

مقایسه روش بیزی (Bayesian) و کلاسیک در برآورد پارامترهای مدل رگرسیون لجستیک با وجود مقادیر گمشده در متغیرهای 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
دکتر کاظم محمد: استاد گروه آمار زیستی و اپیدمیولوژی دانشکده بهداشت و انستیتو تحقیقات بهداشتی دانشگاه علوم پزشکی تهران
دکتر محمدرضا مشکانی: استاد گروه آمار دانشکده علوم ریاضی دانشگاه شهید بهشتی
غلامرضا جندقی: عضو هیات علمی دانشگاه تهران، پردیس قم
دکتر کرامت... نوری: استادیار دانشکده بهداشت و انستیتو تحقیقات بهداشتی دانشگاه علوم پزشکی تهران
دکتر عین... پاشا: دانشیار دانشکده ریاضی دانشگاه تربیت معلم
دکتر کمال 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 بطور وسیعی در تحقیقات پزشکی و اپیدمیولوژیک کاربرد دارد. در بسیاری از مطالعات با مجموعه داده هایی مواجه می شویم که بخشی از آنها گزارش نشده اند یا به عبارت دیگر گمشده می باشند. ساده ترین روش برای تجزیه و تحلیل چنین داده هایی صرف نظر کردن از مواردی دارای مقادیر گمشده و ادامه آنالیز با داده های کامل می باشد که این روش، در عمل کارآمد نیست.

روش کار: در این مطالعه روشی برای آنالیز مدل های رگرسیون لجستیک وقتی که مقادیر متغیر کمکی Z بطور کامل مشاهده شده و مقادیر 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 و بدون گمشدگی اند، عمل می کند.

نتایج: با بکاربردن این تابع درستنمایی، برآورد پارامترها هم به روش کلاسیک ماکزیمم درستنمایی و هم به روش بیزی با استفاده از تکنیک زنجیره های مارکوف مونت کارلویی (MCMC) انجام و مقادیر بدست آمده مقایسه گردیده اند. داده های این مطالعه مربوط به اطلاعات سمع ریه بدست آمده از طرح سلامت و بیماری در شهر تهران می باشد.

نتیجه گیری: برآوردهای حاصل از مدل ارائه شده در این مطالعه و آنالیز 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:

تحقیقات اجتماعی داشته اند. بدلیل غیر خطی بودن مدل رگرسیون لجستیک، تحلیل آن از طریق ماکزیمم درستنمایی صورت می پذیرد؛ چرا که در آن هیچگونه

دردهای اخیر مدل های رگرسیون لجستیک نقش مهمی را در تحلیل داده های پزشکی و اپیدمیولوژیک و

مشکل مواجه سازد. گمشدگی می تواند در متغیرهای پاسخ یا متغیرهای کمکی رخ دهد. در این مطالعه، گمشدن درمقادیر متغیرهای کمکی موردنظر بوده و مکانیزم آن از نوع گمشدگی تصادفی (MAR) می باشد.

در نحوه برخورد با داده های گمشده عمدتاً 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 عمدتاً 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 نمایم. این روش در صورت بالا بودن تعداد آزمودنیهای گمشده، دارای دو اشکال عمده می باشد: اولاً شکل طبیعی توزیع متغیر دارای مقادیر گمشده تغییر می یابد و ثانیاً، 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 در تابع درستنمایی می باشد.

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

محدودیتی برای متغیرهای مستقل درنظر گرفته نمی شود. در مطالعات کوهورت یا مطالعه مقطعی که در آنها جورکردن ملاک عمل نیست از روش ماکزیم درستنمایی غیرشرطی اقدام به برآورد پارامترها می گردد (Kleinbaum D.G., Klein M. ۲۰۰۲). به علاوه در اینگونه مطالعات با مواردی مواجه می شویم که در آنها بخشی از داده ها بدلیل 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 کاملاً تصادفی (MCAR Missing Completely At Random)، گمشدگی تصادفی غیرقابل اغماض (Missing At Random) NI (Non Ignorable) و گمشدگی (Little R.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 از مطالعه حذف و آنالیز را براساس حجم نمونه جدید اجرا نمود بدون اینکه دربرآورد ها اریبی ایجاد شود.

درمکانیزم گمشدگی MAR، گمشدگی در متغیر کمکی (Covariate) بستگی به خود متغیر نداشته بلکه به متغیرهای دیگر بستگی 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 سیگارها، غیر سیگارها بدلیل آگاهی و علاقمندی نسبت به وضعیت سلامت جسمی خود، بیشتر به شرکت در مطالعه 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 نهایتاً دستیابی به یک نتیجه گیری مفید از داده های جمع آوری شده را با

کمکی پیوسته دارای مقادیر گمشده از توزیع 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 و کوپر (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 و اعمال 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؛ از جمله: زلن و پارکر (Zelen M. and Parker R.A. ۱۹۸۶)، نورمینن و موتانن (Nurminen M. and Mutanen P. ۱۹۸۷) و اشبی و همکاران (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 و احتمالاً دارای مقادیر گمشده مورد بررسی قرار دادند.

