مدل رگرسیون لجستیک چند حالته با مقادیر گم شده و کاربرد آن در بررسی بیماری گواتر

کمال اعظم*[1]، دکتر عباس گرامی[2]، دکتر کاظم محمد[3] و دکتر انوشیروان کاظم نژاد[1]

چکیده:

در جمع آوری داده های انبوه بویژه در بررسی سلامت و بیماری در ایران بعضی از متغیرها با عدم پاسخ روبرو می شوند که به اینها داده های گم شده گویند. این داده های گم شده می تواند در متغیر پاسخ یا در متغیرهای کمکی بوجود آید. در این مقاله داده های گم شده در متغیرهای کمکی مورد بررسی است. روش پیشنهادی برای تجزیه و تحلیل مدلهای رگرسیون لجستیک وقتی که متغیر پاسخ *(Y)* چند وضعیتی باشد و متغیر کمکی *(Z)* دارای مشاهدات کامل و متغیر کمکی *(X)* دارای مقادیر گم شده باشد مورد بررسی قرار داده ایم . در اینجا فرض شده است که مقادیر گم شده متغیر کمکی *(X)* به طور تصادفی گم شده اند. برای این منظور تابع درست نمایی برای داده های مشاهده شده را به دست آورده و سپس نتایج آن با روشهای معمول که مبتنی بر حذف مقادیر گم شده هستند و معمولاً در نرم افزارهای متداول نظیر *SPSS* بکار می رود مقایسه شده اند. برای تشریح بیشتر، هر دو روش روی مثالی در مورد بیماری گواتر که دارای پاسخهای چند حالته است به کار برده شد. مقایسه نتایج نشان داد که مدل پیشنهادی بهتر عمل می نماید.

واژگان کلیدی : *داده های گم شده تصادفی، رگرسیون لجستیک ، بیماری گواتر، ماکزیمم درست نمایی ، پاسخهای چند حالته ، تیروئید*

* (عهده دار مکاتبات)

۱ .گروه آمار حیاتی دانشکده پزشکی دانشگاه تربیت مدرس

۲ . پژوهشکده آمار

۳ .گروه اپیدمیولوژی و آمار زیستی دانشکده بهداشت و انستیتو تحقیقات بهداشتی, دانشگاه علوم پزشکی و خدمات بهداشتی درمانی تهران

مقدمه :

رگرسیون لجستیک ابزاری تحلیلی است که عموماً در تحقیقات پزشکی و اپیدمیولوژی مورد استفاده زیادی دارد (*Stuart R.L. et al. 1998*). در تحقیقات اپیدمیولوژی محقق در صدد محاسبهٔ مقدار بخت (برتری) (Odds) و نسبت بخت (نسبت برتری) (Odds Ratio) و خطر نسبی مواجهه در بروز بیماری است. تحلیل با این مقادیر عموماً به وفور در بسیاری از مقالات مشاهده می شود. از آنجا که در معادلات رگرسیون لجستیک پارامترهای برآورد شده منجر به برآورد مقادیر(Odds Ratio) نسبت بخت می شوند، این مقاله بر آن است که مدل رگرسیون لجستیک را در حالت خاصی که متغیر پاسخ بیش از دو حالت دارد و متغیرهای کمکی دارای مقادیر گم شده هستند را مورد بررسی قرار دهد. در بسیاری از داده های پزشکی با مواردی مواجهه می شویم که در آنها بخشی از داده ها گزارش نشده اند از قبیل خودداری از پاسخ، عدم تکمیل کامل پرسشنامه ها یا پرونده ها، ناقص بودن چهارچوب مطالعه و غیره. در این صورت با داده های گم شده سر وکار داریم در این مطالعه فرض بر این است که این گمشدگی به طور تصادفی رخ داده و مستقل از مقادیر مشاهده شده باشد (Missing At Random)*MAR* (Little R.J.A. and Rubin D.B. 2002). به عنوان مثال در تحلیل عوامل مؤثر بر بیماری گواتر ممکن است متغیرهایی همچون جنس ، سن ، محل سکونت ، میزان مصرف ید و غیره مورد سؤال بوده و به دلایل ذکر شده برخی از این سؤالها بدون پاسخ باشند و این عدم پاسخ گویی مثلاً درسؤال میزان مصرف ید نباید متأثر از سن یا جنسیت و یا محل سکونت باشد.

