(MLE) داده های کامل بدون گمشدگی (Full Data) بدست آمده ازS-Plus می باشد،که به اختصار باعنوان FMLE نامیده شده است. (n=۲۰۰۰).

ستون سه - شامل برآوردهای بیزی با تکنیکMCMC برای داده های کامل بدون گمشدگی است که با نگارش برنامه ای درمحیط WinBUGS وبا فرضهای مطرح شده دربخش قبل، محاسبه شده اند و به اختصار باعنوان FMCMC نام گرفته است .

ستون چهار- شامل برآوردهای ماکزیمم درستنمایی داده های کامل باحذف آزمودنیهای دارای مقادیر گمشده از مجموعه کل افراد CC (Complete Case) وآنالیز روی داده های باقی مانده است که با S-Plus محاسبه شده و باعنوان CCMLE نام گذاری شده است. ضمنا مکانیزم گمشدگی درمتغیر x دراین مرحله ازنوع گمشدگی کاملا تصادفی MCAR بود و لذا ۳۰٪ از داده های متغیرx با این روش حذف شدند و آنالیز باحجم نمونه ۱٤۰۰ آانالیز صورت پذیرفت.

ستون پنج- شامل برآوردهای ماکزیمم درستنمایی MLE داده های کامل و با حذف آزمودنی های دارای مقادیر گمشده (CC) است؛ که با استفاده ازتابع درستنمایی تعمیم یافته ساتن وکارول (Satten G.A. and Carol S. ۲۰۰۰)، رابطه (۱٤) ، و بانگارش برنامه ای درمحیط

R حاصل و به اختصار با (SCMLE) نام گذاری شده است.

ستون شش – شامل برآوردهای بیزی با تکنیک MCMC برای داده های کامل و با حذف آزمودنی های دارای مقادیر گمشده (CC) است؛ که بااستفاد از درستنمایی ساتن و کارول ،رابطه (۱٤) و با نگارش برنامه ای درمحیط WinBUGS بدست آمده اند و به اختصار با عنوان SCMCMC ذکر شده است.

شایان ذکر است کـه بـرای بدسـت آوردن برآوردهـای ستونهای ٤ تا ٦ جدول ۱ ، ابتدا ۳۰٪ از مقادیر متغیر x درفایل ۲۰۰۰ تایی به تعداد ٤۰بار مسـتقل از هـم ، بطـور کاملا تصادفی (MCAR) حـذف وسـپس میـانگین و خطای معیار ٤۰ پارامتر درجدول درج شده است.

همانگونه که ملاحظه می شود برآوردهای سـه سـتون آخر برای هر پارامتر تفـاوت معنـی داری بایکـدیگر و بـا مقادیر ستونهای دوم وسوم که مربوط به داده هـای کامـل اسـت نـدارد وایـن بـدین معنـی اسـت کـه اگرگمشـدن داده های یک متغیر بـه صـورت کـاملا تصـادفی رخ داده باشد، آنالیز روی حجـم نمونـه ای کـه درآن آزمودنیهـای دارای مقـادیر گمشـده، حـذف شـده باشـند (در اینجـا n = ۱٤۰۰) با آنالیز روی حجم نمونه کامل(n =۲۰۰۰) یکسـان مـی باشـد. بعبـارت دیگـر در مکـانیزم گمشدگی MCAR کافی است افراد دارای داده گمشـده را حـذف کـرده وتجزیـه و تحلیـل براسـاس داده هـای باقیمانده انجام پذیرد. دراین صورت برآوردهـای حاصـل اریب نمی باشـند و کـاراً هسـتند (Carlin B.P. and Louis T.A. ۲۰۰۰).

نکتـه حـائز اهمیـت در ایـن جـدول ایـن اسـت کـه برآوردهای بیزی ستون های سوم وششم ازلحـاظ عـددی نسبت به برآوردهای مشابه واریانس کمتری دارند.

درمرحله بعد مجددا نمونه ۲۰۰۰ تایی بکاربرده شد و این بار مقادیری از متغیر x بااستفاده از مکانیزم گمشدن تصادفی MAR ، بادرنظرگرفتن مقادیر اختیاری ولی متفاوت، از احتمال گمشدن مقادیر x به ازای سطوح متغیر های d وz بصورت زیر ، حذف گردیدند.