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

در روش بیزی (Bayesian) استنباط درباره پارامترهای مدل بر مبنای توزیع پسین (Posteriordistribution) آنها صورت می پذیرد که تلفیقی است از داده های مشاهده شده و اطلاعات ناشی از مطالعات قبلی و یا تجارب شخصی که با عنوان توزیع پیشین (Prior distribution) شناخته می شود در صورت نامعلوم بودن توزیع پسین، می توان بابهیره گیری از روشهای شبیه سازی، زنجیره های مارکوف مونت کارلویی MCMC (Markov Chain Monte Carlo) برای هر حجم نمونه دلخواه، استنباط دقیقی از پارامترها بدست آورد.

در مطالعه حاضر، از روش سوم و تعیین 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 و MCMC 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 عموماً درارتباط با مطالعات مورد-شاهدی می باشد. فاکس (Fuchs ۱۹۸۲) و لیتل و شلاختر (Little J.A. ۱۹۸۵ 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 عموماً به تکرار نیاز داشته و گاهی 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

متغیر کمکی X احتمالاً در برخی مشاهدات دارای مقادیر گمشده است. بدون اینکه از کلیت مساله کاسته شود، مدل‌های خود را بر اساس یک متغیر Z و یک متغیر X ارائه داده، در ابتدا فرض می‌کنیم که هر دو متغیر با مشاهدات کامل و بدون گمشدگی باشند. در این حالت احتمالات شرطی متغیربیماری به شرط متغیرهای کمکی، با استفاده از مدل لجستیک به صورت روابط ۱ الی ۴ تعریف می‌شوند، که در آن x' و z' سطوحی از متغیر X و متفاوت از سطوح x و z هستند. هدف از تحلیل رگرسیون لجستیک بدست آوردن برآوردی از پارامترهای مدل (در اینجا: $\beta_0, \beta_1, \beta_2, \beta_{12}$) برای تبیین ارتباط بین متغیر پاسخ Y و مجموعه‌ای از متغیرهای کمکی X و Z می‌باشد. در صورت کامل بودن مقادیر متغیرهای X و Z از روشهای استاندارد جهت برآورد این پارامترها استفاده می‌شود. اما اگر فرض کنیم که برخی از مشاهدات متغیر X گمشده باشند در این صورت داریم:

$$\tilde{\theta}(Z) = \frac{P(Y=1 | Z=z)}{P(Y=0 | Z=z)} \quad (5)$$

بعلاوه می‌توان تعریف کرد:

$$\pi(x | z) = P(X=x | Y=0, Z=z) \quad (6)$$

$$\rho(x | z) = P(X=x | Y=1, Z=z) \quad (7)$$

$$P(Y=1 | X=x, Z=z) = \frac{e^{\beta_0 + \beta_1 x + \beta_2 z + \beta_{12} xz}}{1 + e^{\beta_0 + \beta_1 x + \beta_2 z + \beta_{12} xz}} \quad (1)$$

$$P(Y=0 | X=x, Z=z) = \frac{1}{1 + e^{\beta_0 + \beta_1 x + \beta_2 z + \beta_{12} xz}} \quad (2)$$

بنابراین با تعاریف فوق بخت (Odds) و نسبت بخت بیماری (Odds Ratio) به ترتیب عبارت خواهند بود از:

$$\theta(x, z) = \frac{P(Y=1 | X=x, Z=z)}{P(Y=0 | X=x, Z=z)} = e^{\beta_0 + \beta_1 x + \beta_2 z + \beta_{12} xz} \quad (3)$$

$$\psi(x, z, x', z') = \frac{\theta(x, z)}{\theta(x', z')} \quad (4)$$

سی من و ریچارد سون (Seaman S.R. and Richardson S. ۲۰۰۱) روش مولر و روی در را برای متغیرهای کمکی گسسته گسترش دادند و ارتباطی بین روشهای مولر و رویدر وزن و پارکر هنگامی که خطای اندازه گیری وجود ندارد، برقرار کردند. سینها و همکاران (Sinha S. et al. ۲۰۰۴) 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 تعمیم دادند و با روش MCMC، استنباط در مورد پارامترهای مدل را به انجام رساندند.

هیچ یک از مطالعات انجام شده، تحلیل مدل رگرسیون لجستیک غیر شرطی مربوط به مطالعات مقطعی را وقتی داده‌های گمشده در متغیرهای کمکی وجود داشته باشند، به روش کلاسیک و بیزی را مورد مقایسه قرار نداده‌اند.