برای تحلیل داده هایی با این خصوصیات روشهای مختلفی وجود دارد. ساده ترین روش این است که موارد دارای مقادیر گم شده را حذف و تجزیه و تحلیل بر اساس داده های کامل صورت پذیرد. این امر باعث از دست رفتن اطلاعات و حتی در بعضی از موارد سبب ایجاد اریبی می شود (Little R.J.A. and Rubin D.B. 2002). این روش در پیش فرض اکثر نرم افزارهای آماری از جمله SAS ، SPSS و S-Plus وجود دارد (Gao S. and Hui S.L. 1997) . روش دیگر این است که برآوردهایی جانشین مقادیر گم شده گردد و سپس با روشهای استاندارد، تحلیل آماری برای داده های کامل صورت پذیرد. این روش، در صورت بالا بودن تعداد موارد گم شده , دارای دو اشکال عمده می باشد. اولاً شکل طبیعی توزیع متغیر دارای مقادیر گم شده را تغییر می دهد و ثانیاً میانگین ، واریانس و خطای معیار پارامتر (توابع نمونه ای) به دلیل اضافه شدن تعدادی مقادیر یکسان تغییر خواهد یافت.

در مطالعه حاضر، استنباط بر مبنای تابع درست نمایی با در نظر گرفتن مقادیر گم شده صورت می پذیرد و با روشهای متداول برای داده های کامل متفاوت است و بر اساس حذف یا جانهی مقادیر گم شده نمی باشد. طرز عمل برآورد ماکسیمم درست نمایی برای داده های کامل و داده های دارای مقادیر گم شده یکسان است ، با این تفاوت که تابع درست نمایی در حالت با مقادیر گم شده شامل تغییراتی است که در بخش بعدی به آن خواهیم پرداخت.

نویسندگان مختلفی روشهای برخورد با مسائل مربوط به مقادیر گم شده را برای مدلهای رگرسیون لجستیک معرفی کرده اند . بعضی از نویسندگان الگوریتم EM را برای به دست آوردن برآوردهای ماکسیمم درست نمایی در رگرسیون لجستیک با متغیرهای کمکی گسسته یا ترکیبی از متغیرهای گسسته و پیوسته همراه با مقادیر گم شده پیشنهاد کرده اند (Fuchs C. 1982, Little R.J.A. and Schluchter M.D. 1985) . الگوریتم EM عموماً نیاز به تکرار دارد. در صورتی که متغیر کمکی پیوسته باشدو از توزیع نرمال پیروی کند روش ماکسیمم درست نمایی با استفاده از الگوریتم EM نیاز به تکرار ندارد. در یک بررسی با استفاده از روش مونت کارلو ، سه روش آنالیز داده ها، که با استفاده از داده های کامل دیگری جانهی مقادیر گم شده و سومی روش درست نمایی ماکزیمم برای زمانی که دو متغیر کمکی وجود دارد و تنها یکی از شامل مقادیر گم شده بود با یکدیگر مقایسه شدند و نتیجه گرفته شد که روش درست نمایی بهتر از روشهای دیگرعمل می کند (Blackhurst D.W. and Schluchter M.D. 1989).

محققان دیگری روش تحلیل رگرسیون لجستیک را وقتی که متغیر کمکی دارای مقادیر گم شده بود بسط دادند و از متغیرهای جانشین برای پیدا کردن اطلاعی از اثر متغیرهای دارای مقادیرگم شده در مدل استفاده کردند (Sattan G.A. and Kupper L.1993a, Sattan G.A.and Kupper L.1993b). برای تحلیل مطالعات مورد شاهدی جور شده وقتی که بعضی از متغیرهای کمکی

دارای مقـادیر گم شده هستند نیـز روشهـایی ارائـه شـده اسـت (Paik M.C.and Sacco R.L. 2000) . همچنین در مقاله دیگری، محققان با تعیین توزیعی برای متغیر کمکی دارای مقادیر گم شده و اعمـال تغییـراتی در توابع درست نـمایی شرطی و غیر شـرطی رگرسیون لجستیک، برآورد پـارامترها را بهبـود بخشیـدند (Satten G.A. and Carrol R.J. 2000).

علاوه بر موارد فوق، کلاس جدیدی از برآوردها ارائه شده است که بر اساس مدل بندی توزیع متغیرهای کمکی با داده های گم شده ومدل بندی روند گم شدن مقادیر متغیر کمکی پایه گذاری شده است(Rathouz P.J. et al. 2003).

در منابع مورد بررسی، مطالعات انجام شده ، متغیر پاسخ دو وضعیتی در نظر گرفته شده است. مقاله حاضر قصد دارد روشی را برای مطالعه رگرسیون لجستیک با وجود مقادیر گم شده در متغیرهای کمکی وقتی که متغیر پاسخ چند وضعیتی است, مورد بررسی قرار دهد.

مدل : در این بخش به ارائه مدل رگرسیون لجستیک با پاسخهای چند وضعیتی با وجود مقادیر گم شده در متغیر کمکی X و برآورد ماکزیمم درست نمـایی آن می پردازیم.