$P(d=۰, z=۰)=۰،۵۰$, $p(d=۰, z=۱)=۰،۱۰$,
$p(d=۱, z=۰)=۰،۳۰$, $p(d=۱, z=۱)=۰،۲۰$

با احتساب این احتمالات درصد کل گمشدگی درمتغیر x تقریبا برابر ۰/۲۷ می باشد.

با ٤۰ بار تکرار این عمل به صورت مستقل از هم ، تعداد ٤۰ فایل داده ایجاد گردید که توسط هرسه نرم افزار مذکور، کلیه مراحل تجزیه وتحلیل فوق ، روی آنها انجام پذیرفت که خلاصه نتایج بامحاسبه میانگین و خطای معیار از روی این تکرار ها به عنوان مقادیر برآورد شده پارامتر ها درجدول ۲ درج گردیده است. ترتیب ستونها در جدول ۲ مشابه جدول ۱ است.

ستون چهارم این جدول مربوط به حالتی است که کلیه اطلاعات آزمودنیهای دارای مقدار گمشده درمتغیر x آن ، حذف شده و آنالیز روی داده های کامل باحجم نمونه کمتر (دراینجا ۱٤٤۲فرد) انجام پذیرفته است. با مقایسه مقادیر این ستون با ستون دوم ملاحظه می شود که مقادیر پارامترهای $\beta_۱$ و $\beta_۲$ تفاوت چشمگیری بامقدار واقعی خود دارند واریبی ایجاد شده می تواند به شدت تفسیر نتایج را تحت تاثیر قرار دهد. جهت اصلاح این اریبی تابع درستنمایی ساتن و کارول مورد استفاده قرار گرفت و مقادیر ستونهای ۵ و ٦ جدول محاسبه شدند. مشاهده می شود که درمقایسه ستون پنجم با ستون دوم تفاوت بین برآوردها هنوز قابل ملاحظه بوده وشاید علت آن حساسیت زیاد تابع درستنمایی به مکانیزم گمشدگی باشد. اما با مقایسه ستون ششم با ستون سوم و با ستون دوم مشاهده می شود، برآوردها به واقعیت بسیار نزدیکتر و قابل اتکا می باشند. نکته مهم این است که واریانس برآوردها نیز کوچکتر از واریانس های ستون سوم می باشند.

بنابراین درصورتی که باداده های گمشده با مکانیزم MAR مواجه باشیم با استفاده از تکنیک MCMC روی توزیع پسینی که از ترکیب تابع درستنمایی تعمیم یافته ساتن وکارول و پیشین نا آگاهی بخش حاصل می شود، می توان به برآورد های ، نااریب ونزدیک به

برآورد های داده های کامل نسبت به حالت حذف آزمودنیها از مطالعه ، دست یافت.

مثال :مطالعه سلامت و بیماری درایران ، شهرتهران – داده مثالی این مطالعه مربوط به طرح ملی سلامت و بیماری درایران (نوربالا، احمدعلی و محمد، کاظم ۱۳۸۰) است که درسال ۱۳۸۰ در کل کشور به اجرا گذاشته شد.

به منظور بررسی مشکلات تنفسی ، متغیر سمع ریه از این مطالعه انتخاب گردید که در آن سمع ریه به وسیله گوشی طبی ریتینگ تشخیص داده شده و ثبت اطلاعات براساس سه حالت زیر انجام پذیرفته بود، ۱- رال: صداهایی موزیکال منقطع ریه مانند صدای راه رفتن دربرف، ۲- ویزینگ : صدای موزیکال مداوم که معمولا نشان دهنده انسداد راههای تنفسی می باشد، ۳- طبیعی: درصورت عدم سمع هریک از صداهای فوق سمع ریه طبیعی تلقی شد (نوربالا و محمد ۱۳۸۰).

از این مطالعه اطلاعات مربوط به افراد ۱۵ سال به بالای شهر تهران با حجم نمونه $n = ٦٢٣٨$ ، و با انتخاب متغیر سمع ریه (d) به عنوان متغیر وابسته ومتغیرهای سیگارکشیدن (X) و جنس (Z) به عنوان متغیر های کمکی، فایل داده ها جهت انجام آنالیز تهیه گردید.