روش کار:

مدل و تعاریف: فرض کنید متغیر Y_i نشان دهنده متغیر پاسخ دوقضیعی متناظر با فرد i ام باشد، بطوریکه $Y_i=1$ به معنی بیمار بودن فرد i ام و $Y_i=0$ به معنی سالم بودن فرد i ام باشد. به علاوه، فرض می‌کنیم بردار متغیر کمکی Z دارای مقادیر بطور کامل مشاهده شده و بردار

در صورتی که متغیر کمکی X دارای مقادیر گمشده باشد، متغیر نشانگر Δ_i را به این صورت تعریف می‌کنیم: $\Delta_i=1$ اگر X_i مشاهده باشد و $\Delta_i=0$ اگر X_i مشاهده نشده باشد. باین تعریف، می‌توان تابع درستنمایی داده‌های کامل را بصورت زیر بازنویسی کرد:

$$P(Y, X, \Delta | Z) = P(Y | Z) P(\Delta | Y, Z) P(X | Y, Z, \Delta) \quad (11)$$

تحت شرط گمشدن تصادفی MAR می‌توان فرض کرد که، $P(X | Y, Z, \Delta) = P(X | Y, Z)$ (Little and R.J.A. and Rubin D.B. ۲۰۰۲). همچنین فرض می‌کنیم که احتمال گمشدگی $P(\Delta | Y, Z)$ بستگی به بردار پارامترها، β ندارد. [۹] بنابراین با حذف این عبارت از درستنمایی فوق، درستنمایی غیر شرطی اصلاح شده برای داده‌های مشاهده شده عبارت خواهد بود از:

$$L(\beta) = \prod_{i=1}^n \tilde{\theta}(Z_i)^{Y_i} [1 + \tilde{\theta}(Z_i)]^{-1} \times \pi(X_i | Z_i)^{\Delta_i(1-Y_i)} \rho_1(X_i | Z_i)^{\Delta_i Y_i} \quad (12)$$

در رابطه فوق توزیع احتمال $\pi(x|z)$ نامعلوم می‌باشد، در صورتیکه X و Z مقادیر شمارا و محدودی را اختیار کنند توزیع مناسبی که می‌توان برای $\pi(\cdot)$ مفروض دانست عبارتست از (Satten G.A. and Carol R.J. ۲۰۰۰):

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

$$\tilde{\theta}(z) = \sum_x \theta(x, z) \cdot \pi(x | z) \quad (8)$$

که در آن، مجموع روی همه مقادیر ممکن متغیر X صورت می‌پذیرد. دومین نتیجه مقاله مذکور عبارتست از:

$$\rho(x | z) = \frac{\pi(x | z) \theta(x, z)}{\sum_x \pi(x | z) \theta(x, z)} \quad (9)$$

اگر X متغیر پیوسته باشد در این صورت به جای مجموعها باید از انتگرالها استفاده شود و عبارتهای $\pi(x | z)$ و $\rho(x | z)$ نیز توابع چگالی احتمال در نظر گرفته شوند.

تابع درستنمایی، توزیع پیشین و توزیع پسین: می‌توان نتایج بخش قبل را برای تعریف تابع درستنمایی جهت برآورد پارامترهای مورد نظر بوسیله داده‌های مشاهده شده، بکار برد. تابع درستنمایی رگرسیون لجستیک استاندارد درحالتیکه مقادیر X و Z بطور کامل مشاهده شده باشند، عبارتست از:

$$L(\beta) = \prod_{i=1}^n \frac{[\theta(X_i, Z_i)]^{Y_i}}{1 + \theta(X_i, Z_i)} \quad (10)$$

$$\pi(x | z) = \frac{e^{\gamma_{xz}}}{\sum_{x'} e^{\gamma_{x'z}}} = \frac{e^{\gamma_0 + \gamma_1 x + \gamma_2 z + \gamma_{12} xz}}{\sum_{x'} e^{\gamma_0 + \gamma_1 x' + \gamma_2 z + \gamma_{12} x'z}} = \frac{e^{\gamma_1 x + \gamma_{12} xz}}{\sum_{x'} e^{\gamma_1 x' + \gamma_{12} x'z}} \quad (13)$$

بایکارگیری روابط (۳) تا (۹) در رابطه (۱۳) و بازنویسی مجدد تابع درستنمایی (۱۲) داریم:

$$l(\beta | X, Z) = \prod_{i=1}^n \frac{\left(e^{\beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_{12} x_i z_i} \right)^{\Delta_i y_i}}{1 + \sum_x \left[e^{\beta_0 + (\beta_1 + \gamma_1)x + \beta_2 z_i + (\beta_{12} + \gamma_{12})x z_i} \cdot \frac{1}{\sum_x e^{\gamma_1 x + \gamma_{12} x z_i}} \right]} \quad (14)$$

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

با فرض معلوم بودن مشاهدات (X, Z) می توان به کمک قضیه بیز توزیع β به شرط (X, Z) رابدست آورد (Gilks et al 1۹۹۷):

$$\Pi(X, Z, \beta) = L(\beta | X, Z) \Pi(\beta) \quad (16)$$

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 و تنها توزیع پسین پایه استنباط درباره پارامترهاست.

