مجله اپيدميولوژي ايران
دوره ۱ شماره ۱ پاييز ۱۳۸٤ صفحات ۷۲ - ۶۵

تحليل درستنمايي ماكزيمم مدل رگرسيون لجستيک در حالتي كه داده هاي متغيرهای پيشگو كامل نيستند ولي متغيرهای كمكي وجود دارند

محمد امين پورحسينقلي : دانشجوی کارشناسی ارشد آمار زيستی, گروه آمار زيستی دانشکده پيراپزشکی , دانشگاه علوم پزشکی شهيدبهشتی : نويسنده رابط amin_phg@yahoo.com
دكتر حميد علوي مجد : استاديار, گروه آمار زيستی , دانشکده پيراپزشکی , دانشگاه علوم پزشکی شهيدبهشتی
دكتر عليرضا ابدي : استاديار گروه پزشکی اجتماعی و بهداشت , دانشکده پزشکی , دانشگاه علوم پزشکی شهيدبهشتی
سيمين پروانه وار: دانش آموخته کارشناسی ارشد مامايی , دانشگاه علوم پزشکی شهيدبهشتی

دريافت : ۸۴/۴/۵ پذيرش : ٨٤/٥/٢١

چكيده :

مقدمه و هـدف: داده هاي گمشده در بسياري از مطالعات آماري از جمله مدلهاي رگرسيوني وجود دارند و باعث كاهش دقت برآورد مي شوند. تا كنون روشهاي گوناگوني براي مقابله با مشكل داده هاي گمشده ابداع شده كه عموما بر داده هاي گمـشدة متغيـر پاسـخ متمركز بوده است حال آنكه متغيرهاي پيشگو نيز مي توانند دستخوش تغيير و از دست رفتن اطلاعات شوند.

مواد و روشها: در اين تحقيق ضمن بررسي روش جانهي داده هاي گمشده با استفاده از الگوريتم EM و متغير كمكي, نتايج حاصل از اين روش را با روش تحليل مورد كامل در يک مدل رگرسيون لجـستيک پيرامـون عوامـل مـؤثر بـر انتخـاب نـوع زايمـان مقايـسه مي كنيم.

يافته ها: داده هاي مورد استفاده در اين مقاله از يک مطالعه توصيفي پيرامون عوامل مرتبط با انتخـاب نـوع زايمـان در زنـان مراجعـه كننده به مراكز بهداشتي و درماني شهر تهران بدست آمده است. حجم نمونه دراين تحقيق ۳۸۵ نفر بوده و از روش نمونه گيري چند مرحله اي انتخاب شدندو مشخصات فردي، سوابق مامايي، نوع نگرش و عوامل اجتماعي نمونه ها از طريـق پرسـشنامه ثبـت شـدند. براي مقايسه ميزان كارايي دو روش، برآورد انحراف معيار پارامترها مورد استناد قرار گرفت.

بحث و نتيجه گيری : نتايج حاصل نشان مي دهد روش تحليل درستنمايي با الگوريتم EM در مقايسه با روش مـورد كامـل كـارايي بهتري دارد. مشكل داده هاي گمشده در بسياري از مطالعات آماري وجود دارد و موجب اريبي و كاهش كارايي مـي شـوند. در ايـن بررسي نشان داده ايم استفاده از الگوريتم EM براي جانهي گمشده هادر يک مدل رگرسيون لجستيک با متغيرهاي توضـيحي گسـسته و سپس تحليل مدل، از روش مورد كامل كه مستلزم حذف گمشده ها به همراه قسمتهايي از اطلاعات است كـاراتر اسـت. از سـوي ديگر اگر متغير توضيحي ناكامل پيوسته باشد بدست آوردن مدل، روشي متفاوت مي طلبد و يا مي توان با تبديل آن به متغيري گسسته از روش قبل استفاده كرد.