فـرض می کنیم که Y_i نشـان دهنده متغیر پـاسخ چند وضعیتی باشد که سه مقدار ۰ ، ۱ و ۲ را بگیرد. همچنین فرض می کنیم که X و Z دو متغیر کمکی با مشاهدات کامل باشند، حالت کلی در مدل اشباع رگرسیون لجستیک چند حالته, احتمـالات شـرطی متغیـر پـاسخ با مقـادیر مختـلف به شـرط متغیـرهای کمکی به صـورت زیـر تعـریـف مـی شـود (Hosmer D.W. and Lemeshow Jr.S. 1989, Kleinbaum D.G. and Klein M. 2002).

$$(۱)\quad \pi_0(x) = P(Y=0 \mid X=x, Z=z) = \frac{1}{1+\exp(\beta_{10}+\beta_{11}x+\beta_{12}z+\beta_{13}xz)+\exp(\beta_{20}+\beta_{21}x+\beta_{22}z+\beta_{23}xz)}$$

$$(۲)\quad \pi_1(x) = P(Y=1 \mid X=x, Z=z) = \frac{\exp(\beta_{10}+\beta_{11}x+\beta_{12}z+\beta_{13}xz)}{1+\exp(\beta_{10}+\beta_{11}x+\beta_{12}z+\beta_{13}xz)+\exp(\beta_{20}+\beta_{21}x+\beta_{22}z+\beta_{23}xz)}$$

$$(۳)\quad \pi_2(x) = P(Y=2 \mid X=x, Z=z) = \frac{\exp(\beta_{20}+\beta_{21}x+\beta_{22}z+\beta_{23}xz)}{1+\exp(\beta_{10}+\beta_{11}x+\beta_{12}z+\beta_{13}xz)+\exp(\beta_{20}+\beta_{21}x+\beta_{22}z+\beta_{23}xz)}$$

چون Y_i سه وضعیتی در نظر گرفته شده است بنابراین دو بخت (Odds) و دو نسبت بخت (یا نسبت برتری Odds Ratio) به صورت زیر تعریف می شود:

$$(۴)\quad \theta_1(x,z) = \frac{P(Y=1 \mid X=x, Z=z)}{P(Y=0 \mid X=x, Z=z)} = \exp(\beta_{10}+\beta_{11}x+\beta_{12}z+\beta_{13}xz)$$

$$(۵)\quad \theta_2(x,z) = \frac{P(Y=2 \mid X=x, Z=z)}{P(Y=0 \mid X=x, Z=z)} = \exp(\beta_{20}+\beta_{21}x+\beta_{22}z+\beta_{23}xz)$$

و

$$(۶)\quad \psi_1(x,z,x',z') = \frac{\theta_1(x,z)}{\theta_1(x',z')}$$

$$\psi_2(x, z, x', z') = \frac{\theta_2(x, z)}{\theta_2(x', z')} \qquad (۷)$$

هدف تحلیل رگرسیون لجستیک بدست آوردن برآورد پارامترهای مدل (در اینجا $\beta_{23}, \beta_{22}, \beta_{21}, \beta_{20}$ ، $\beta_{13}, \beta_{12}, \beta_{11}, \beta_{10}$) بـرای تـوصیـف رابـطه بیـن متغیـر وابسته Y و مجمـوعه ای از متغیـرهای کمـکی می بـاشد (Armitage P. and Colton T. 1997) . در مقاله حاضر دو متغیر کمکی X وZ در نظر گرفته ایم . در صورتی که این دو متغیر کمکی به طـور کـامل بـرای تمام افـراد مشاهده شده باشند از روشهای اسـتاندارد جهت برآورد پارامترها استفاده می شود. حـال فـرض می کنیم که بـرخی از مقـادیر X مشاهده نشـده باشد و به عبـارت دیـگر برای متغیـر کمـکی X دادهٔ گم شده داشته باشیم، در ایـن صورت متـغیر نشـانگر Δ_i را به صورت زیر تعریف می کنیم :

۱ = Δ_i اگر X_i مشـاهده شـده بـاشـد و ۰ = Δ_i اگـر X_i مشـاهـده نشـده باشـد. بـدین تـرتیب مقـدار بخـت (Odds) و نسبـت بخـت (Odds Ratio) بدون حضور متغیر تصادفی X در مدل رگرسیون لجستیک عبارت خواهد بود از :

$$\tilde{\theta}_1(Z) = \frac{P(Y=1 \mid Z=z)}{P(Y=0 \mid Z=z)} \qquad (۸)$$

$$\tilde{\theta}_2(Z) = \frac{P(Y=2 \mid Z=z)}{P(Y=0 \mid Z=z)} \qquad (۹)$$

$$\tilde{\psi}_1(z, z') = \frac{\tilde{\theta}_1(z)}{\tilde{\theta}_1(z')} \qquad (۱۰)$$

$$\tilde{\psi}_2(z, z') = \frac{\tilde{\theta}_2(z)}{\tilde{\theta}_2(z')} \qquad (۱۱)$$