دراین مطالعه با ضرب نمودن چگالی توزیع نرمال هرپارامتر در درستنمایی (۱۴) شکل توزیع پسین حاصل، نامعلوم بود و عملاً ناچار به استفاده از الگوریتم متروپولیس (Metropolis)، با معرفی توزیع پیشنهادی (Proposal distribution) نرمال شدیم. بنابراین در هر مرحله از شبیه سازی MCMC به ترتیب از توابع چگالی شرطی کامل هر پارامتر $\beta_0, \beta_1, \beta_2, \beta_{12}, \gamma_1, \gamma_2$ ، نمونه گیری به عمل آمد و براساس آن استنباط درباره پارامتر مذکور صورت پذیرفت.

که تابعی از پارامترهای $\beta_0, \beta_1, \beta_2, \beta_{12}, \gamma_1, \gamma_2$ می باشد. با لگاریتم گیری از تابع فوق و مشتق گیری از آن نسبت به تک تک پارامترها و مساوی صفر قرار دادن آنها دستگاه معادلات عددی حاصل می گردد که بدلیل غیر خطی بودن معادلات بایکارگیری روشهای عددی اقدام به محاسبه MLE می نمائیم. برآوردهای حاصل از بکارگیری درستنمایی مذکور بسیار کارآتر از روش استاندارد است که در آن 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 $\beta = (\beta_0, \beta_1, \beta_2, \beta_{12}, \gamma_1, \gamma_2)$ تعیین شود و سپس بطور متوالی از آنها نمونه گیری به عمل آید. دراین مطالعه، توزیع پیشین (Prior distribution) پارامترهای مدل را ناآگاهی بخش با توزیع $\beta_i \sim N(\mu = 0, \sigma^2 = 10^6)$ در نظر گرفته و توزیع احتمال مشاهدات را تابع درستنمایی (۱۴)، $L(\beta | X, Z)$ قرار دادیم. برای استنباط بیزی لازم است توزیع توام مشاهدات و پارامترها را بصورت زیر محاسبه نمائیم.

$$\Pi(X, Z, \beta) = L(\beta | X, Z) \Pi(\beta) \quad (15)$$

نتایج:

مقایسه برآورد پارامترها به روش MCMC و MLE: در بخش قبل، متدولوژی جامعی از دو روش بیزی MCMC 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. در این بخش بازائه مثالی ساختگی، بحث را آغاز می کنیم و دربخش بعد با یک نمونه واقعی از طرح سلامت و بیماری مطلب را ادامه خواهیم داد. جدول زیر راکه در آن X و Z متغیرهای کمکی و d متغیر وابسته یا پاسخ است در نظر بگیرد:

x	z	d	n
۰	۰	۲۸۰	۶۶۰
۰	۱	۴۰۰	۶۶۰
۱	۰	۲۲۰	۳۴۰
۱	۱	۱۸۰	۳۴۰

به عنوان مثال سطر اول این جدول نشان دهنده این است که از کل ۶۶۰ نفر موجود در طبقه $X=0$ ، $Z=0$ تعداد ۲۸۰ نفر بیمار می باشند. به این ترتیب، یک نمونه ۲۰۰۰ تایی خواهیم داشت که ازجهاتی با داده های نوعی مطالعه حقیقی مشابه است که در آن توزیع متغیر Z به صورت ۵۰٪-۵۰٪ و توزیع متغیر X به صورت ۶۶٪-۳۴٪ و کلیه مقادیر نسبت بخت (OR) بین ۲ تا ۳ می باشند.

برای آنالیز داده های فوق ازسه بسته نرم افزاری متداول S-Plus ۲۰۰۰، R نسخه ۱/۹ و WinBUGS نسخه ۴/۱ استفاده گردید. WinBUGS نرم افزار توانمندی است که درسالهای اخیر توسط اسپیگل هالتر وهمکاران (Spiegelhalter D. et al. ۲۰۰۳) جهت اجرای استنباط های بیزی باتکنیک

MCMC طراحی و به صورت آزاد در اختیار محققین و بخصوص آمارشناسان بیزی قرار گرفته است. در مجموع با پنج روش پارامترهای $\beta_1, \beta_2, \beta_3, \beta$ با استفاده از مجموعه داده های فوق و نرم افزارهای مذکور بر آورد گردیدند. که نتایج آن درجدول ۱ درج شده است. ضمناً از ذکر مقادیر برآوردشده برای پارامترهای γ_1, γ_2 بدلیل اینکه اصولاً 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 از S-Plus می باشد، که به اختصار باعنوان FMLE نامیده شده است. ($n=2000$).