كليد واژه ها : مدل رگرسيون لجستيك، الگوريتمEM ، تحليل مورد كامل، داده گمشده، متغير كمكي، سزارين

مقدمه

در بسياري از تحقيقات پزشكي با متغيرهايي مواجه مي شويم كه قسمتي از اطلاعاتشان به دلايل مختلفي چون عدم پاسخ، ناكامل بودن چارچوب بررسي، ازدست رفتن اطلاعات موجود در پرونده ها و ... از دست رفته اند. به اين گونه داده ها در آمار داده گمشده گويند و در صورت ناديده گرفتن آنها استنباط آماري را دچار مشكل مي كنند و بخصوص در برآورد به روش درستنمايي، ايجاد اريبي كرده و كارايي را كاهش مي دهند(۳,۲,۱). روشهاي گوناگوني براي مقابله با مشكل داده هاي گمشده ابداع شده كه بر اساس آنها گمشده ها جانهي مي شوند. يكي از اين روشها استفاده از الگوريتم EM است(۴). الگوريتم EM يك روش تكرار شونده است كه در هر تكرار دو گام را شامل مي شود. گام E(گام اميد رياضي) كه داده گمشده به شرط داده هاي مشاهده شده محاسبه مي شوند اين

اميدهاي رياضي را به جاي داده هاي گمشده قرار مي دهند و پارامترهاي مورد نظر برآورد مي شوند. در گام بعدي يعني گام M (گام ماكزيمم كردن) بعد از جايگذاري اعداد اوليه بجاي داده هاي گمشده به شرط داده هاي مشاهده شده لگاريتم تابع درستنمايي را حداكثر مي كنيم. اين مكانيسم آنقدر تكرار مي شود تا به همگرايي ميان پارامترهاي برآورد شده در تكرارها برسيم.

با اين حال بيشترين روشهاي جانهي متوجه داده هاي گمشده در متغير پاسخ بوده است حال آنكه متغيرهاي پيشگو نيز مي توانند دستخوش تغيير و گمشدگي شوند. از اينرو داده هاي گمشده در متغيرهاي پيشگو نيز مورد توجه برخي آمار دانان قرار گرفته است.

ابراهيم روش درستنمايي ماكزيمم را براي مدلهاي عمومي رگرسيوني پيشنهاد مي كندكه با بهره گيري از الگوريتم EM گمشده هاي متغيرهاي مستقل گسسته برآورد مي شوند(۵).

در اكثرمطالعات آماري معمولا متغيرهايي بيش از آنچه براي طراحي مدل مورد نياز است در مورد آزمودني ها جمع آوري مي شوند. به اين متغيرها كه مقادير آنها جمع آوري و ثبت شده ولي به عنوان يك متغير پيشگو در مدل استفاده نشده اند متغيرهاي كمكي گويند. در حالي كه متغير پيشگو كامل نيست اين متغيرها ممكن است كاملا مشاهده شده باشند. استفاده از متغيرهاي كمكي كه داده هايشان كامل است در حالتي كه متتغيرهاي پيشگو كامل نيستند مي تواند موجب افزايش كارايي آناليز مدل شود.

در اين تحقيق، برآورد درستنمايي را در مدل رگرسيون لجستيك بررسي مي كنيم كه متغيرهاي مستقل گسسته آن ناكامل اند ولي متغيركمكي گسسته آن كامل است(۶)و با استفاده از داده هاي تحقيقي كه درباره علل تمايل زنان باردار به نوع زايمان است (۷) كارايي اين روش را براساس انحراف معيار برآوردها با روش تحليل مورد كامل مقايسه مي كنيم.

۱- تحليل مورد كامل: در اين روش همه واحدهايي كه داراي مقادير گمشده اند كنار گذاشته مي شوند. شايد اين متداول ترين روش براي حل مشكل داده هاي گمشده باشد ولي در كل شيوه خوبي تلقي نمي شود زيرا با حذف همه واحدهاي داراي داده هاي گمشده، اندازه نمونه به صورتي قابل ملاحظه كوچك شده و منجر به كاهش دقت مي شود. همچنين اگر واحدهايي كه از تحليل حذف مي شوند با آنهايي كه باقي مي مانند تفاوت زيادي داشته باشند، ممكن است برآوردهاي حاصل به شدت اريب شوند(۸,۲).