متغیر سمع ریه به دو وضعیت ۰= طبیعی و ۱= غیرطبیعی ، متغیر گسسته جنس بصورت ۰=مرد و ۱=زن و متغیر سیگار کشیدن بصورت ۰= غیرسیگاری و ۱= سیگاری کدگذاری شدند. درکل نمونه ۲۱٦ نفر (۳/۵٪) سمع ریه غیر طبیعی داشتند . همچنین ۷٦۸ نفر (۱۲/۳٪) سیگاری و ۳۵٤۹ نفر (۵٦/۹٪) زن بودندو درهیچیک ازمتغیرها داده گمشده نداشتیم. بعلاوه خطر سمع ریه غیر طبیعی از زن به مرد به میزان ۲/۳٦ برابر و ازغیر سیگاری به سیگاری ٦/۱۹ برابر بوده و کلیه ضرائب مدل رگرسیون لجستیک برای داده های کامل معنی دار بودند $(p= ۰/۰۰۰)$.

برای ایجاد فایلی با مقادیر گمشده درمتغیر سیگار کشیدن با هدف استفاده از مکانیزم گمشدن تصادفی MAR درطبقات متغیرهای سمع ریه و جنس به ترتیب ازطبقه $(d=۰, z= ۰)$ ۳۰٪ ، از طبقه $(d=۰, z= ۱)$ ٦٪ ، از طبقه $(d=۱, z= ۰)$ ۱۸٪ و ازطبقه $(d=۱, z= ۱)$ به میزان ۱۲٪ داده ها به تصادف حذف شدند و به این ترتیب درمجموع ، درحدود ۲۰٪ ازکل داده ها به طور تصادفی(MAR) ازمتغیر سیگار کشیدن گمشده فرض شدند؛ سپس با پنج روش مطرح شده دربخش قبل ، به برآورد پارامترهای مدل رگرسیون لجستیک اقدام شد که نتایج آن درجدول ۳ آمده است. ترتیب ستونها در جدول ۲ مشابه جدول ۱ است.

دراین جدول نیز ستونهای دوم و سوم همانند جداول ۱ و ۲ مربوط به داده های کامل وستون چهارم تا ششم مربوط به داده های با مقادیر گمشده است. همانگونه که ملاحظه می شود کلیه برآوردهای بیزی (SCMCMC) ستون ششم که با بکارگیری درستنمایی ساتن و کارول مربوط به داده های گمشده بدست آمده اند ، نسبت به سایر برآوردها به برآوردهای داده های کامل نزدیک ترند و دقت مشابهی دارند.

برای توضیح بیشتر، به عنوان مثال مقادیر برآورد شده برای پارامتر $\beta_۲$ مورد بررسی قرار می گیرند. درصورتی که نمونه کامل ٦۲۳۸ تایی بدون گمشدگی بکاربرده شود، با استفاده ازمدل رگرسیون لجستیک مقدار برآورد پارامتر متغیر جنس به روش ماکزیمم درستنمایی درستون دو (و نیز ستون سه به روش بیزی) برابر ۰/۸٦ می گردد، بنابراین درسال ۱۳۸۰ مشکل سمع ریه غیرطبیعی درمردان تهرانی ۱۵ سال به بالا ۲/۳٦ برابر زنان همین گروه سنی است (OR= ۲/۳٦). حال اگر درحدود ۲۰٪ گمشدگی MAR درمتغیر سیگار کشیدن وجود داشته باشد و آزمودنیهای با مقادیر گمشده از مطالعه حذف گردند، حجم نمونه به ۵۰۱۳ نفر خواهد رسید. برآورد ماکزیمم درستنمایی پارامتر $\beta_۲$ به ازای این حجم از داده های کامل درستون چهارم جدول ۳ برابر ۰/۷۰ می باشد؛ به این معنا که مقدار نسبت بخت (Odds Ratio) از ۲/۳٦ به ۲/۰۱ تغییر می کند. این

اتفاقی است که ممکن است درعمل رخ دهد ومنجربه نتیجه گیری‌های غلط گردد. اما بااستفاده از درستنمایی ساتن وکارول به روش بی‌زی در ستون ششم جدول ۳ ملاحظه می شودکه مقدار این پارامتر برابر ۰/۸۳ می گردد وازآنجا نسبت بخت برابر ۲/۲۹ محاسبه می شود و لذا با این روش مقدار اریبی برآورد به مقدار قابل توجهی تصحیح می گردد.

بحث :

همانگونه که در بخش های قبل ملاحظه شد مسئله گمشدگی داده ها، بامکانیزم MAR ، درمطالعات مقطعی ازاهمیت ویژه ای برخوردار است؛ بطوریکه حذف آزمودنیهایی که دارای مقادیرگمشده اند منجر به برآوردهای اریب از پارامترهای مدل شده ونهایتا نتیجه گیری های دور از واقع حاصل می گردد.