ستون سه - شامل برآوردهای بیزی با تکنیک MCMC برای داده های کامل بدون گمشدگی است که با نگارش برنامه ای درمحیط WinBUGS و با 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 (Complete Case) CC و آنالیز روی داده های باقی مانده است که با S-Plus محاسبه شده و باعنوان CCMLE نام گذاری شده است. ضمناً مکانیزم گمشدگی درمتغیر X دراین مرحله از نوع گمشدگی کاملاً تصادفی MCAR بود و لذا ۳۰٪ از داده های متغیر X با این روش حذف شدند و آنالیز باحجم نمونه ۱۴۰۰ آنالیز 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 و کارول (Satten G.A. and Carol ۲۰۰۰، S)، رابطه (۱۴)، و بانگارش برنامه ای درمحیط

$$P(d=0, z=0)=0,50, \quad p(d=0, z=1)=0,10,$$

$$p(d=1, z=0)=0,30, \quad p(d=1, z=1)=0,20$$

با احتساب این احتمالات درصد کل گمشدگی درمتغیر X تقریباً برابر ۰/۲۷ می باشد.

با ۴۰ بار تکرار این عمل به صورت مستقل از هم، تعداد ۴۰ فایل داده ایجاد گردید که توسط هر سه نرم افزار مذکور، کلیه مراحل تجزیه و تحلیل فوق، روی آنها انجام پذیرفت که خلاصه نتایج بامحاسبه میانگین و خطای معیار از روی این تکرارها به عنوان مقادیر برآورد شده پارامترها در جدول 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 (در اینجا ۱۴۴۲ فرد) انجام پذیرفته است. با مقایسه مقادیر این ستون با ستون دوم ملاحظه می شود که مقادیر پارامترهای β_1 و β_2 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 ۶ جدول محاسبه شدند. مشاهده می شود که درمقایسه ستون پنجم با ستون دوم تفاوت بین برآوردها هنوز قابل ملاحظه بوده و شاید 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 در صورتی 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 و کارول و پیشین نا آگاهی بخش حاصل می شود، می توان به برآورد های، ناریب و نزدیک به

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 (۱۴) و با نگارش برنامه ای در محیط WinBUGS بدست آمده اند و به اختصار با عنوان SCMMCMC ذکر شده است.

شایان ذکر است که برای بدست آوردن برآوردهای ستونهای ۴ تا ۶ جدول ۱، ابتدا ۳۰٪ از مقادیر متغیر X در فایل ۲۰۰۰ تایی به تعداد ۴۰ بار مستقل از هم، بطور کاملاً تصادفی (MCAR) حذف و سپس میانگین و خطای معیار ۴۰ پارامتر در جدول درج شده است.

همانگونه که ملاحظه می شود برآوردهای سه ستون آخر برای هر پارامتر تفاوت معنی داری بایکدیگر و با مقادیر ستونهای دوم و سوم که مربوط به داده های کامل است ندارد و این بدین معنی است که اگر گمشدن داده های یک متغیر به صورت کاملاً تصادفی رخ داده باشد، آنالیز روی حجم نمونه ای که در آن آزمودنیهای دارای مقادیر گمشده، حذف شده باشند (در اینجا $n = 1400$) با آنالیز روی حجم نمونه کامل ($n = 2000$) یکسان می باشد. بعبارت دیگر در مکانیزم گمشدگی MCAR کافی است افراد دارای داده گمشده را حذف کرده و تجزیه و تحلیل براساس داده های باقیمانده انجام پذیرد. در این صورت برآوردهای حاصل اریب نمی باشند و کارآ هستند (Carlin B.P. and Louis T.A. ۲۰۰۰).

نکته حائز اهمیت در این جدول این است که برآوردهای بیزی ستون های سوم و ششم از لحاظ عددی نسبت به برآوردهای مشابه واریانس کمتری دارند. در مرحله بعد مجدداً نمونه ۲۰۰۰ تایی بکاربرده شد و این بار مقادیری از متغیر X با استفاده از مکانیزم گمشدن تصادفی MAR، بادر نظر گرفتن مقادیر اختیاری ولی متفاوت، از احتمال گمشدن مقادیر X به ازای سطوح متغیر های d و z بصورت زیر، حذف گردیدند.

