دانشور
پزشکی

روشی برای برآورد پارامترهای مدل رگرسیون لجستیک با وجود مقادیر گمشده در متغیر کمکی و کاربرد آن در بررسی بیماری گواتر

نویسندگان: کمال اعظم[1]، دکتر عباس گرامی[2]، دکتر کاظم محمد[3]، دکتر انوشیروان کاظم‌نژاد[4]، دکتر غلامرضا جندقی[5] و مسعود کریملو[6]

۱. دانشجوی دکترای آمار زیستی دانشگاه تربیت مدرس
۲. دانشیار پژوهشکده آمار
۳. استاد گروه آمار زیستی و اپیدمیولوژی دانشکده بهداشت دانشگاه علوم پزشکی تهران
٤. دانشیار دانشگاه تربیت مدرس
٥. استادیار دانشگاه تهران - پردیس قم
٦. مربی دانشگاه علوم بهزیستی و توانبخشی

دوماهنامه علمی - پژوهشی
دانشگاه شاهد
سال دوازدهم - شماره ۵۴
دی ۱۳۸۳

چکیده

سابقه و هدف: در جمع‌آوری داده‌های انبوه، بعضی از متغیرها با عدم پاسخ روبه‌رو می‌شوند که به این‌ها داده‌های گمشده می‌گویند. این داده‌های گمشده می‌توانند در متغیر پاسخ یا در متغیرهای کمکی به‌وجود آیند.

روش بررسی: در این مقاله داده‌های گمشده در متغیرهای کمکی مورد بررسی است و روش پیشنهادی برای تجزیه و تحلیل مدل‌های رگرسیون لجستیک را وقتی که متغیر پاسخ Y دو وضعیتی (بیمار و سالم) با مشاهدات کامل باشد و متغیر کمکی Z دارای مشاهدات کامل و متغیر کمکی X دارای مقادیر گمشده باشد مورد بررسی قرار داده‌ایم. در این‌جا فرض می‌کنیم مقادیر گمشده متغیر کمکی X تصادفی بوده است.

نتایج و یافته‌ها: برای این منظور، تابع درستنمایی برای داده‌های مشاهده شده را تعریف و برآوردهای حاصل را به روش ماکسیمم درستنمایی به‌دست آورده، سپس نتایج آن‌را با نرم‌افزار استاندارد S-Plus که داده‌های دارای مقادیر گمشده را حذف می‌کند مقایسه کردیم. برای تشریح بیش‌تر، هر دو روش را روی مثالی از بیماری گواتر اعمال کردیم.

بحث: مقایسه نتایج نشان داد که برآوردهای به‌دست آمده از مدل پیشنهادی نسبت به برآوردهای به‌دست آمده از نرم‌افزار که مبتنی بر حذف کلیه اطلاعات فرد دارای داده گمشده است دقت بیش‌تری دارند.

واژه‌های کلیدی: داده‌های گمشده، رگرسیون لجستیک، بیماری گواتر، ماکسیمم درستنمایی

مقدمه

رگرسیون‌لجستیک، ابزاری تحلیلی‌است که‌عموماً در تحقیقات پزشکی و اپیدمیولوژی استفاده زیاد دارد [۱]. در بسیاری از داده‌های پزشکی با مواردی مواجهه می‌شویم که در آن‌ها بخشی از داده‌ها گزارش نشده‌اند، از قبیل خودداری از پاسخ، عدم تکمیل کامل پرسشنامه‌ها یا پرونده‌ها، ناقص بودن چهارچوب مطالعه و غیره که در این صورت با داده‌های گمشده سرکار داریم. در این مطالعه فرض بر این است که این گمشدن به‌طور تصادفی رخ داده، مستقل از مقادیر مشاهده شده است (Missing At Random: MAR) [۲]. به‌عنوان مثال در تحلیل عوامل مؤثر بر بیماری گواتر ممکن است متغیرهایی همچون جنس، سن، محل سکونت، میزان مصرف ید و غیره مورد سؤال باشد و به دلایل ذکر شده برخی از این سؤالات بدون پاسخ باشند که این عدم پاسخ‌گویی مثلاً در سؤال میزان مصرف ید نباید متأثر از سن یا جنسیت و یا محل سکونت باشد.

برای تحلیل داده‌هایی که چنین خصوصیاتی دارند روش‌های مختلفی وجود دارد. ساده‌ترین روش این است که موارد دارای مقادیر گمشده را حذف کنیم و تجزیه و تحلیل براساس داده‌های کامل صورت پذیرد. این امر باعث از دست رفتن اطلاعات و حتی در بعضی از موارد سبب ایجاد اریبی می‌شود [۲]. این روش در پیش فرض اکثر نرم‌افزارهای آماری از جمله SAS، SPSS و S-Plus وجود دارد [۳]. روش دیگر این است که برآوردهایی جانشین مقادیر گمشده گردند و سپس با روش‌های استاندارد، تحلیل آماری برای کل داده‌ها، شامل مشاهده شده و گم شده صورت پذیرد. این روش که یک جانهی است نه برآورد، در صورت بالا بودن تعداد موارد گمشده، دارای دو اشکال عمده است، به این نحو که اولاً شکل طبیعی توزیع متغیر دارای مقادیر گمشده را تغییر می‌دهد و ثانیاً میانگین، واریانس و خطای معیار پارامتر (توابع نمونه‌ای) به‌دلیل اضافه شدن تعدادی مقادیر یکسان تغییر خواهد یافت [۲].