۲- تحليل درستنمايي با استفاده از الگوريتم EM و متغير كمكي

۲-۱- الگوريتم EM : يكي از روش هايي كه براي برآورد داده هاي گمشده ابداع شده الگوريتم EM است.

الگوريتم EM يك روش محاسباتي عمومي براي برآوردهاي حداكثر درستنمايي تحت داده هاي ناكامل است. تاريخچة EM به سال ۱۹۲۶ و مقاله ايي از مك كندريك باز مي گردد ولي اولين بار بوسيله دمپستر و همكاران ابداع شد(۴). نام اين الگوريتم بر گرفته از دو گام اين الگوريتم است: گام (E) كه محاسبة مقادير مورد انتظار براي گمشده هاست و گام (M) كه محاسبة برآوردهاي حداكثر درستنمايي پارامترها با فرض كامل بودن داده هاست كه براساس دو مرحلة پايه گذاري شد.

الف) اگر ما مقادير گمشده را بدانيم مي توانيم پارامترها را برآورد كنيم.

ب) اگر پارامترها را بدانيم مي توانيم مقادير گمشده را با مقادير مورد انتظار جايگذاري كنيم.

۲-۲- مدل رگرسيون لجستيك با داده هاي كمكي

براي مشاهدات دو حالتي $y_1, y_2, \dots, y_n$ و ماتريس x كه مجموعه ايي از متغيرهاي گسسته مستقل است مدل لجستيك زير را داريم.

$$\log it\left(E\left[Y_i \middle| x_i\right]\right) = x_i^1 \beta \quad (۱)$$

كه بعد β ، $P \times 1$ است. حداكثر درستنمايي براساس احتمال شرطي $f\left(Y \middle| X, \beta\right)$ بدست مي آيد ولي اگر x دچار داده هاي گمشده باشد روش ML مؤثر نيست. ابراهيم روش درستنمايي ماكزيمم را پيشنهاد مي كند كه با استفاده از الگوريتم EM انجام شده و بجاي مدل سازي احتمال Y به شرط X از احتمال توأم آنها نيز استفاده مي كند(۵).

$$f(Y, X \mid \Omega) = f\left(Y \middle| X, \beta\right) f\left(X \middle| \gamma\right) \quad (۲)$$

که $\Omega = (\beta, \gamma)$

زير مجموعه ايي از اين متغيرها را وارد مدل مي کنند و عموما تعدادي از متغيرها علي رغم گردآوري اطلاعاتشان، در مدل استفاده نمي شوند. به اين گونه متغيرها، متغيرهاي کمکي گويند(۶).

اگر متغيرکمکي A موجود باشد مي توانيم تابع درستنمايي داده هاي کامل را به صورت زير بنويسيم:

اگر x به طور کامل مشاهده شده باشد $f(X;Y)$ دردرستنمايي براي β شرکت نمي کند و اگر داده گمشده در x باشد برآورد β ممکن است دچار اريبي و کاهش کارايي شود. از سوي ديگر معمولا محققين مجموعه ايي از متغيرهاي مستقل را در شروع تحقيق بررسي و اطلاعات آنها را ثبت مي کنند ولي فقط

$$f(Y, X, A|\Omega^*) = f(Y|X, A, \beta^*) f(X, A|\gamma^*) \quad (۳)$$

که $\Omega^* = (\beta^*, \gamma^*)$

و* را براي نشان دادن وابستگي به توزيع A بکار مي بريم.

اگر فرض کنيم

$f(Y|X, A, \beta^*) = f(Y|X, \beta)$ (يعني استقلال شرطي y و A تحت x) روش ابراهيم را براي مدل مي توان بکار برد.

واچ اين برآوردگر را وقتي متغير کمکي براي بهبود بخشيدن به مدل استفاده مي شود پيشنهاد کرده است و

زماني بکار مي رود که استقلال شرطي برقرار باشد(۹). اما استقلال شرطي هميشه برقرار نيست. اگر استقلال شرطي برقرار نباشد فاکتور بندي رابطهٔ (۳) طبيعي نيست زيرا ضرايب رگرسيوني براي X در $E(Y|X, A)$ عموماً برابر β در مدل (۱) نيستند.