MAR در طبقات متغیرهای سمع ریه و جنس به ترتیب از طبقه $(d=0, z=0)$ ۳۰٪، از طبقه $(d=1, z=0)$ ۱۸٪ و از طبقه $(d=1, z=1)$ به میزان ۱۲٪ داده ها به تصادف حذف شدند و به این ترتیب در مجموع، در حدود ۲۰٪ از کل داده ها به طور تصادفی (MAR) از متغیر سیگار کشیدن گم شده فرض شدند؛ سپس با پنج روش مطرح شده در بخش قبل، به برآورد پارامترهای مدل رگرسیون لجستیک اقدام شد که نتایج آن در جدول ۳ آمده است. ترتیب ستونها در جدول ۲ مشابه جدول ۱ است.

در این جدول نیز ستونهای دوم و سوم همانند جداول ۱ و ۲ مربوط به داده های کامل و ستون چهارم تا ششم مربوط به داده های با مقادیر گم شده است. همانگونه که ملاحظه می شود کلیه برآوردهای بیزی (SCMCMC) ستون ششم که با بکارگیری درستنمایی ساتن و کارول مربوط به داده های گم شده 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 β_4 مورد بررسی قرار می گیرند. در صورتی که نمونه کامل ۶۲۳۸ تایی بدون گمشدگی بکاربرده شود، با استفاده از مدل رگرسیون لجستیک مقدار برآورد پارامتر متغیر جنس به روش ماکزیمم درستنمایی درستون دو (و نیز ستون سه به روش بیزی) برابر ۰/۸۶ می گردد، بنابراین در سال ۱۳۸۰ مشکل سمع ریه غیرطبیعی در مردان تهرانی ۱۵ سال به بالا ۲/۳۶ برابر زنان همین گروه سنی است ($OR=2/36$). حال اگر در حدود ۲۰٪ گمشدگی MAR در متغیر سیگار کشیدن وجود داشته باشد و آزمودنیهای با مقادیر گم شده از مطالعه حذف گردند، حجم نمونه به ۵۰۱۳ نفر خواهد رسید. برآورد ماکزیمم درستنمایی پارامتر β_4 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 ۲/۳۶ به ۲/۰۱ تغییر می کند. این

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

مثال: مطالعه سلامت و بیماری در ایران، شهر تهران - داده مثالی این مطالعه مربوط به طرح ملی سلامت و بیماری در ایران (نوربالا، احمدعلی و محمد، کاظم ۱۳۸۰) است که در سال 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 معمولاً نشان دهنده انسداد راههای تنفسی می باشد، ۳- طبیعی: در صورت عدم سمع هر یک 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 $n = 6238$ ، و با انتخاب متغیر سمع ریه (d) به عنوان متغیر وابسته و متغیرهای سیگار کشیدن (X) و جنس (Z) به عنوان متغیرهای کمکی، فایل داده ها جهت انجام آنالیز تهیه گردید.

متغیر سمع ریه به دو وضعیت $=0$ طبیعی و $=1$ غیرطبیعی، متغیر گسسته جنس بصورت $=0$ مرد و $=1$ زن و متغیر سیگار کشیدن بصورت $=0$ غیرسیگاری و $=1$ سیگاری کدگذاری شدند. در کل نمونه ۲۱۶ نفر $(3/5)$ سمع ریه غیر طبیعی داشتند. همچنین ۷۶۸ نفر $(12/3)$ سیگاری و ۳۵۴۹ نفر $(56/9)$ زن بودند و در هیچیک از متغیرها داده گم شده نداشتیم. بعلاوه خطر سمع ریه غیر طبیعی از زن به مرد به میزان ۲/۳۶ برابر و از غیر سیگاری به سیگاری ۶/۱۹ برابر بوده و کلیه ضرائب مدل رگرسیون لجستیک برای داده های کامل معنی دار بودند ($p=0/000$).

برای ایجاد فایلی با مقادیر گم شده در متغیر سیگار کشیدن با هدف استفاده از مکانیزم گمشدن تصادفی

کلاسیک نشان دادند که برآوردهای مدل رگرسیون لجستیک در مطالعات مورد-شاهدی با بکارگیری تابع درستنمایی تعمیم یافته بهبود می یابد.

نتیجه گیری:

بنابراین در صورت مواجهه شدن با داده های گمشده اولین قدم بازنگری و مشاهده مجدد واحدهای مورد مطالعه و تکمیل مقادیر گمشده است. در مرحله بعد باید با محاسبه احتمالات گمشدگی در سطوح مختلف متغیرها، نسبت به تشخیص گمشدگی تصادفی MAR اطمینان حاصل نمود. در صورت مثبت بودن پاسخ، استفاده از مدل های تعمیم یافته مطروحه در این مقاله به لحاظ اجتناب از نتیجه گیری های نادرست، توصیه می گردد. هرچند که در حال حاضر هیچیک از نرم افزارهای آماری موجود قادر به تحلیل چنین مدل هایی نیستند و لازم است که برنامه کامپیوتری ویژه ای تهیه شود.