در مطالعه‌حاضر، استنباطی بر مبنای تابع درستنمایی با در نظر گرفتن مقادیر گمشده صورت می‌پذیرد که با روش‌های متداول برای داده‌های کامل متفاوت است و براساس حذف یا جانهی مقادیر گمشده نیست. طرز عمل برآورد ماکسیمم درستنمایی برای داده‌های کامل و داده‌های دارای مقادیر گمشده یکسان است، با این تفاوت که تابع درستنمایی در حالت با مقادیر گمشده، شامل تغییراتی است که در بخش بعدی به آن خواهیم پرداخت.

نویسندگان مختلفی روش‌های برخورد با مسائل مربوط به مقادیر گمشده را برای مدل‌های رگرسیون لجستیک معرفی کرده‌اند. لیتل و اشلاختر (Little and Schluchter) [۴]، و فوکس (Fuchs) [۵] از الگوریتم EM برای به‌دست آوردن برآوردهای ماکسیمم درستنمایی در رگرسیون لجستیک با متغیرهای کمکی گسسته‌یا ترکیبی از متغیرهای گسسته و پیوسته همراه با مقادیر گمشده استفاده کردند. الگوریتم EM عموماً نیاز به تکرار دارد. بلکهارست و اشلاختر (Bluckhurst and Schluchter) [۶] پیشنهاد کرده‌اند درصورتی‌که متغیر کمکی پیوسته باشد و از توزیع نرمال پیروی کند روش ماکسیمم درستنمایی با استفاده از الگوریتم EM نیاز به تکرار ندارد. آن‌ها با استفاده‌از روش‌مونت‌کارلو (Mont Carlo)، روش آنالیز داده‌های کامل، روش جانهی مقادیر گمشده و نیز روش درستنمایی را برای زمانی که دو متغیر کمکی وجود دارد – به‌صورتی‌که یکی از متغیرها تمام مقادیرش مشخص است و دیگری با مقادیر گمشده باشد - با هم مقایسه کردند و نتیجه گرفتند که روش درستنمایی بهتر از روش‌های دیگر عمل می‌کند [۳].

ساتن‌و کوپر (Satten and Kupper) در دو مقاله، روش تحلیل رگرسیون لجستیک را وقتی که متغیر کمکی دارای مقادیر گمشده بود بسط دادند و از متغیرهای جانشین برای پیدا کردن اطلاعی از اثر متغیرهای دارای مقادیر گمشده در مدل استفاده کردند [۷و۸]. پیک و ساکو (Paik and Sacco) روش‌هایی را برای تحلیل

مطالعات مورد شاهدی جور شده، وقتی که بعضی از متغیرهای کمکی دارای مقادیر گمشده هستند، ارائه دادند [۹]. ساتن و کارول در مقاله‌ای با تعیین توزیعی برای متغیر کمکی دارای مقادیر گمشده و اعمال تغییراتی در توابع درستنمایی شرطی و غیرشرطی رگرسیون لجستیک، برآورد پارامترها را بهبود بخشیدند [۱۰].

کارول، ساتن و راتور (Carrol, Satten and Rathour) مسأله‌ای از داده‌های با متغیرهای کمکی دارای مقادیر گمشده را در مدل‌های رگرسیون لجستیک شرطی براساس این‌که اولاً مدل‌بندی توزیع متغیرهای کمکی دارای مقادیر گمشده باشد و ثانیاً با استفاده از مدل‌بندی روند گمشدن مقادیر متغیر کمکی، کلاس جدیدی از برآوردها را ارائه داده‌اند [۱۱].

مدل

در این بخش به ارائه مدل رگرسیون لجستیک با وجود مقادیر گمشده در متغیر کمکی **X** و برآورد ماکزیمم درستنمایی آن می‌پردازیم.

فرض می‌کنیم که $\mathbf{Y_i}$ نشان‌دهنده متغیر پاسخ دو وضعیتی متناظر فرد **i** ام باشد که $\mathbf{Y_i}=1$، اگر فرد **i** ام بیمار باشد و $\mathbf{Y_i}=0$، اگر فرد **i** ام سالم باشد. همچنین فرض می‌کنیم که **X** و **Z** دو متغیر کمکی با مشاهدات کامل باشند. در حالت کلی، مدل اشباع رگرسیون لجستیک، احتمالات شرطی متغیر بیماری به شرط متغیرهای کمکی به‌صورت زیر تعریف می‌شود:

$$P(Y=1\,|\,X=x\ ,Z=z)=\frac{\exp(\beta_0+\beta_1 x+\beta_2 z+\beta_3 xz)}{1+\exp(\beta_0+\beta_1 x+\beta_2 z+\beta_3 xz)} \quad (۱)$$

و

$$P(Y=0\,|\,X=x\ ,Z=z)=\frac{1}{1+\exp(\beta_0+\beta_1 x+\beta_2 z+\beta_3 xz)} \quad (۲)$$