روش پيشنهاد شده ديگري که به وسيله واچ براي داده هاي کامل بکار رفته بصورت زير است(۱۰):

$$f(Y, X, A|\theta) = f(A|Y, X, \alpha) f(Y|X, \beta) f(X|\gamma) \quad (٤)$$

$\theta = (\alpha, \beta, \gamma)$ مجموعه اين از پارامترهاي درستنمايي لگاريتمي تابع

$$\sum_i l_{a,y,x}(\theta|a_i, y_i, x_i) = \sum_i \{ l_{a|y,x}(\alpha|a_i, y_i, x_i) + l_{y|x}(\beta|y_i, x_i) + l_x(\gamma|x_i) \} \quad (۵)$$

است.

۲-۳- برآورد درستنمايي ماکزيمم تحت متغيرهاي مستقل ناکامل و اطلاعات کمکي

اگر داده ها کامل باشند $\prod f(Y_i|X_i, \beta)$ براي برآورد β براحتي ماکزيمم مي شود. در اين حالت β جداي از α و γ با رگرسيون لجستيك

برآورد مي شود. اما اگر Xi هاي گمشده باشند هر سه قسمت درستنمايي بايد برآورد شوند. با فرض اينکه گمشده ها به صورت تصادفي حادث شده اند درستنمايي لگاريتمي زير را داريم(۶):

$$L^0_{a,y,x}(\theta) = \sum_i \log \sum_{xmissx} \{ L_{a|y,x}(\alpha|a_i, y_i, x_i) \times L_{y|x}(\beta|y_i, x_i) L_x(\gamma|x_i) \}$$

که $X_{miss,i}$ نشان دهندهٔ قسمتهايي از x_i است که گمشده اند و $\sum_{xmiss,i}$ مجموع مکان نمونه ايي $X_{miss,i}$ است.

بوسيلهٔ اين تابع براوردهاي براي β بدست مي آوريم که تحت متغير کمکي و بدون فرض استقلال شرطي ماکزيمم شده است.

بر اساس روش ابراهيم (۵) متغيرهاي پيشگو $X = (x_1, x_2, ..., x_p)$ متغيرهاي گسسته تصادفي با توزيع چندگانه تحت پارامتر

$\gamma=(\gamma_1,\gamma_2,\dots,\gamma_r)$ هستند. مثلاً اگر سه متغير دو حالتي وجود داشته باشد بعد γ برابر است با $r = 2^3 - 1 = 7$. در حالت كلي فرض كنيد $c_1,\dots,c_p$ نشان دهنده تعداد رده ها براي هر متغير باشد. در نتيجه γ داراي بعد $r = c_1\times\dots\times c_p - 1$ است.

فرض كنيد $f(Y|X,\beta)$ چگالي متغير پاسخ و $f(A|X,Y,\alpha)$ و $f(X|Y)$ توزيع هاي چند جمله ايي با پارامترهاي γ,α (مجزا از β) باشند. اگر هيچ محدوديتي به توزيع X تحميل نشود تحت داده هاي كامل، γ بوسيلة $r+1$ خانه شمارشي مشاهده شده از رده هاي x در يك جدول توافقي p بعدي بدست مي آيد. به طور مشابه مدل چند جمله ايي براي برآورد $f(A|X,Y,\alpha)$ مي تواند مورد استفاده قرار گيرد و رگرسيون لجستيك را وقتي X كاملا مشاهده شده است براي برآورد β بكار رود.

اما در حالتي كه X ناكامل است از الگوريتم EM (۴)براي برآورد پارامترها در درستنمايي داده هاي مشاهده شده استفاده مي كنيم. اين الگوريتم تكرار شونده شامل دو مرحله است. در مرحله E اميد $L_{a,y,x}(\theta|a,Y,X)$ براساس پارامترهاي $\theta^{(t)}$ و داده هاي مشاهده شده محاسبه مي شوند.