اتفاقی است که ممکن است در عمل رخ دهد و منجر به نتایج گمراهی های غلط گردد. اما با استفاده از درستنمایی ساتن و کارول به روش بی‌زی در ستون ششم جدول ۳ ملاحظه می شود که مقدار این پارامتر برابر $0/83$ می گردد و از آنجا نسبت بخت برابر $2/29$ 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، با مکانیزم MAR، در مطالعات مقطعی از اهمیت ویژه ای برخوردار است؛ بطوریکه حذف آزمودنی هایی که دارای مقادیر گمشده اند منجر به برآوردهای اریب از پارامترهای مدل شده و نهایتاً نتیجه گیری های دور از واقع حاصل می گردد.

برآوردهای حاصل از مدل ارائه شده در این مطالعه و آنالیز آن به روش بی‌زی نسبت به سایر روش های برآورد، مقادیر نزدیکتری به برآوردهای مدل استاندارد برای داده های کامل داشتند، که این مسئله توسط سینها و همکاران (۲۰۰۳) نیز در مورد مطالعات مورد-شاهدی نشان داده شده است. ساتن و کارول (۲۰۰۰) نیز به روش

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 کاملاً تصادفی (MCAR) درمتغیر کمی X بدست آمده اند. (اعدادداخل پرانتزها خطای معیار برآوردها می باشند)

پارامتر	FMLE* S-Plus	FMCMC WinBUGS	CCMLE S-Plus	SCMLE R	SCMCMC WinBUGS
β	-۰/۳۰۵۴ (۰/۰۷۸۸)	-۰/۳۰۶۸ (۰/۰۷۵۹)	-۰/۲۹۲۶ (۰/۰۶۴۶)	-۰/۳۰۴۰ (۰/۰۳۰۷)	-۰/۳۰۸۱ (۰/۰۲۷۰)
β_1	۰/۹۱۱۵ (۰/۱۳۸۱)	۰/۹۲۳۸ (۰/۱۳۶۶)	۰/۹۰۱۰ (۰/۰۷۷۵)	۰/۹۰۵۹ (۰/۰۸۰۷)	۰/۹۱۸۰ (۰/۰۷۹۸)
β_2	۰/۷۳۶۲ (۰/۱۱۲۰)	۰/۷۳۸۷ (۰/۱۹۳۶)	۰/۷۳۳۰ (۰/۰۸۰۴)	۰/۷۳۹۶ (۰/۰۴۳۹)	۰/۷۴۶۸ (۰/۰۳۶۹)
β_{12}	-۱/۲۲۴۵ (۰/۱۹۲۹)	-۱/۲۳۹۰ (۰/۱۰۸۱)	-۱/۲۲۶۸ (۰/۱۱۲۱)	-۱/۲۳۵۲ (۰/۱۱۹۰)	-۱/۲۴۷۱ (۰/۱۱۱۷)

*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 یا حذف حدود ۲۷٪ گمشدگی تصادفی (MAR) درمتغیر کمی X بدست آمده اند. (اعدادداخل پرانتزها خطای معیار برآوردها می باشند)

پارامتر	FMLE* S-Plus	FMCMC WinBUGS	CCMLE S-Plus	SCMLE R	SCMCMC WinBUGS
β	-۰/۳۰۵۴ (۰/۰۷۸۸)	-۰/۳۰۶۸ (۰/۰۷۵۹)	۰/۰۲۹۴ (۰/۰۳۰۸)	-۰/۵۰۲۳ (۰/۰۳۵۸)	-۰/۳۱۵۵ (۰/۰۳۱۷)
β_1	۰/۹۱۱۵ (۰/۱۳۸۱)	۰/۹۲۳۸ (۰/۱۳۶۶)	۰/۹۱۶۳ (۰/۱۰۰۷)	۱/۴۴۶۷ (۰/۱۰۱۶)	۰/۹۳۹۲ (۰/۱۰۱۱)
β_2	۰/۷۳۶۲ (۰/۱۱۲۰)	۰/۷۳۸۷ (۰/۱۹۳۶)	۰/۲۸۳۱ (۰/۰۳۹۰)	۰/۹۹۶۱ (۰/۰۴۴۰)	۰/۷۴۷۴ (۰/۰۴۰۶)
β_{12}	-۱/۲۲۴۵ (۰/۱۹۲۹)	-۱/۲۳۹۰ (۰/۱۰۸۱)	-۱/۲۲۲۲ (۰/۱۱۶۷)	-۱/۹۳۵۷ (۰/۱۲۰۰)	-۱/۲۴۸۸ (۰/۱۲۰۴)