به علاوه, تعاریف زیر را می سازیم.

$$\rho_0(X \mid Z) = P(X = x \mid Y = 0, Z = z) \qquad (۱۲)$$

$$\rho_1(X \mid Z) = P(X = x \mid Y = 1, Z = z) \qquad (۱۳)$$

$$\rho_2(X \mid Z) = P(X = x \mid Y = 2, Z = z) \qquad (۱۴)$$

همان طور که مشاهده می شود توابع احتمال $\rho_0(X \mid Z)$ ، $\rho_1(X \mid Z)$ و $\rho_2(X \mid Z)$ به ترتیب نشان دهندهٔ توزیع احتمـال مقادیر متغیر X به شرط Y=0 ، Y=1 و Y=2 و معلوم بودن متغیر کمکی Z می باشد . با استفاده از قضیه احتمال بیز و فرمولهای ۱۲، ۱۳ و ۱۴

فرمولهای ۸ و ۹ به صورت زیر تغییر می کند.

$$\tilde{\theta}_1(z) = \sum_x \theta_1(x, z) . \rho_0(x \mid z) \qquad (۱۵)$$

$$\tilde{\theta}_2(z) = \sum_x \theta_2(x, z) . \rho_0(x \mid z) \qquad (۱۶)$$

که در آن مجموع روی همهٔ مقادیر ممکن متغیر X می باشد و همچنین خواهیم داشت :

(۱۷)
$$\rho_1(\mathrm{X}|Z)=\frac{\theta_1(X,Z)\rho_0(X|Z)}{\sum_x \theta_1(x,Z)\rho_0(x|Z)}$$

(۱۸)
$$\rho_2(\mathrm{X}|Z)=\frac{\theta_1(X,Z)\rho_0(X|Z)}{\sum_x \theta_1(x,Z)\rho_0(x|Z)}$$

حال با استفاده از موارد فوق به شرح زیر تابع درست نمایی را تشکیل می دهیم.

تابع درست نمایی و برآورد ماکزیمم درست نمایی با مقادیر گم شده در متغیر کمکی تابع درست نمایی در مدل رگرسیون لجستیک وقتی که Y متغیر پاسخ چند وضعیتی و متغیرهای کمکی X و Z به طور کامل مشاهده شده باشند عبارتست از :

(۱۹)
$$L(\beta)=\prod_{i=1}^{n}\left\{-[\pi_0(x_i)]^{y_{0i}}[\pi_1(x_i)]^{y_{1i}}[\pi_2(x_i)]^{y_{2i}}\right\}$$

که در آن $\sum_{j=0}^{2} y_{ji}=1$ برای هر i می باشد (Hosmer D.W. and Lemeshow Jr. S. 1989).

در صورتی که متغیر کمکی دارای مقادیر گم شده باشد تابع درست نمایی را به صورت زیر به دست می آوریم (Little R.J.A. and Rubin D.B. 2002):

(۲۰)
$$P(Y,X,\Delta|Z)=P(Y|Z).\ \mathrm{P}(\Delta|Y,Z)\mathrm{P}(\mathrm{X}\mid \mathrm{Y},\mathrm{Z},\Delta)$$

با استفاده از (۱۵) الی (۱۹) تابع درست نمایی فوق در حالتی که داده های گم شده داشته باشیم به شکل زیر تغییر می یابد.

(۲۱)
$$L(\beta)=\prod_{i=1}^{n}\left\{-[\pi_0(z_i)]^{y_{0i}}[\pi_1(z_i)]^{y_{1i}}[\pi_2(z_i)]^{y_{2i}}\right\}$$
$$[\rho_0(X_i|\mathrm{Z_i})]^{\Delta_i y_{0i}}[\rho_1(X_i|\mathrm{Z_i})]^{\Delta_i y_{1i}}[\rho_2(X_i|\mathrm{Z_i})]^{\Delta_i y_{2i}}$$

با توجه به اینکه $\sum_{j=0}^{2} y_{ji}=1$ و استفاده از روابط (۱۵) الی (۱۸) تابع درست نمایی به شکل زیر خواهد شد.