در مدل فوق می‌توان تعداد متغیر کمکی را بیش‌تر از دو نیز در نظر گرفت. بنابراین بخت (Odds) و نسبت بخت (یا نسبت برتری Odds Ratio) بیماری به ترتیب عبارتند از:

$$\theta(x,z)=\frac{P(Y=1\,|X=x,Z=z)}{P(Y=0\,|X=x,Z=z)}=\exp(\beta_0+\beta_1 x+\beta_2 z+\beta_3 xz) \quad (۳)$$

و

$$\psi(x,z,x',z')=\frac{\theta(x,z)}{\theta(x',z')} \quad (۴)$$

که در آن x' و z' نقاطی متفاوت از x و z هستند. هدف تحلیل رگرسیون لجستیک به‌دست آوردن برآورد از پارامترهای مدل (در این‌جا $\beta_0,\beta_1,\beta_2,\beta_3$) برای توصیف رابطه بین متغیر وابسته، **y**، مجموعه‌ای از متغیرهای کمکی، **x** و **z** است [۱۲]. در صورتی‌که **x** و **z** به‌طور کامل برای تمام افراد مشاهده شده باشد از روش‌های استاندارد جهت برآورد پارامترها استفاده می‌شود. حال فرض می‌کنیم که برخی از مقادیر **x** مشاهده نشده باشد و به عبارت دیگر برای متغیر کمکی x داده گمشده داشته باشیم. بدین‌ترتیب مقدار بخت (Odds) در مدل رگرسیون‌لجستیک عبارت‌خواهد بود از:

$$\tilde{\theta}(Z)=\frac{P(Y=1\,|Z=z)}{P(Y=0\,|Z=z)} \quad (۵)$$

به‌علاوه تعریف می‌کنیم:

$$\pi(x\,|z)=P\,(X=x\,|Y=0,Z=z) \quad (۶)$$

$$\rho(x\,|z)=P\,(X=x\,|Y=1,Z=z) \quad (۷)$$

همان‌طورکه مشاهده می‌شود توابع احتمال $\pi(x\,|\,z)$ و $\rho(x\,|\,z)$ به‌ترتیب نشان‌دهنده توزیع احتمال مقادیر متغیر **x** در افراد سالم و بیمار است. نتیجه مطالعه ساتن و کوپر [۷و۸] با استفاده از قضیه بیز عبارت است از:

$$\tilde{\theta}(z)=\sum_x \theta(x,z).\pi(x\,|z) \quad (۸)$$

که در آن، مجموع روی همه مقادیر ممکن متغیر **x** است. دومین نتیجه مقاله مذکور عبارت است از:

$$\rho(x\,|z)=\frac{\pi(x\,|z).\theta(x,\,z)}{\sum_x \pi(x\,|z)\,\theta(x,z)} \quad (۹)$$

حال با استفاده از موارد فوق به شرح زیر تابع درستنمایی را تشکیل می‌دهیم.

برآورد ماکسیمم درستنمایی با مقادیر گمشده در متغیر کمکی

تابع درستنمایی در مدل رگرسیون لجستیک استاندارد که متغیرهای کمکی x و z به‌طور کامل مشاهده شده باشند عبارت است از [۱۲]:

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

درصورتی‌که متغیر کمکی‌دارای‌مقادیر گمشده باشد، متغیر نشانگر Δ_i را به‌صورت زیر تعریف می‌کنیم:

$\Delta_i = ۱$ اگر x_i مشاهده شده باشد و $\Delta_i = ۰$ اگر x_i مشاهده نشده باشد. طبق تعریف لیتل و روبین (Little and Rubin) [۲] تابع درستنمایی را به‌صورت زیر به‌دست می‌آوریم:

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

با استفاده از روابط ۵ تا ۹ در رابطه فوق، تابع درستنمایی در حالتی که داده‌های گمشده داشته باشیم به شکل زیر تغییر می‌یابد [۲]:

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

از طرفی توزیع $\pi(x|z)$ نامعلوم است. در صورتی که x و z دارای مقادیری شمارا و محدود باشند. ساتن و کارول توزیعی از خانواده نمایی به شکل زیر را برای $\pi(x|z)$ پیشنهاد داده‌اند [۱۰]:

$$\pi(x|z)=\frac{e^{\gamma xz}}{\sum_{x'}e^{\gamma x'z}}=\frac{e^{\gamma_0+\gamma_1 x+\gamma_2 z+\gamma_3 xz}}{\sum_{x'}e^{\gamma_0+\gamma_1 x+\gamma_2 z+\gamma_3 xz}}=\frac{e^{\gamma_1 x+\gamma_3 xz}}{\sum_{x'}e^{\gamma_1 x'+\gamma_3 x'z}} \qquad (۱۳$$