در اين وضعيت، داده هاي مشاهده شده شامل $(y_i,x_{obs,i},a_i)$ اند كه $X_{obs,i}$ نشان دهنده مجموعه ايي با بعد كوچكتر يا مساوي p متغير مستقل در X_i است و اگر i امين عضو مشاهده شده در متغيرهاي مستقل باشد معادل X_i خواهد بود.

اين اميد با $Q(\theta|\theta^{(t)})$ نشان داده مي شود و داريم(۶):

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{r+1} w_{ij}^{(t)} L_{a,y,x}(\theta|a_i,y_i,x^j)$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{r+1} w_{ij}^{(t)}\left\{l_{a|y,x}(\alpha|a_i,y_i,x^j) + l_{y|x}(\beta|y_i,x^j) + L_x(\gamma|x^j)\right\} \quad (۶)$$

كه x^j، j امين نمونه ممكن از متغيرهاي مستقل، $L_{a,y,x}(\theta|a_i,y_i,x_i)$ درستنمايي لگاريتمي داده هاي كامل براي θ با i امين مشاهده تحت X_i ارزيابي شده با x^j است و $w_{ij}^{(t)} = p(x^j|a_i,y_i,x_{obs,i},\theta^{(t)})$ ميتواند وزني از j امين نمونه براي i امين مشاهده در t امين تكرار باشد. در نتيجه $W_{it}^{(t)}=1$. البته بيشتر اين w_{ij} ها صفر خواهند شد.

پس هر مقدار $X_{obs,i}$ تحت x^j مناسب تشخيص داده نشود مورد استفاده قرار نمي گيرد. (ستون x^j و $X_{obs,i}$ زماني متناسب اند كه اعضاي متناظر با ستون x^j برابر داده هاي مشاهده شده $X_{obs,i}$ باشد).

اگر همه متغيرهاي مستقل براي iامين عضو، مشاهده شده باشند فقط يك قسمت مخالف صفر وجود خواهد داشت و وزنها براي مقدار مشاهده شده $x^j = x_{obs,i}$ برابر يك خواهد بود.

وزنها را مي توان از قانون بيز به صورت زير محاسبه كرد(۶):

$$W_{ij}^{(t)} = P\left(x^{j} \middle| a_i, y_i, x_i, \theta^{(t)}\right)$$

$$= \begin{cases} 0 & if\, x^{j}\ is\ \ not\ \ compatibeal\ with\ \ \ x_i \\ \dfrac{p\left(y_i \middle| x_i^{j}\right) P\left(a_i \middle| x_i^{j}, y_i\right) p\left(x_i^{j}\right)}{\sum_{k \in obs,i} p\left(y_i \middle| x_i^{k}\right) p\left(a_i \middle| x_i^{k}, y_i\right) p\left(x_i^{k}\right)} & \\ & if\, x^{j}\ is\ \ \ \ \ compatibeal\ with\ \ \ x_i \end{cases} \quad (٧)$$

كه x^{j} نشان دهنده يـك بـردار p تـايي اسـت كـه اجزاء مشاهده شده اش $X_{obs,i}$ و اجزاء باقي مانـده آن بر اساس j امـين نمونـه بـراي متغيرهـاي گمـشده است.

براي مرحله M معادلة (۶) را به عنـوان يـك تابع از θ و بوسيله يافتن راه حل معادلـه درسـتنمايي لگـاريتمي داده هـاي كامـل بـا اسـتفاده از ايـن وزنهـا ماكزيمم مـي كنـيم. اكنـون مـي تـوانيم از رگرسـيون لجستيك وزن داده شده براي برآورد β استفاده كنيم و براي برآورد γ, α نيز از خانه هاي شمارش شـده مورد انتظار بهره مي گيريم.

براي اولين تكرار، وزنها را مي توان از رابطه (۷) و نخستين برآوردها بـراي θ را بـا فـرض كامـل بودن داده ها بدست آورد. اين مراحـل آنقـدر تكـرار مي شوند كه به برآوردهاي درستنمايي ماكزيمم همگرا شوند و اختلاف ميان پارامترها در تكرارهاي بعد معني دار نباشد.