*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) درمتغیر کمی X. (اعدادداخل پراترها خطای معیار برآوردها می باشند)

برآورد پارامترها 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 از مبدا	-۴/۰۷۰۰ (۰/۱۳۰۶)	-۴/۰۷۹ (۰/۱۳۱۱)	-۳/۹۲۵۴ (۰/۱۴۶۰)	-۴/۱۱۳۸ (۰/۱۹۴۰)	-۴/۰۷۳۰ (۰/۱۳۷۹)
β_1 سیگار کشیدن (X)	۱/۸۲۴۴ (۰/۳۷۴۱)	۱/۷۸۹۰ (۰/۳۶۲۴)	۱/۹۷۹۵ (۰/۴۰۵۲)	۲/۱۶۷۹ (۰/۲۱۲۰)	۱/۶۸۹۰ (۰/۴۱۴۲)
β_2 جنس (Z)	۰/۸۵۷۶ (۰/۱۷۴۳)	۰/۸۶۰۸ (۰/۱۷۲۱)	۰/۶۹۷۱ (۰/۱۸۹۱)	۰/۹۹۰۵ (۰/۲۴۴۵)	۰/۸۳۲۹ (۰/۱۹۱۹)
β_{12} X*Z	-۰/۷۵۱۴ (۰/۴۱۱۱)	-۰/۷۱۵۶ (۰/۳۹۷۵)	-۱/۰۰۴۲ (۰/۴۴۳۹)	-۱/۲۹۷۷ (۰/۲۷۵۵)	-۰/۵۷۴۹ (۰/۴۵۰۲)

*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)

partially observed covariate. *Comm. Statist. Simul.* ۱۸(۱):۱۶۳-۱۷۷.

Carlin B.P. and Louis T.A. (۲۰۰۰) Bayes and Empirical Bayes Methods for Data Analysis. Second edition, Chapman and Hall.

Fleiss J.L., Levin B. and Paik M.C. (۲۰۰۳) Statistical Methods for Rates and Proportions. Third Edition, John Wiley and Sons.

Fuchs C. (۱۹۸۲) Maximum likelihood estimation and model selection in contingency tables with missing data. *J. Amer. Statist. Assoc.* ۷۷: ۲۷۰- ۲۷۸.

Gilks W.R., Richardson S. and Spiegelhalter D.J. (۱۹۹۷) Markov Chain

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*, ۴۲:۳۸۵-۳۸۹.
Blackhurst D.W. and Schluchter M.D. (۱۹۸۹) Logistic regression with a

- Satten G.A. and Carroll R.J. (۲۰۰۰) Conditional and unconditional categorical regression models with missing covariates. *Biometrics*. ۵۶: ۳۸۴-۳۸۸.
- 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*. ۴۴: ۴۲۹ – ۴۴۰.
- Seaman S.R. and Richardson S. (۲۰۰۱) Bayesian Analysis of case-control studies with categorical covariates. *Biometrika*. ۸۸: ۱۰۷۳-۱۰۸۸.
- Sinha S., Mukherjee B. and Ghosh M. (۲۰۰۴) Bayesian Semiparametric Modeling for Matched Case-Control Studies with Multiple Disease States. *Biometrics*. ۶۰: ۴۱-۴۹.
- Spiegelhalter D., Thomas A., Best N. and Lunn D. (۲۰۰۳) WinBUGS ۱.۴ Manual
- Zelen M. and Parker R.A. (۱۹۸۶) Case-Control Studies and Bayesian inference. *Statistics in Medicine*. ۵: ۲۶۱-۲۶۹.
- 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. (۱۹۸۵) Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*. ۷۲: ۴۹۷- ۵۱۲.
- Muller P., Parmigiani G., Schildkraut J. and Tardella L. (۱۹۹۹) A Bayesian Hierarchical Approach for combining case-control and prospective studies. *Biometrics*. ۵۵: ۸۵۸-۸۶۶.
- Muller P. and Roeder K. (۱۹۹۷) A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika*. ۸۴: ۵۲۳-۵۳۷.
- Nurminen M. and Mutanen P. (۱۹۸۷) Exact Bayesian analysis of two proportions. *Scandinavian Journal of Statistics*. ۱۴: ۶۷-۷۷.
- Paik M.C. and Sacco R.L. (۲۰۰۰) Matched Case–Control data analyses with missing covariates. *Applied Statistics*. ۴۹: ۱۴۶-۱۵۶.
- 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.