(۲۲)
$$=\prod_{i=1}^{n}\left\{\frac{[P_0(X_i|\mathrm{Z_i})]^{\Delta_i y_{0i}}[P_1(X_i|\mathrm{Z_i})]^{\Delta_i y_{1i}}\left[\sum_x \rho_0(x_i|z_i)\theta_1(x_i,z_i)\right]^{(1-\Delta_i)y_{1i}}\times[\rho_0(x_i|z_i)\theta_1(x_i,z_i)]^{\Delta_i y_{2i}}\left[\sum_x \rho_0(x_i|z_i)\theta_1(x,z_i)\right]^{(1-\Delta_i)y_{2i}}}{1+\sum_x \rho_0(x_i|z_i)\theta_1(x,z_i)+\sum_x \rho_0(x_i|z_i)\theta_2(x,z_i)}\right.$$

از طرفی توزیع $\rho_0(x|z)$ نامعلوم می باشد. در صورتی که X و Z دارای مقادیری متعدد اما محدود باشند می توان توزیعی از خانواده نمایی به شکل زیر را برای $\rho_0(x|z)$ در نظر گرفت که نتایج جالب زیر را به دست می دهد (Satten G.A.and Carrol R.J. 2000):

(۲۳)
$$\rho_0(x|z)=\frac{e^{\gamma xz}}{\sum_{x'} e^{\gamma x'z}}=\frac{e^{\gamma_0+\gamma_1 x+\gamma_2 z+\gamma_3 xz}}{\sum_{x'} e^{\gamma_0+\gamma_1 x+\gamma_2 z+\gamma_3 xz}}=\frac{e^{\gamma_1 x+\gamma_3 xz}}{\sum_{x'} e^{\gamma_1 x'+\gamma_3 x'z}}$$

با به کارگیری روابط (۱۵) تا (۱۸) و رابطهٔ (۲۳) و بازنویسی مجدد تابع درست نمایی (۲۲)، عبارت حاصل تابعی از پارامترهای

γ_3 ، γ_1، β_{23} ، β_{22} ، β_{21} ، β_{20} ، β_{13} ، β_{12} ، β_{11} ، β_{10} خواهد شد. پس از لگاریتم گیری از تابع درست نمایی نهایی ، با مشتق گیری نسبت به تک تک پارامترها و مساوی صفر قرار دادن آنها، دستگاه ۱۰ معادلهٔ ۱۰ مجهولی حاصل، به دلیل غیر خطی بودن، به روش معمول قابل حل نمی باشد و عملاً نیاز به کارگیری روشهای عددی برای برآورد پارامترها است. بدین منظور با استفاده از مدل ارائه شده مثال زیر را برای آن در نظر گرفتیم.

مثال : مطالعهٔ سلامت و بیماری در ایران ـ استان قزوین

دادهٔ مثالی این مطالعه مربوط به طرح ملی سلامت و بیماری در ایران است که در سال ۱۳۸۰ در کل کشور به اجرا گذاشته شد. در این مطالعه که نتایج مقایسه استانی در سال ۱۳۸۱ منتشر گردید اطلاعات مربوط به تیروئید قابل مشاهده که در جدول شیوع برخی از بیماریهای غیر واگیر در استانهای کشور این مجموعه ارائه گردیده حاکی از این است که استان قزوین از لحاظ شیوع بیماری گواتر با میزان شیوع ۱۱/۴٪ دارای مقام نخست است و پس از آن استانهای کردستان و یزد به ترتیب با میزانهای شیوع ۱۰/۸٪ و ۹/۶٪ قرار دارد. این میزان شیوع در مناطق شهری استان قزوین ۶/۷٪ و در مناطق روستایی ۱۷/۸٪ و در مردان ۹/۹٪ و در زنان ۱۲/۶٪ می باشد (نوربالا و محمد ۱۳۸۱). در مطالعهٔ قبلی سلامت و بیماری که نتایج آن در سال ۱۳۷۰ منتشر گردید درصد اطلاعات تیروئید قابل رؤیت شهرستان قزوین که جزیی از اطلاعات استان زنجان را تشکیل می داد نیز حاکی از بالا بودن این میزان بود (زالی و همکاران ۱۳۷۰). این مطلب موجب گردید که این استان جهت بررسی تأثیر متغیرهای مستقل متنوع موجود در پرسشنامه طرح مذکور که بعضاً برخی از آنها دارای مقادیر گم شده بودند انتخاب شود. پس از تحلیل رگرسیون لجستیک مشخص گردید که متغیرهای کمکی محل سکونت و جنس ارتباط معنی داری با متغیر پاسخ دارند. در این مطالعه متغیر پاسخ وضعیت تیروئید به سه حالت، افراد سالم شامل گروه صفر وضعیت تیروئید ، افراد گروه یک شامل گروه ۱A وضعیت تیروئید و گروه دو شامل گروههای ۱B و بالاتر وضعیت تیروئید هستند (زالی وهمکاران ۱۳۷۳). در این استان، افراد مورد بررسی از جهت وضعیت غدهٔ تیروئید شامل ۷۵۸ نفر بودند که ۶۰٪ مبتلا به یکی از درجات غدهٔ تیروئید بودند (نور بالا و محمد ۱۳۸۰). در اینجا متغیر پاسخ Y=۰ شامل افراد سالم و Y=۱ شامل افراد با وضعیت تیروئید ۱A و Y=۲ شامل افراد با وضعیت تیروئید ۱B و بالاتر می باشند. اگرچه این یک متغیر کیفی رتبه ای می باشد ولی در این بررسی فقط حالت کیفی اسمی چند حالته این متغیر در نظر گرفته شد. متغیرهای جنس (Z) و محل سکونت (X) به ازای همهٔ افراد به طور کامل مشاهده شده بود و عملاً دادهٔ گم شده وجود نداشت. جهت دستیابی به اهداف تحقیق حاضر با وجود محدودیتهایی مثل زمان مورد نیاز برای اجرای برنامه، درصدهای گم شدگی، حجم بالای داده ها و ظرفیت محدود حافظهٔ کامپیوتر از بین ۷۵۸ نفر، نمونه ای به حجم ۱۲۰ نفر به صورت کاملاً تصادفی انتخاب و درصدهای معینی از آن به تصادف حذف و سپس برآورد پارامترهای مدل محاسبه شد. نتایج به دست آمده در بخش بعد ارائه گردیده است.