روش كار

براي مقايسه كارايي دو روش تحليل مورد كامل و تحليل درستنمايي با الگوريتم EM از اطلاعات تحقيقي استفاده كرده ايم كه در مورد عوامل مرتبط با انتخاب نوع زايمان در زنان مراجعه كننده به مراكز بهداشتي و درماني شهر تهران (۷) است. اين تحقيق يک مطالعه توصيفي بوده است و در آن نمونه اي به حجـــم ۳۸۵ نفر از مادران باردار با سن حاملگي ۲۸ هفته و بالاتر از مراكز بهداشتي و درماني شهر تهران انتخاب شدند. روش نمونـــــه گيري از نوع چنــد مرحله اي بوده و مشخصات فردي، سوابق مامايي، نوع نگرش و عوامل اجتماعي نمونه ها از طريق پرسشنامه ثبت شدند. متغير پاسخ در اين تحقيق نوع زايمان انتخاب شده بوسيله زنان باردار است كه به دو سطح زايمان طبيعي و سزارين تقسيم شده است. متغيرهاي پيشگويي كه براي ساختن مدل انتخاب شده اند شش متغير دوحالتي هستند شامل شغل آزمودني، شركت در كلاسهاي آموزشي دوران بارداري، نوع حاملگي، نگرش زنان باردار به زايمان طبيعی , داشتن سابقه تولد نوزاد باوزن بيشتر از ۴۰۰۰ گرم و داشتن سابقه تولد نوزاد باوزن كمتر از ۲۵۰۰ گرم (جدول یک).

در متغير داشتن سابقه تولد نوزاد باوزن كمتر از ۲۵۰۰، ۲۵ درصد از داده ها به عنوان داده گمـشده بـه صورت تصادفي حذف شـدند. همچنـين داشـتن سـابقه زايمان پـس از موعـد نيـز بـه همـراه سـاير متغيرهـا در پرسشنامه ثبت شده و اطلاعات آن كامل بود كه در مـدل وارد نشده و به عنوان يك متغير كمكي مورد استفاده قرار گرفت (اين متغير نيز دو حالتی است).

ابتدا با روش تحليـل مـورد كامـل, قـسمتهاي گمـشده از مطالعـه حـذف و پارامترهـا بـرآورد شـدند. درمرحله بعد, از برآورد درستنمايي ماكزيمم تحت متغير كمكي براي برآورد پارامترهاي مدل استفاده كرديم. يـك برنامه كامپيوتري تحت نرم افزارs-plus براي اين منظور مورد استفاده قرار گرفته و برای مقايسه ميزان كـارايي دو روش برآورد انحراف معيار پارامترها مـورد اسـتناد قـرار گرفت كه انحراف معيارها در روش تحليل درستنمايي بـا الگوريتم EM از روش لوئيس (۱۱) برآورد شدند.

یافته ها

نتایج حاصل از برآورد انحراف معیارهاي دو روش تحلیل مورد کامـل و جـانهي بـا الگـوریتم EM نشان مي دهد که انحراف معیارهاي برآورد شده بـراي تمام پارامترهادر روش جانهي با الگوریتم EM کمتر از انحراف معیارهایي هـستند کـه در روش تحلیـل مـورد کامل بدست آمده است که نشان ميدهد کارایي بـرآورد پارامترها در این روش بهتر از روش تحلیل مورد کامل است و علي رغم اینکـه هـیچ قـسمت از اطلاعـات از مدل حذف نميشوند پارامترهایي با انحـراف معیارهـاي پایینتر برآورد ميـشوند و درنتیجـه فاصـله اطمینانهـاي کوتاهتري براي پارامترها بدست مي آیـد(جـدول دو). اگـر چـه اختلافـات قابـل تـوجهي نیـز در بـرآورد پارامترهاي لجستیک در دو روش مشاهده مي شود کـه احتمالا ناشـي از اریبـي حاصـل از حـذف قـسمتي از اطلاعات در روش تحلیل مورد کامل است(۲).