یا با به‌کارگیری روابط ۳ تا ۹ و رابطه ۱۳ و بازنویسی مجدد تابع درستنمایی ۱۲، عبارت حاصل تابعی از پارامترهای β_0 و β_1 و β_2 و β_3 و γ_1 و γ_3 خواهد شد. پس از لگاریتم‌گیری از تابع درستنمایی نهایی، شکل تابع به‌صورت زیر است [۱۴]:

$$\ell(\beta)=\sum_{i=1}^{n}\Big\{\Delta_i Y_i(\beta_0+\beta_1 X_i+\beta_2 Z_i+\beta_3 X_i Z_i)-\ln\left[1+\sum_x e^{\beta_0+(\beta_1+\gamma_1)x+\beta_2 Z_i+(\beta_3+\gamma_3)xZ_i}\frac{1}{\sum_x e^{\gamma_1 x+\gamma_3 xZ_i}}\right]$$

$$+Y_i(1-\Delta_i)\ln\left[\sum_x e^{\beta_0+(\beta_1+\gamma_1)x+\beta_2 Z_i+(\beta_3+\gamma_3)xZ_i}\cdot\frac{1}{\sum_x e^{\gamma_1 x+\gamma_3 xZ_i}}\right]+\Delta_i(\gamma_1 X_i+\gamma_3 X_i Z_i)-\Delta_i\ln\sum_x e^{\gamma_1 x+\gamma_3 xZ_i}\Big\}$$

با مشتق‌گیری نسبت به تک تک پارامترها و مساوی صفر قرار دادن آن‌ها، دستگاه ۶ معادله ۶ مجهول حاصل، به‌دلیل غیرخطی بودن معادلات آن، به روش معمول قابل حل نیست و عملاً به‌کارگیری روش‌های عددی برای برآورد پارامترها ضرورت دارد.

روش برآورد: در مقاله حاضر، ابتدا با گرفتن مشتقات مراتب اول و دوم از رابطه ۱۴ نسبت به پارامترهای مذکور، دستگاه معادلات غیرخطی را ایجاد می‌کنیم و سپس در محیط نرم‌افزار آماری S-Plus و نیز نرم‌افزار R دستگاه به روش‌های تکرار طی مراحل زیر حل می‌شود:

مرحله اول: در این مرحله، متغیرهای x و z هر دو دارای مشاهدات کامل بودند و با دقت $e = ۰/۰۰۰۱$ مقادیر برآورد شده پارامترها با برنامه پیشنهادی و برنامه استاندارد موجود در نرم‌افزار مقایسه شدند.

مرحله دوم: در این مرحله، بخشی از متغیر x به‌صورت کاملاً تصادفی به داده گمشده تبدیل گردید و با دقتی ثابت مجدداً پارامترها با هر دو برنامه برآورد و مقایسه شدند.

مرحله سوم: در این مرحله، علاوه بر تغییر درصد گمشدگی در متغیر x دقت برآورد نیز تغییر داده شد و مقایسه‌های لازم صورت پذیرفت.

مثال: مطالعه سلامت و بیماری در ایران: استان قزوین

داده مثالی این مطالعه مربوط به طرح ملی سلامت و بیماری در ایران است که در سال ۱۳۸۰ در کل کشور به اجرا گذاشته شد. اطلاعات این مطالعه در مورد استان قزوین، شامل متغیر پاسخ دو حالتی بیماری گواتر به همراه دو متغیر کمکی محل سکونت و جنس که ارتباط معناداری با بیماری گواتر نشان دادند، مورد بررسی قرار گرفت [۱۳]. بیماری گواتر از شایع‌ترین بیماری‌های استان قزوین است که به درجه بزرگی غده تیروییید بستگی دارد. در این‌جا افراد سالم شامل گروه صفر وضعیت تیروییید هستند و افراد بیمار شامل گروه‌های 1A و بالاتر، یعنی گواتر کلی هستند [۱۴]. در این استان، افراد مورد بررسی از جهت وضعیت غده تیروییید ۷۵۸ نفر بودند که ۶۰ درصد بیماری گواتر داشتند. همچنین متغیرهای جنس (z) و محل سکونت (x) به ازای همه افراد به‌طور کامل مشاهده شده بود و عملاً داده گمشده وجود نداشت. جهت دستیابی به اهداف تحقیق حاضر با وجود محدودیت‌هایی مثل زمان مورد نیاز برای اجرای برنامه، درصدهای گمشدگی، میزان دقت متفاوت، حجم بالای داده‌ها و ظرفیت محدود حافظه کامپیوتر از این ۷۵۸ نفر، داده‌ای به حجم ۱۰۰ نفر به صورت کاملاً تصادفی انتخاب و درصدهای معینی از آن به تصادف حذف گردید و سپس برآورد پارامترهای مدل محاسبه شد. نتایج به‌دست آمده در بخش بعد ارائه گردیده است.

نتایج

در این بخش، چندین مرحله تجزیه و تحلیل روی این داده‌ها به تناسب اهداف تحقیق صورت پذیرفت. پیش از ارائه نتایج، ذکر این نکته‌ها حائز اهمیت است که هدف از تجزیه و تحلیل داده‌ها در این مطالعه، صرفاً ارزیابی مدل جدید و برنامه نرم‌افزاری تهیه شده برای آن است و بررسی رابطه معناداری بین متغیرهای کمکی با متغیر پاسخ مورد نظر نیست.