برآوردهای حاصل از مدل ارائه شده دراین مطالعه وآنالیز آن به روش بیزی نسبت به سایر روشهای برآورد، مقادیر نزدیکتری به برآوردهای مدل استاندارد برای داده های کامل داشتند، که این مسئله توسط سینها و همکاران (۲۰۰۳) نیزدرمورد مطالعات مورد- شاهدی نشان داده شده است. ساتن و کارول (۲۰۰۰) نیز به روش کلاسیک نشان دادند که برآوردهای مدل رگرسیون لجستیک در مطالعات مورد- شاهدی با بکارگیری تابع درستنمایی تعمیم یافته بهبود می یابد.

نتیجه گیری:

بنابراین درصورت مواجه شدن با داده های گمشده اولین قدم بازنگری ومشاهده مجدد واحدهای موردمطالعه وتکمیل مقادیر گمشده است. درمرحله بعد باید بامحاسبه احتمالات گمشدگی درسطوح مختلف متغیرها، نسبت به تشخیص گمشدگی تصادفیMAR اطمینان حاصل نمود. درصورت مثبت بودن پاسخ ، استفاده از مدلهای تعمیم یافته مطروحه دراین مقاله به لحاظ اجتناب از نتیجه گیری های نادرست، توصیه می گردد. هرچندکه درحال حاضر هیچیک از نرم افزارهای آماری موجود قادر به تحلیل چنین مدلهایی نیستند ولازم است که برنامه کامپیوتری ویژ ه ای تهیه شود.

جدول ۱- مقادیر برآوردهای پارامترها به روشهای FMLE , FMCMC برای داده های کامل وبدون گمشدگی و مقادیر برآوردهای پارامترها به روشهای CCMLE و SCMLE و SCMCMC با حذف ۳۰٪ گمشدگی کاملا تصادفی (MCAR) درمتغیرکمکی X بدست آمده اند. (اعدادداخل پرانتزها خطای معیار برآوردها می باشند)

برآورد پارامتر ها با ۳۰٪ گمشدن کاملا تصادفی (MCAR) درمتغیر X					برآورد پارامترها برای داده های کامل
پارامتر	SCMCMC WinBUGS	SCMLE R	CCMLE S-Plus	FMCMC WinBUGS	FMLE* S-Plus
β_0	-۰/۳۰۸۱ (۰/۰۲۷۰)	-۰/۳۰۴۰ (۰/۰۳۰۷)	-۰/۲۹۲۶ (۰/۰۶۴۶)	-۰/۳۰۶۸ (۰/۰۷۵۹)	-۰/۳۰۵۴ (۰/۰۷۸۸)
β_1	۰/۹۱۸۰ (۰/۰۷۹۸)	۰/۹۰۵۹ (۰/۰۸۰۷)	۰/۹۰۱۰ (۰/۰۷۷۵)	۰/۹۲۳۸ (۰/۱۳۶۶)	۰/۹۱۱۵ (۰/۱۳۸۱)
β_2	۰/۷۴۶۸ (۰/۰۳۶۹)	۰/۷۳۹۶ (۰/۰۴۳۹)	۰/۷۳۳۰ (۰/۰۸۰۴)	۰/۷۳۸۷ (۰/۱۹۳۶)	۰/۷۳۶۲ (۰/۱۱۲۰)
β_{12}	-۱/۲۴۷۱ (۰/۱۱۱۷)	-۱/۲۳۵۲ (۰/۱۱۹۰)	-۱/۲۲۶۸ (۰/۱۱۲۱)	-۱/۲۳۹۰ (۰/۱۰۸۱)	-۱/۲۲۴۵ (۰/۱۹۲۹)

*FMLE (Full data Maximum Likelihood Estimation)
FMCMC (Full data Marcov Chain Monte Carlo estimation)
CCMLE (Complete Case Maximum Likelihood Estimation)
SCMLE (Satten and Carol Maximum Likelihood Estimation)
SCMCMC (Satten and Carol Marcov Chain Monte Carlo estimation)

جدول ۲- مقادیر برآوردهای پارامترها به روشهای FMLE , FMCMC برای داده های کامل وبدون گمشدگی و مقادیر برآوردهای پارامترها به روشهای CCMLE و SCMLE و SCMCMC با حذف حدود ۲۷٪ گمشدگی تصادفی (MAR) درمتغیرکمکی X بدست آمده اند. (اعدادداخل پرانتزها خطای معیار برآوردها می باشند)