بحث

مـشکل داده هـاي گمـشده در بـسیاري از مطالعات آماري وجود دارد و اگر بدون توجه به وجود گمشده ها، مدل رابرازش دهیم و ضـرایب رگرسـیوني را با روش درستنمایي ماکزیمم برآورد کنیم گمشده هـا بر برآورد ها اثر منفي گذاشته و موجب اریبي و کاهش کارایي مي شـوند(۳,۲,۱). در ایـن تحقیـق روشـي را بررسي کردیم که با بهره گیري از متغیري کمکي آنـالیز درستنمایي ماکزیمم را در مدل لجستیك انجام داده و با اسـتفاده از الگـوریتم EM بـرآورد پارامترهـا را در حالتیکه متغیر پاسخ کامـل نیـست محاسـبه مـي کنـد. نتایج حاصل از بررسي این روش در مـدل مربـوط بـه عوامل موثر بر انتخاب نوع زایمان نشان داد که بـرآورد انحراف معیار پارامترها در ایـن روش از تحلیـل مـورد کامل کمتر است که این مساله با یافته هاي مدل کاربردي هورتون و لایرد تطابق دارد(۶).

وجود متغیري کمکي که اطلاعات آن بـه طـور کامل موجود است ولي مـستقیما بـه عنـوان یـك متغیـر پیشگو در مدل مورد استفاده قـرار نگرفتـه میتوانـد بـراي افزایش کارایي مدل و برآورد دقیقتر پارامترها مفید باشـد. انتخاب متغیر کمکـي بـستگي بـه اطلاعـات موجـود در تحقیق و نظر محقق دارد ولي هرچقدر ارتباط میان متغیر کمکي و متغیري که دچار گمشدگي شده بیشتر باشد این روش مؤثرتر خواهد بـود(۶). از طـرف دیگـر بـرآورد انحراف معیار پارامترهایي که بـا EM بدسـت آمـده انـد مشکل است. لوئیس روشي براي برآورد انحراف معیار پیشنهاد کرده که در این تحقیق مورد استفاده قـرار گرفتـه است(۱۱). همچنین از روش بوت استراپ نیز مي تـوان براي برآورد انحراف معیارها استفاده کرد.

توزیعهـاي $f(A|X,Y,\alpha), f(X;\delta)$ از طریق برآورد پارامترهاي توزیع چند جملـه ایـي بـرآورد مي شوندهرچند گمشده ها مي توانند تاثیر نـامطلوبي بـر برآوردداشته باشند.

از سـوي دیگـر اگـرY متغیـري پیوسـته باشـد بدسـت آوردن مـدلي بـراي $f(A|Y,X)$ روشـي متفاوت مي طلبد یا مي توان با تبدیل کردن Y به متغیري گسسته از روش قبل استفاده کرد.

توسعه این روش به سایر مـدلهاي رگرسـیوني قابل بررسي است و برخي از این مدلها – بدون اسـتفاده از متغیر کمکي – بوسیله ابراهیم و همکاران مورد توجـه قـرار گرفتـه اسـت. همچنـین توجـه بـه نـوع مکانیـسم گمشدگي (۱۲) در استفاده از روش جـانهي بـا الگـوریتم EM مهم اسـت زیـرا ایـن روش ماننـد اکثـر روشـهاي جانهي (۱۳,۲) صرفا در مکانیسمهاي گمشدگي تـصادفي مناسب عمل مي کند.