نتایج :

در این بخش ، چندین مرحلهٔ تجزیه و تحلیل روی این داده ها به تناسب اهداف تحقیق صورت پذیرفت. پیش از ارائه نتایج ، ذکر این نکته حائز اهمیت است که، هدف از تجزیه و تحلیل داده ها در این مطالعه، صرفاً ارزیابی مدل جدید و برنامه نرم افزاری تهیه شده برای آن می باشد و بررسی رابطهٔ معنی داری بین متغیرهای کمکی با متغیر پاسخ مورد نظر نیست.

همان گونه که اشاره شد متغیرهای مدل عبارتنداز : یک متغیر پاسخ سه حالتی وضعیت تیروئید با مقادیر Y=۲ وضعیت تیروئید ۱B و بالاتر، Y =۱ وضعیت تیروئید ۱A 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 =۱ برای ساکنان شهری و x =۰ برای ساکنان روستایی. ابتدا مدل رگرسیون لجستیک با وجود متغیرهای جنسیت و محل سکونت روی داده های مورد مطالعه برازش شد و در سطح معنی داری α =۰/۰۵ کلیه متغیرها و اثر متقابل آنها در مدل باقی ماندند. نتایج در جدول ۱ خلاصه شده است . اعداد داخل جدول، برآورد ماکسیمم درست نمایی پارامترهای مدل در وضعیت های گوناگون به همراه خطای معیار برآوردها را نشان می دهد. اعداد ستونهای اول و دوم مربوط به برآورد پارامترها برای داده های کامل می باشد. مقایسهٔ برآورد های استاندارد و متداول نظیر SPSS با برآورد های برنامهٔ پیشنهادی برای مدل جدید که در محیط S-Plus و نرم افزار R نوشته شده است

نشان می دهد که برآوردها برای تمام پارامترها تقریباً یک سان بوده و با عنایت به این نکته که خطای معیار محاسبه شده توسط برنامه پیشنهادی از خطای معیار محاسبه شده برنامه های استاندارد به مراتب کمتر می باشد و این تأییدی بر درست بودن برآوردهای به دست آمده از مدل پیشنهادی برنامهٔ کامپیوتری نوشته شده می باشد. پس از حذف ۳۵٪ از داده های متغیر محل سکونت به صورت کاملاً تصادفی و اجرای مجدد هر دو برنامه روی داده های ناقص، خروجی های به دست آمده در ستون سوم و چهارم درج گردیده است. مقایسه مقادیر این دو ستون با یکدیگر و با ستونهای اول و دوم حاکی از آن است که برآوردهای حاصل از مدل پیشنهادی، نسبت به برآوردهای روشهای استاندارد که مبتنی بر حذف موارد گم شده می باشند نظیر برنامهٔ SPSS ، به مراتب به نتایج داده های کامل، نزدیکتر است. جهت بررسی بیشتر و ارزیابی دقیقتر مدل جدید با مدل استاندارد, تجزیه و تحلیل کاملتری صورت پذیرفت ابتدا برآوردهای ماکزیمم درست نمایی پارامترها را بر حسب مدل جدید و مــدل استاندارد (با استفاده از نرم افزار SPSS) با ده بار تکرار با درصدهای گم شدگی ۲۰٪ ، ۲۵٪ ، ۳۰٪ و ۳۵٪ انجام گردید. سپس برای تأیید دقیقتر ابتدا برآوردهای دو مدل را از برآوردهای داده های کامل کسر نموده و برای مقادیر حاصل تحلیل واریانس دو عاملی انجام شد. در این تحلیل متغیر وابسته برآورد پارامترهای مختلف مدل و عوامل مورد بررسی عبارت بودند از میزان درصدهای گم شدگی و نوع مدل. نتایج نشان داد که تنها عامل نوع مدل در سطح ۰/۰۰۱ معنی دار نشان داد. بدین معنا که در مجموع اختلاف برآورد پارامترها بین دو نوع مدل تفاوت دارد و برآوردهای مدل جدید به برآوردهای مدل با داده های کامل نزدیکترند. این نشانهٔ مثبتی از مناسب بودن مدل جدید برای تحلیل داده ها می باشد