همان‌گونه که اشاره شد در این مثال، یک متغیر پاسخ دو حالتی بیماری گواتر با مقادیر Y=۱ برای بیمار و Y=۰ برای افراد سالم و دو متغیر کمکی جنسیت با مقادیر z=۰ برای مردان و z=۱ برای زنان و متغیر محل سکونت با مقادیر x=۰ برای ساکنان شهری و x=۱ برای ساکنان روستایی در نظر گرفته شد. ابتدا مدل رگرسیون لجستیک با وجود متغیرهای جنسیت و محل سکونت روی داده‌های مورد مطالعه برازش شد و در سطح معنا داری α =۰/۰۵ کلیه متغیرها به‌علاوه اثر متقابل آن‌ها در مدل باقی ماندند. نتایج در جدول ۱ خلاصه شده است. اعداد داخل جدول، برآورد پارامترهای مدل در وضعیت‌های گوناگون به همراه انحراف معیار برآوردها را نشان می‌دهد. اعداد ستون‌های اول و دوم مربوط به برآورد پارامترها برای داده‌های کامل با دقت e=۰/۰۰۰۱ است. مقایسه برآوردهای برنامه استاندارد S-Plus با برآوردهای برنامه پیشنهادی برای مدل جدید که در محیط S-Plus و نرم‌افزار R نوشته شده نشان می‌دهد که با دقت مذکور هر دو جواب برای تمام پارامترها تقریباً یکسان است و این تأییدی بر درست بودن مدل جدید و برنامه کامپیوتری نوشته شده محسوب می‌شود. به‌علاوه در بیش‌تر موارد، برآورد انحراف معیار پارامترهای مدل جدید همان‌طور که در جدول ۱ و ۲ مشاهده می‌شود از برآورد انحراف معیار پارامترها توسط مدل استاندارد کم‌تر است. پس از حذف ۲۰ درصد از داده‌های متغیر محل سکونت به‌صورت کاملاً تصادفی و اجرای مجدد هر دو برنامه روی داده‌های ناقص، خروجی‌های به‌دست آمده در ستون سوم و چهارم درج گردیده است. مقایسه مقادیر این دو ستون با یکدیگر و با ستون‌های اول و دوم حاکی است که برآوردهای حاصل از مدل جدید و برنامه کامپیوتری طراحی شده، نسبت به برآوردهای حاصل از نرم‌افزار استاندارد S-

Plus با وجود داده‌های گمشده، به مراتب به نتایج داده‌های کامل نزدیک‌تر است.

برای بررسی بیش‌تر و ارزیابی دقیق‌تر مدل جدید و برنامه نرم‌افزاری تهیه شده، تجزیه و تحلیل کامل تری روی داده‌ها صورت پذیرفت که نتایج آن در جدول ۲ ملاحظه می‌گردد. در این جدول، برآورد پارامترها مجدداً بر حسب دقت‌های مختلف، e، (۰/۰۵ و ۰/۰۱ و ۰/۰۰۵) با در نظر گرفتن درصد گمشدگی متفاوت، r، (۲۰، ۲۵، ۳۰ و ۳۵ درصد) در هر دو برنامه ارائه شده است. توجه به اعداد جدول حاکی از اختلاف بین برآوردهای دو مدل است. برای تأیید دقیق‌تر ابتدا برآوردهای دو مدل را از برآوردهای داده‌های کامل کسر کرده، برای مقادیر حاصل تحلیل واریانس سه عاملی چند متغیره انجام شد. در این آزمون، متغیر وابسته پارامترهای مختلف مدل و عوامل مورد بررسی عبارت بودند از: ۱. میزان دقت برآورد (e) ۲. میزان درصدهای گمشدگی تصادفی (r) ۳. نوع مدل (MS). نتایج نشان می‌دهد که تنها عامل نوع مدل معنادار است (p=۰/۰۰۲)؛ بدین‌نحو که در مجموع، اختلاف برآورد پارامترها بین دو نوع مدل تفاوت دارد و برآوردهای مدل جدید به برآوردهای مدل با داده کامل نزدیک‌تر هستند که این نشانه مثبتی از مناسب بودن مدل جدید برای تحلیل داده‌ها است.

نتیجه‌گیری

مثال انجام شده نشان می‌دهد برآوردهای جدید به مقادیر برآوردها براساس داده‌های کامل نزدیک‌تر هستند، ضمن این‌که از واریانس کم‌تری نیز برخوردارند.

جدول ۱: تجزیه و تحلیل داده‌های مربوط به بیماری گواتر استان قزوین* (اعداد داخل جدول حاصل نتایج نرم‌افزارهای S- Plus و برنامه مدل جدید است که توسط محقق در محیط R نوشته شده است)