جدول ۱: متغیرهاي توضیحي مدل

متغیر	رده ها
شغل آزمودني	شاغل=۱ ، خانه دار=۰
شركت در كلاسهاي آموزشي دوران بارداري	شركت =۱ ، عدم شركت= ۰
نوع حاملگي	خواسته=۱ ، ناخواسته=۰
نگرش زنان باردار به زايمان طبيعی	نگرش مثبت = ۰
	نگرش منفي = ۱
سابقه تولد نوزاد باوزن بيشتر از ۴۰۰۰ گرم	سابقه دارد=۱ ، سابقه ندارد=۰
سابقه تولد نوزاد باوزن كمتر از ۲۵۰۰ گرم	سابقه دارد=۱ ، سابقه ندارد=۰

جدول ۲- برآورد پارامترها و انحراف معيار آنها از دو روش درستنمايي الگوريتم EM و تحليل مورد كامل

	روش درستنمايي با الگوريتم EM و متغير كمكي		روش تحليل مورد كامل	
پارامترهاي مدل	انحراف معيار	برآورد ضريب رگرسيوني	انحراف معيار	برآورد ضريب رگرسيوني
عرض از مبدا	۰/۲۷۲۸	-۱/۳۱۷۲	۰/۴۶۲	-۲/۲۴۲
شغل آزمودنی	۰/۴۰۰۶	-۱/۲۴۹۶	۰/۹۳۵	۰/۹۸۵
شركت در كلاسهاي آموزشي دوران بارداري	۰/۲۶۸۳	-۰/۵۳۳۹	۰/۴۵۵	۰/۱۶۶
نوع حاملگي	۰/۲۲۷	-۰/۳۲۲۱	۰/۳۲۱	-۰/۰۸۳
نگرش زنان باردار به زايمان طبيعی	۰/۲۸۶۴	-۰/۵۷۹۴	۰/۳۵۴	۰/۴۶۰
سابقه تولد نوزاد باوزن بيشتر از ۴۰۰۰ گرم	۰/۲۸۴۳	۲/۴۸۳۸	۰/۳۸۳	۲/۵۳۱
سابقه تولد نوزاد باوزن كمتر از ۲۵۰۰ گرم	۰/۳۸۶۷	-۰/۳۴۸۸	۰/۸۱۳	۰/۲۱۰

References

۱- پروانه وار سیمین. بررسي عوامل مرتبط با انتخاب نوع زايمان در زنان باردار مراجعه كننده به مراكز بهداشتي و درماني شهر تهران سال ۱۳۸۲. پايان نامه كارشناسي ارشد رشته مامايي, دانشگاه علوم پزشكي شهيد بهشتي ۱۳۸۲

2- Little, R. J. Biostatistical Analysis with Missing Data. In *the Encyclopedia of Biostatistics*, Armitage, P. A. and Colton, T. , Eds., Wiley, Chichester U. K. , 1998

3- Levy, P. S. and Lemeshow, S. Sampling of Populations Methods and Applications. Third Edition John Wiley and Sons 1999;393-416.

4- Little, R. J. and Rubin, D. B. Statistical Analysis with Missing Data. Newyork: John Wiley and Sons.1987 .

5- Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum, likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society,1977; Series B39: 1–22.

6- Ibrahim, J. G. Incomplete data in generalized linear, models. Journal of the American Statistical Association, 1990, 85, 765–769.

7- Horton. N. J. and Laird. N. M. (2001) Maximum Likelihood Analysis of Logistic Regression Models with Incomplete Covariate Data and Auxiliary Information. Biometrics 2001, 57, 34–42.

8- Glynn, R. J. and Laird N. M. Regression Estimates and Missing Data: Complete Case Analysis. Unpublished Manuscript, Department of Biostatistics, Harvard University 1983.

9- Vach, W. Some Issues in Estimating the Effect of Prognostic Factors from Incomplete Covariate Data. Statistics in Medicine 1997 16, 57-72.

10- Vach, W. Logistic Regression with Missing Values in the Covariates. Berlin: Springer-Verlag 1994.

11- Louis, T. A. Finding the Observed Information Matrix When Using the EM Algorithm. Journal of the Royal Statistical Society, 1982 Serios B 44, 226-233.

12- Rubin, D. B. Inference and Missing Data. Biometrika 1976 63, 581-592.

13- Saleh A. M. Some Methods for Dealing with Missing Data in Sample Surveys. Invited Papers Proceedings of the 7th Iranian Statistical Conference, 2004, 313-324