بحث و نتیجه گیری :

ساتن و کارول در مقاله خود با استفاده از روش مذکور برای زمانی که متغیر پاسخ دو حالتی بود تحلیل را با متغیر کمکی بدون دادهٔ گم شده و با دادهٔ گم شده را با یکدیگر مقایسه نمودند و عملاً به این نتیجه رسیدند که مدل مذکور برای داده های گم شده در متغیر کمکی نسبت به روشهایی که مبنای کار آن اصولاً جانهی و یا حذف موردی که دارای مقادیر گم شده است بهتر عمل می کند. همان گونه که در جدول ۱ ملاحظه می شود به کارگیری مدل مذکور در حالتی که پاسخ سه وضعیتی بود عملاً منجر به برآوردهای با صحت بیشتر و واریانس کمتر شد. علاوه بر این با انجام آزمون کروسکال والیس، انحراف معیار پارامترهای با مدل جدید و مدل استاندارد اختلاف معنی دار داشتند و این خود تأییدی بر بالا بودن دقت برآوردهای مدل جدید نسبت به مدل استاندارد است.

تشکر و قدر دانی :

از جناب آقای دکتر مسعود کریملو که در تنظیم مقاله نظرات ارزشمندی ارائه نموده اند قدر دانی و سپاسگزاری می شود.

جدول ۱ - تجـزیه و تحلیـل داده های مـربوط به وضعیت تیروئید در استان قزوین*، مقایسه برآوردهای ماکزیمم درست نمایی با روش پیشنهادی و مدل استاندارد در حالت داده های کامل و حالتی که متغیر کمکی X ، ۳۵٪ داده های گم شده دارد

				برآورد پارامتر ها برای داده های کامل		برآورد پارامترها با ۳۵٪ گم شدگی تصادفی در متغیر محل سکونت	
لجیت[1]	متغیرها		پارامترها	مدل جدید	مدل استاندارد	مدل جدید	مدل استاندارد
	Intercept	عرض از مبدأ	β_{10}	-۰/۹۹۳	-۰/۹۹۳	-۱/۰۳۵	-۰/۶۹۳
				(۰/۳۶۸)[2]	(۰/۳۷۰)	(۰/۳۸۶)	(۰/۴۳۳)
	Sex (Z)	جنس (z)	β_{11}	۱/۴۶۳	۱/۴۶۳	۱/۵۰۲	۱/۰۹۹
۱				(۰/۶۱۷)	(۰/۶۸۰)	(۰/۶۵۸)	(۰/۷۷۷)
	Area (X)	محل سکونت	β_{12}	۱/۱۹۰	۱/۱۹۴	۱/۲۷۸	۱/۲۸۱
				(۰/۵۴۲)	(۰/۵۸۲)	(۰/۵۶۵)	(۰/۷۰۶)
	Area* Sex (x * z)	محل سکونت*جنس	β_{13}	-۲/۲۶۳	-۲/۲۷۰	-۲/۴۰۹	-۲/۳۸۰
				(۰/۸۳۵)	(۰/۹۶۰)	(۰/۹۱۱)	(۱/۱۳۶)
	Intercept	عرض از مبدأ	β_{20}	-۱/۹۱۰	-۱/۹۱۰	-۱/۸۱۰	-۱/۳۸۶
				(۰/۵۳۴)	(۰/۵۳۶)	(۰/۵۸۲)	(۰/۵۵۹)
	Sex (Z)	جنس (z)	β_{21}	۲/۴۹۷	۲/۴۹۷	۲/۲۹۰	۱/۷۹۲
				(۰/۷۱۹)	(۰/۷۷۳)	(۰/۷۸۳)	(۰/۸۵۴)
	Area (X)	محل سکونت	β_{22}	۲/۱۰۶	۲/۱۱۰	۲/۰۶۹	۱/۸۵۶
۲				(۰/۶۶۶)	(۰/۶۹۹)	(۰/۷۱۹)	(۰/۷۹۸)
	Area* Sex (x * z)	محل سکونت*جنس	β_{23}	-۲/۸۹۱	-۲/۸۹۹	-۲/۸۱۳	-۲/۵۴۹
				(۰/۸۸۱)	(۱/۰۰۱)	(۰/۹۸۱)	(۱/۱۶۰)

* از داده های طرح سلامت و بیماری سال ۱۳۸۰ ـ کل کشور

۱- منظور از لجیت ۱ و ۲ لگاریتم گرفتن از فرمولهای (۴) و (۵) یعنی لجیت ۱ برابر با $\ln\theta_1(x,z)$ و لجیت ۲ برابر با $\ln\theta_2(x,z)$ می باشد.