برآورد پارامتر ها با ۲۷٪ گمشدن تصادفی (MAR) درمتغیر X					برآورد پارامترها برای داده های کامل
پارامتر	SCMCMC WinBUGS	SCMLE R	CCMLE S-Plus	FMCMC WinBUGS	FMLE* S-Plus
β_0	-۰/۳۱۵۵ (۰/۰۳۱۷)	-۰/۵۰۲۳ (۰/۰۳۵۸)	۰/۰۲۹۴ (۰/۰۳۰۸)	-۰/۳۰۶۸ (۰/۰۷۵۹)	-۰/۳۰۵۴ (۰/۰۷۸۸)
β_1	۰/۹۳۹۲ (۰/۱۰۱۱)	۱/۴۴۶۷ (۰/۱۰۱۶)	۰/۹۱۶۳ (۰/۱۰۰۷)	۰/۹۲۳۸ (۰/۱۳۶۶)	۰/۹۱۱۵ (۰/۱۳۸۱)
β_2	۰/۷۴۷۴ (۰/۰۴۰۶)	۰/۹۹۶۱ (۰/۰۴۴۰)	۰/۲۸۳۱ (۰/۰۳۹۰)	۰/۷۳۸۷ (۰/۱۹۳۶)	۰/۷۳۶۲ (۰/۱۱۲۰)
β_{12}	-۱/۲۴۸۸ (۰/۱۲۰۴)	-۱/۹۳۵۷ (۰/۱۲۰۰)	-۱/۲۲۲۲ (۰/۱۱۶۷)	-۱/۲۳۹۰ (۰/۱۰۸۱)	-۱/۲۲۴۵ (۰/۱۹۲۹)

*FMLE (Full data Maximum Likelihood Estimation)
FMCMC (Full data Marcov Chain Monte Carlo estimation)
CCMLE (Complete Case Maximum Likelihood Estimation)
SCMLE (Satten and Carol Maximum Likelihood Estimation)
SCMCMC (Satten and Carol Marcov Chain Monte Carlo estimation)

جدول ۳- برآورد پارامترهای مدل لجستیک داده های سمع ریه بر حسب سیگار کشیدن و جنس به روشهای FMCMC , FMLE برای داده های کامل و بدون گمشدگی و روشهای CCMLE و SCMLE و SCMCMC یا حذف حدود ۲۰٪ گمشدگی تصادفی (MAR) درمتغیرکمکی X . (اعدادداخل پرانتزها خطای معیار برآوردها می باشند)

	برآورد پارامتر ها با ۲۰٪ گمشدن تصادفی (MAR) درمتغیر X				برآورد پارامترها برای داده های کامل
پارامتر	FMLE S-Plus	FMCMC WinBUGS	CCMLE S-Plus	SCMLE R	SCMCMC WinBUGS
β_0 عرض ازمبدا	-٤/٠٧٠٠ (٠/١٣٠٦)	-٤/٠٧٩ (٠/١٣١١)	-٣/٩٢٥٤ (٠/١٤٦٠)	-٤/١١٣٨ (٠/١٩٤٠)	-٤/٠٧٣٠ (٠/١٣٧٩)
β_1 سیگارکشیدن (X)	١/٨٢٤٤ (٠/٣٧٤١)	١/٧٨٩٠ (٠/٣٦٢٤)	١/٩٧٩٥ (٠/٤٠٥٢)	٢/١٦٧٩ (٠/٢١٢٠)	١/٦٨٩٠ (٠/٤١٤٢)
β_2 جنس (Z)	٠/٨٥٧٦ (٠/١٧٤٣)	٠/٨٦٠٨ (٠/١٧٢١)	٠/٦٩٧١ (٠/١٨٩١)	٠/٩٩٠٥ (٠/٢٤٤٥)	٠/٨٣٢٩ (٠/١٩١٩)
β_{12} X*Z	-٠/٧٥١٤ (٠/٤١١١)	-٠/٧١٥٦ (٠/٣٩٧٥)	-١/٠٠٤٢ (٠/٤٤٣٩)	-١/٢٩٧٧ (٠/٢٧٥٥)	-٠/٥٧٤٩ (٠/٤٥٠٢)

*FMLE (Full data Maximum Likelihood Estimation)
FMCMC (Full data Marcov Chain Monte Carlo estimation)
CCMLE (Complete Case Maximum Likelihood Estimation)
SCMLE (Satten and Carol Maximum Likelihood Estimation)
SCMCMC (Satten and Carol Marcov Chain Monte Carlo estimation)

منابع :

خیری، سلیمان. فقیه زاده، سقراط. مشکانی، محمدرضا. محمودی، محمود. (۱۳۸۲) تحلیل مدلهای شکنندگی همبسته به روش بیزی. رساله دکتری آمارزیستی، دانشگاه تربیت مدرس، دانشکده پزشکی.