برآورد پارامترها برای داده‌های کامل		برآورد پارامترها با ۲۰ درصد گمشدن تصادفی در متغیر محل سکونت		پارامترها	متغیرها
مدل استاندارد	مدل جدید	مدل استاندارد	مدل جدید		
-۰/۷۳۳۱۷	-۰/۹۷۹۹۶	-۰/۹۸۰۸۲۹	-۰/۹۸۰۸۵۲	β_0	عرض از مبدأ
(۰/۵۲۲۷۱)**	(۰/۵۱۰۲۱)	(۰/۴۷۸۵۹)	(۰/۴۹۰۱۷)		
۱/۵۰۳۳۹	۱/۷۹۵۴۷	۱/۷۹۱۷۵۹	۱/۷۹۱۷۹۲	β_1	جنس (z)
(۰/۷۰۲۷۹)	(۰/۶۹۴۷۸)	(۰/۷۶۸۱۹)	(۰/۷۵۸۴۴)		
۱/۸۱۵۳۰	۲/۴۳۲۰۹	۲/۲۶۸۶۷۳	۲/۲۶۸۷۲۵	β_2	محل سکونت (x)
(۰/۶۸۲۰۲)	(۰/۶۹۲۱۱)	(۰/۶۲۲۸۹)	(۰/۶۰۲۴۱)		
-۲/۳۸۲۷۹	-۲/۹۷۸۷۹	-۲/۷۹۱۹۲۱	-۲/۷۹۱۹۶۶	β_3	محل سکونت* جنس (x * z)
(۰/۹۱۰۴۹)	(۰/۹۲۴۷۹)	(۰/۹۴۵۹۹)	(۰/۹۶۹۴۷)		

* از داده‌های طرح سلامت و بیماری سال ۱۳۸۰ ـ کل کشور

** اعداد داخل پرانتز خطای معیار برآورد پارامترها است.

جدول ۲: برآورد ماکسیمم درستنمایی پارامترها بر حسب مدل استاندارد و مدل جدید به تفکیک درصدهای گمشدن تصادفی در متغیر محل سکونت با دو بار تکرار