۲- اعداد داخل پرانتز خطای استاندارد است

منابع :

زالی, محمد رضا. محمد, کاظم. مسجدی, محمدرضا (۱۳۷۰). بررسی سلامت و بیماری در ایران ، معاونت پژوهشی وزارت بهداشت .

زالی, محمد رضا. محمد, کاظم. اعظم, کمال. مسجدی, محمد رضا (۱۳۷۳). وضعیت تیروئید درایران بر اساس نتایج طرح سلامت و بیماری ، مجله علمی نظام پزشکی، دوره سیزدهم از ۱۱۳ تا ۱۲۲.

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

نوربالا, احمد علی. محمد, کاظم (۱۳۸۱). بررسی سلامت و بیماری در ایران ، مقایسه استانی ۱۳۷۸ ، انتشارات مرکز ملی تحقیقات علوم پزشکی کشور

Armitage P. and Colton T. (1997) Encyclopedia of biostatistic, *John Wiley, New York.*

Blackhurst D.W. and Schluchter M. D. (1989), Logistic regression with a partially observed covariate, Comm. *Statist. Simul* , **18**(1): 163-177.

Fuchs C. (1982) Maximum likelihood estimation and model selection in contingency tables with missing data, *J. Amer. Statist. Assoc,* **77:** 270- 278.

Gao S. and Hui S.L. (1997) Logistic regression models with missing covariate value for complex survey data, *Statistics in Medicine.* **16:** 2419-2428.

Hosmer D.W. and Lemeshow Jr. S. (1989) Applied logistic regression , John Willey and Sons .

Kleinbaum D.G. and Klein M.(2002) Logistic Regression A Self – Learning Text , *Second Edition Springer.*

Little R.J.A. and Rubin D.B. (2002) Statistical analysis with missing data , John Wiley and Sons, Second Edition, New York.

Little R.J.A. and Schluchter M.D. (1985) Maximum likelihood estimation for mixed continuous and categorical data with missing values , *Biometrika* , **72:** 497- 512.

Paik M.C. and Sacco R.L. (2000) Matched case – control data analyses with missing covariates, *Applied Statistics* , **49:** 146-156.

Rathouz P.J. , Satten G.A. and Carrol R.J.(2003) Semiparametric inference in matched case – control studies with missing covariate data , *Biometrika* .

Satten G.A. and Carroll R.J. (2000) Conditional and unconditional categorical regression models with missing covariates , *Biometrics*, **56:** 384-388.

Satten G.A. and Kupper L. (1993a) Inferences about exposure – disease associations using probability of exposure information , *J. Amer. Statist. Assoc* , **88:** 200-208.

Satten G.A. and Kupper L. (1993b) Conditional regression analysis of the odds ratio between two binary variables when one is not measured with certainty, A method for epidemiologic studies, *Biometrics*, **44:** 429 – 440.

Stuart R.L., Michael P. and Marium E. (1998) Inference using conditional logistic egression with missing covariates, *Biometrics,* **54:** 295 – 303.

MULTINOMIAL LOGISTIC REGRESSION MODEL WITH MISSING DATA AND ITS APPLICATION TO GOITER DISEASE DATA

Azam K.,*[1] MSD; Gerami A.,[2] Ph.D; Mohammad K.,[3] Ph.D; Kazemnejad A.,[1] Ph.D

In large–scale sampling opeartions (e.g. nation-wide health surveys) we always face the problem of non-response item(s) and/or non-response unit(s). In fitting a model to the data we have two groups of variables, namely dependent and independent variables. Non-response may occur for any of these groups of variables. In this paper we assume *Y* as a categorical dependent variable with three levels, *Z and X* as independent variables from any kind: scale, categorical, ordinal, etc. We have complete data on the first two variables and we assume that the missing items follow a random pattern (*MAR*). Then a model is devised based on the likelihood function for the whole data set (including missing values) and estimates of parameters are compared with those obtained by statistical programs such as *SPSS*, which are only based on completely observed data and ignore units with missing data. Our results show that the likelihood-based model is superior to the standard approach utilized by the software packages. The comparison is made using data on thyriod disease (goiter) obtained by a health survey in Gazvin province.

Key words: *Missing At Random, Logistic Regression, Goiter Disease, Maximum Likelihood, Polytomous Outcome, Thyroid*

*. Author to whom all correspondence should be addressed.
1. Department of Biostatistices, Tarbiat Modarress University, Iran.
2. Statistical Research Center.
3.Department of Epidemiology and Biostatistices, School of Public Health and Institute of Public Health
Research Terhan University of Medical Sciences.