نوربالا، احمدعلی. محمد، کاظم.(۱۳۸۰) بررسی سلامت و بیماری در ایران، انتشارات مرکز ملی تحقیقات علوم پزشکی کشور.

Ashby D., Hutton J.L. and McGee M.A. (۱۹۹۳) Simple Bayesian analysis for case-controlled studies in cancer epidemiology. *Statistician,* ٤٢:٣٨٥-٣٨٩.

Blackhurst D.W. and Schluchter M.D. (۱۹۸۹) Logistic regression with a partially observed covariate. *Comm. Statist. Simul.* ١٨(١):١٦٣-١٧٧.

Carlin B.P. and Louis T.A. (٢٠٠٠) Bayes and Empirical Bayes Methods for Data Analysis. Second edition, Chapman and Hall.

Fleiss J.L., Levin B. and Paik M.C. (٢٠٠٣) Statistical Methods for Rates and Proportions. Third Edition, John Wiley and Sons.

Fuchs C. (١٩٨٢) Maximum likelihood estimation and model selection in contingency tables with missing data. *J. Amer. Statist. Assoc.* ٧٧: ٢٧٠- ٢٧٨.

Gilks W.R., Richardson S. and Spiegelhalter D.J. (١٩٩٧) Markov Chain

Satten G.A. and Carroll R.J. (۲۰۰۰) Conditional and unconditional categorical regression models with missing covariates. *Biometrics.* ٥٦: ۳۸٤-۳۸۸.

Satten G.A. and Kupper L. (۱۹۹۳a) Inferences about exposure – disease associations using probability of exposure information. *J. Amer. Statist. Assoc.* ۸۸: ۲۰۰-۲۰۸.

Satten G.A. and Kupper L. (۱۹۹۳b) Conditional regression analysis of the exposure–disease odds ratio using known probability – of –exposure values, *Biometrics.* ٤٤: ٤۲۹ – ٤٤۰.

Seaman S.R. and Richardson S. (۲۰۰۱) Bayesian Analysis of case-control studies with categorical covariates. *Biometricka.* ۸۸:۱۰۷۳-۱۰۸۸.

Sinha S., Mukherjee B. and Ghosh M. (۲۰۰٤) Bayesian Semiparametric Modeling for Matched Case-Control Studies with Multiple Disease States. *Biometrics.* ٦۰: ٤۱-٤۹.

Spiegelhalter D., Thomas A., Best N. and Lunn D. (۲۰۰۳) WinBUGS ۱.٤ Manual

Zelen M. and Parker R.A. (۱۹۸٦) Case-Control Studies and Bayesian inference. *Statistics in Medicine.* ٥:۲٦۱-۲٦۹.

Monte Carlo in Practice. Chapman and Hall.

Little R.J.A. and Rubin, D.B. (۲۰۰۲) Statistical analysis with missing data. Second Edition, John Wiley & Sons, New York.

Little R.J.A. and Schluchter M.D. (۱۹۸٥) Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika.* ۷۲: ٤۹۷- ٥۱۲.

Muller P., Parmigiani G., Schildkraut J. and Tardella L. (۱۹۹۹) A Bayesian Hierarchical Approach for combining case-control and prospective studies. *Biometrics.* ٥٥:۸٥۸-۸٦٦.

Muller P. and Roeder K. (۱۹۹۷) A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika.* ۸٤:٥۲۳-٥۳۷.

Nurminen M. and Mutanen P. (۱۹۸۷) Exact Bayesian analysis of two proportions. *Scandinavian Journal of Statistics.* ۱٤:٦۷-۷۷.

Paik M.C. and Sacco R.L. (۲۰۰۰) Matched Case–Control data analyses with missing covariates. *Applied Statistics.* ٤۹: ۱٤٦-۱٥٦.

Rathouz P.J., Satten G.A. and Carrol R.J. (۲۰۰۳) Semiparametric inference in matched case – control studies with missing covariate data. *Biometrika.* In press.