		با دقت e=۰/۰۵				با دقت e=۰/۰۱				با دقت e=۰/۰۰۵			
		β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
r=٪۲۰	m_1	-۲/۷۲۸ (۰/۲۰۸)	۲/۳۴۷ (۰/۱۲۷)	۱/۷۴۳ (۰/۱۶۲)	-۰/۹۶۹ (۰/۰۸۴)	-۱/۰۰۳ (۰/۲۸۸)	۲/۳۱۸ (۰/۵۵۷)	۱/۷۴۴ (۰/۴۹۱)	-۲/۷۶۳ (۰/۶۲۵)	-۱/۰۴۴ (۰/۱۹۲)	۲/۱۳۱ (۰/۴۸۲)	۱/۸۳۱ (۰/۲۷۱)	-۲/۴۵۳ (۰/۵۴۷)
	s_1	-۲/۷۰ (۰/۶۰۴)	۲/۲۶۵ (۰/۵۳۴)	۱/۷۳۶ (۰/۳۴۰)	-۰/۹۸۴ (۰/۳۰۰)	-۰/۷۷۸ (۰/۲۶۹)	۲/۱۵۶ (۰/۴۲۹)	۱/۵۰۰ (۰/۳۵۷)	-۲/۵۵۷ (۰/۵۷۲)	-۰/۹۸۷ (۰/۱۳۲)	۲/۲۶۱ (۰/۳۳۴)	۱/۷۶۸ (۰/۲۲۵)	-۲/۶۳۴ (۰/۴۷۷)
	m_2	-۲/۸۰۲ **(۰/۳۹۸)**	۲/۲۴۷ **(۰/۱۵۲)**	۱/۶۹۹ **(۰/۳۱۰)**	-۰/۹۶۷ **(۰/۱۳۱)**	-۱/۰۹۳ (۰/۴۲۱)	۱/۸۲۳ (۰/۴۷۷)	۲/۲۶۵ (۰/۶۵۲)	-۲/۶۳۸ (۰/۷۲۵)	-۰/۸۹۱ (۰/۲۱۱)	۲/۲۸۰ (۰/۲۴۷)	۱/۵۸۹ (۰/۷۲۳)	-۲/۷۷۸ (۰/۸۲۵)
	s_2	-۲/۸۴۳ (۰/۶۲۱)	۲/۱۹۶ (۰/۵۱۸)	۱/۶۷۹ (۰/۰۸)	-۰/۹۰۳ (۰/۴۴۸)	-۰/۸۰۴ (۰/۳۸۹)	۱/۵۳۹ (۰/۳۹۹)	۱/۹۶۰ (۰/۴۹۵)	-۲/۳۷۴ (۰/۵۸۶)	-۰/۹۰۳ (۰/۱۹۱)	۲/۲۵۸ (۰/۲۵۷)	۱/۵۸۸ (۰/۵۹۱)	-۲/۷۰۶ (۰/۶۷۱)
r=٪۲۵	m_1	-۲/۸۳۰ (۰/۱۹۲)	۲/۲۷۰ (۰/۰۴۸)	۱/۷۷۴ (۰/۵۰۹)	-۰/۹۹۲ (۰/۲۲۶)	-۰/۹۶۹ (۰/۲۲۲)	۱/۸۲۱ (۰/۳۷۲)	۲/۲۶۰ (۰/۴۹۶)	-۲/۸۵۰ (۰/۵۸۱)	-۱/۰۲۲ (۰/۱۴۲)	۱/۷۹۲ (۰/۲۲۷)	۲/۲۵۲ (۰/۳۵۷)	-۲/۶۹۸ (۰/۵۴۱)
	s_1	-۲/۷۲۶ (۰/۴۹۹)	۲/۱۵۶ (۰/۱۹۹)	۱/۶۳۸ (۰/۲۸۶)	-۰/۸۳۶ (۰/۰۸۴)	-۱/۰۲۰ (۰/۲۲۸)	۱/۸۷۷ (۰/۳۱۱)	۲/۲۹۶ (۰/۳۹۸)	-۲/۸۵۱ (۰/۵۷۹)	-۰/۹۱۳ (۰/۱۷۵)	۱/۶۳۵ (۰/۲۲۱)	۲/۱۰۱ (۰/۴۱۱)	-۲/۵۳۷ (۰/۶۲۷)
	m_2	-۲/۸۲۳ (۰/۱۹۳)	۲/۳۳۸ (۰/۱۴۰)	۱/۷۶۷ (۰/۲۱۱)	-۱/۰۲۷ (۰/۱۳۰)	-۰/۹۷۵ (۰/۳۱۱)	۱/۷۸۹ (۰/۴۲۹)	۲/۲۶۶ (۰/۵۲۱)	-۲/۸۵۴ (۰/۶۱۵)	-۰/۹۴۹ (۰/۳۳۱)	۱/۶۴۰ (۰/۴۷۷)	۲/۰۳۲ (۰/۵۷۲)	-۲/۲۰۷ (۰/۶۲۱)
	s_2	-۳/۰۰۹ (۰/۷۲۲)	۲/۴۸۹ (۰/۵۲۳)	۱/۹۱۴ (۰/۶۱۴)	-۱/۱۷۰ (۰/۴۳۶)	-۰/۹۰۶ (۰/۳۰۱)	۱/۷۴۰ (۰/۳۴۸)	۲/۲۷۸ (۰/۴۹۲)	-۲/۸۸۲ (۰/۵۴۱)	-۰/۸۱۰ (۰/۱۹۷)	۱/۴۹۷ (۰/۲۴۲)	۱/۹۶۵ (۰/۳۴۱)	-۲/۲۶۴ (۰/۵۱۷)
r=٪۳۰	m_1	-۲/۶۱۲ (۰/۳۴۷)	۲/۱۶۵ (۰/۰۸۴)	۱/۷۵۷ (۰/۲۰۷)	-۰/۹۳۵ (۰/۰۸۲)	-۱/۰۲۱ (۰/۰۹۱)	۱/۷۳۷ (۰/۲۱۴)	۲/۱۶۹ (۰/۴۸۵)	-۲/۵۶۰ (۰/۷۴۲)	-۰/۹۳۸ (۰/۲۷۱)	۱/۶۵۲ (۰/۳۱۷)	۲/۲۲۲ (۰/۴۷۲)	-۲/۶۵۵ (۰/۴۴۷)
	s_1	-۲/۴۰۵ (۰/۵۴۸)	۲/۰۷۴ (۰/۲۵۹)	۱/۶۶۳ (۰/۴۵۷)	-۰/۸۲۳ (۰/۱۹۵)	-۱/۰۴۶ (۰/۱۸۱)	۱/۸۲۲ (۰/۲۵۱)	۲/۳۸۸ (۰/۴۹۶)	-۲/۸۴۵ (۰/۷۴۸)	-۰/۸۲۲ (۰/۲۱۲)	۱/۵۲۷ (۰/۲۷۹)	۲/۰۳۰ (۰/۴۱۳)	-۲/۴۷۲ (۰/۵۲۰)
	m_2	-۲/۷۳۷ (۰/۳۸۲)	۲/۲۹۶ (۰/۱۱۴)	۱/۶۹۲ (۰/۵۶۲)	-۰/۹۸۲ (۰/۲۴۹)	-۱/۰۳۲ (۰/۳۶۷)	۱/۸۰۶ (۰/۷۶۳)	۲/۲۸۶ (۰/۳۴۰)	-۲/۷۵۸ (۰/۶۶۱)	-۰/۸۶۵ (۰/۱۰۱)	۱/۶۳۴ (۰/۲۷۱)	۲/۲۰۸ (۰/۴۲۱)	-۲/۷۰۸ (۰/۴۹۹)
	s_2	-۲/۵۶۸ (۰/۸۵۵)	۲/۱۰۷ (۰/۳۸۱)	۱/۶۳۶ (۰/۷۲۸)	-۰/۸۸۳ (۰/۳۳۷)	-۱/۱۳۰ (۰/۳۲۴)	۱/۸۳۱ (۰/۳۸۷)	۲/۶۱۵ (۰/۴۷۸)	-۳/۰۹۰ (۰/۵۵۶)	-۰/۹۹۸ (۰/۰۹۹)	۱/۷۳۸ (۰/۲۶۵)	۲/۲۱۰ (۰/۴۲۵)	-۲/۶۷۴ (۰/۴۵۹)
r=٪۳۵	m_1	-۲/۴۵۴ (۰/۵۱۷)	۲/۲۰۵ (۰/۲۱۷)	۱/۶۲۳ (۰/۵۵۹)	-۱/۰۰۹ (۰/۲۸۹)	-۰/۸۶۶ (۰/۱۸۷)	۱/۴۰۹ (۰/۲۲۹)	۲/۴۲۹ (۰/۴۲۱)	-۲/۷۱۴ (۰/۵۷۲)	-۱/۰۱۹ (۰/۱۹۷)	۱/۸۴۴ (۰/۲۷۵)	۲/۳۱۱ (۰/۴۲۷)	-۲/۹۷۳ (۰/۵۹۲)
	s_1	-۲/۳۶۸ (۰/۴۴۲)	۲/۰۰۸ (۰/۳۰۵)	۱/۴۵۷ (۰/۲۲۱)	-۰/۸۰۹ (۰/۳۲۸)	-۰/۷۵۴ (۰/۲۱۰)	۱/۲۰۸ (۰/۳۱۱)	۲/۲۲۹ (۰/۳۴۷)	-۲/۳۸۹ (۰/۴۸۷)	-۱/۰۳۹ (۰/۱۱۷)	۱/۸۵۰ (۰/۲۵۷)	۲/۲۷۷ (۰/۴۲۴)	-۲/۹۳۰ (۰/۵۲۶)
	m_2	-۲/۴۹۵ (۱/۲۷۵)	۲/۰۷۴ (۰/۵۴۸)	۱/۷۲۲ (۰/۸۷۳)	-۰/۹۹۸ (۰/۳۵۶)	-۱/۰۰۱ (۰/۳۴۳)	۱/۸۱۲ (۰/۳۲۹)	۲/۱۸۱ (۰/۴۷۲)	-۲/۷۳۵ (۰/۶۹۴)	-۰/۹۹۱ (۰/۱۹۲)	۱/۸۱۵ (۰/۴۲۱)	۲/۰۴۰ (۰/۳۹۴)	-۲/۳۷۱ (۰/۵۲۱)
	s_2	-۳/۰۱۰ (۰/۴۵۸)	۲/۳۵۳ (۰/۴۴۷)	۱/۸۳۰ (۰/۳۰۶)	-۱/۰۳۶ (۰/۳۷۲)	-۱/۰۷۲ (۰/۲۴۲)	۱/۸۷۹ (۰/۳۱۱)	۲/۲۴۵ (۰/۴۱۲)	-۲/۹۳۸ (۰/۵۸۷)	-۰/۷۹۶ (۰/۱۷۲)	۱/۶۲۰ (۰/۳۰۲)	۲/۱۷۸ (۰/۴۵۰)	-۲/۷۱۰ (۰/۴۰۵)

m_i برآورد پارامترها تحت مدل جدید در تکرار i ام

s_i برآورد پارامترها تحت برنامه استاندارد *s-plus* در تکرار i ام

r درصدهای مختلف گمشدگی در متغیر کمکی

منابع

1. Stuart R.L., Michael P. and Marium E., "Inference using conditional logistic regression with missing covariates". Biometrics 1998, 54, 295–303.
2. Little R.J.A. and Rubin, D.B., "Statistical analysis with missing data". John Wiley & Sons, Second Edition, New York, 2002.
3. Gao S. and Hui S.L., "Logistic regression models with missing covariate value for complex survey data". Statistics in Medicine, 1997, 16, 2419-2428.
4. Little R.J.A. and Schluchter, M.D., "Maximum likelihood estimation for mixed continuous and categorical data with missing values". Biometrika, 1985, 72, 497-512.
5. Fuchs, C. "Maximum likelihood estimation and model setection in contingency tables with missing data". J. Amer. Statist. Assoc. 1982, 77, 270-278.
6. Blackhurst, D.W. and Schluchter, M.D. "Logistic regression with a partially observed covariate". Comm. Statist. Simul. 1989, 18(1), 163-177.
7. Satten, G.A. and Kupper L. "Inferences about exposure – disease associations using probability of exposure information". J. Amer. Statist. Assoc. 1993a, 88, 200-208.
8. Satten G.A. and Kupper L., "Conditional regression analysis of the odds ratio between two binary variables when one is not measured with certainty " A method for epidemiologic studies. Biometries 1993b, 44, 429–440.
9. Paik M.C. and Sacco R.L. "Matched case – Control data analyses with missing covariates" Applied Statistics, 2000, 49, 146-156.
10. Satten, G.A. and Carroll R.J., "Conditional and unconditional categorical regression models with missing covariates". Biometrics, 2000, 56, 384-388.
11. Rathouz P.J., Satten G.A. and, Carrol R.J., "Semiparametric inference in matched case – control studies with missing covariate data". Biometrika, 2003.
12. Armitage, P. and Colton, "Encyclopedia of biostatistics". John Wiley, New York 1997, pp.2316–2327.

۱۳. نوربالا، احمدعلی؛ محمد، کاظم «بررسی سلامت و بیماری در ایران- سال ۱۳۸۰»؛ انتشارات مرکز ملی تحقیقات علوم پزشکی کشور.

۱۴. زالی، محمدرضا؛ محمد، کاظم؛ اعظم، کمال؛ مسجدی، محمدرضا «وضعیت تیرویید در ایران براساس نتایج طرح سلامت و بیماری» مجله علمی نظام پزشکی، دوره سیزدهم از ۱۱۳ تا ۱۲۲ سال ۱۳۷۳.