مجله اپیدمیولوژی ایران
دوره ۱ شماره ۲ و ۱ پاییز ۱۳۸۴ صفحات ۷۲ - ۶۵

تحلیل درستنمایی ماکزیمم مدل رگرسیون لجستیک در حالتی که داده های متغیرهای پیشگو کامل نیستند ولی متغیرهای کمکی وجود دارند

محمد امین پورحسینقلی : دانشجوی کارشناسی ارشد آمار زیستی، گروه آمار زیستی دانشکده پیراپزشکی ، دانشگاه علوم پزشکی شهیدبهشتی : نویسنده رابط amin_phg@yahoo.com
دکتر حمید علوی مجد : استادیار، گروه آمار زیستی ، دانشکده پیراپزشکی ، دانشگاه علوم پزشکی شهیدبهشتی
دکتر علیرضا ابدی : استادیار گروه پزشکی اجتماعی و بهداشت ، دانشکده پزشکی ، دانشگاه علوم پزشکی شهیدبهشتی
سیمین پروانه وار: دانش آموخته کارشناسی ارشد مامایی ، دانشگاه علوم پزشکی شهیدبهشتی

دریافت : ۸۴/۴/۵ پذیرش : ۸۴/۵/۲۱

چکیده :

مقدمه و هدف: داده های گمشده در بسیاری از مطالعات آماری از جمله مدلهای رگرسیونی وجود دارند و باعث کاهش دقت برآورد می شوند. تا کنون روشهای گوناگونی برای مقابله با مشکل داده های گمشده ابداع شده که عموما بر داده های گمشدهٔ متغیر پاسخ متمرکز بوده است حال آنکه متغیرهای پیشگو نیز می توانند دستخوش تغییر و از دست رفتن اطلاعات شوند.

مواد و روشها: در این تحقیق ضمن بررسِی روش جانهی داده هاِی گمشده با استفاده از الگوریتم EM و متغیر کمکی، نتایج حاصل از این روش را با روش تحلیل مورد کامل در یک مدل رگرسیون لجستیک پیرامون عوامل مؤثر بر انتخاب نوع زایمان مقایسه می کنیم.

یافته ها: داده هاِی مورد استفاده در این مقاله از یک مطالعه توصیفِی پیرامون عوامل مرتبط با انتخاب نوع زایمان در زنان مراجعه کننده به مراکز بهداشتِی و درمانی شهر تهران بدست آمده است. حجم نمونه دراین تحقیق ۳۸۵ نفر بوده و از روش نمونه گِیرِی چند مرحله اِی انتخاب شدندو مشخَصات فردِی، سوابق مامایی، نوع نگرش و عوامل اجتماعِی نمونه ها از طرِیق پرسشنامه ثبت شدند. برای مقایسه میزان کارایی دو روش، برآورد انحراف معیار پارامترها مورد استناد قرار گرفت.

بحث و نتیجه گیری : نتایج حاصل نشان می دهد روش تحلیل درستنمایی با الگوریتم EM در مقایسه با روش مورد کامل کارایی بهترِی دارد. مشکل داده هاِی گمشده در بسیاری از مطالعات آماری وجود دارد و موجب اریبی و کاهش کارایی می شوند. در این بررسِی نشان داده ایم استفاده از الگوریتم ĒM برای جانهی گمشَده هادر یک مدل رگرسیون لجستیک با متغیرهاِی توضیحِی گسسته و سپس تحلیل مدل، از روش مورد کامل که مستلزم حذَف گمشده ها به همراه قسمتهایی از اطلاعات است کاراتر است. از سوی دیگر اگر متغیر توضیحِی ناکامل پیوسته باشد بدست آوردن مدل، روشی متفاوت می طلبد و یا می توان با تبدیل آن به متغیری گسسته از روش قبل استفاده کرد.

کلید واژه ها : مدل رگرسیون لجستیک، الگوریتمEM ، تحلیل مورد کامل، داده گمشده، متغیر کمکی، سزارین

مقدمه

در بسیاری از تحقیقات پزشکی با متغیرهایی مواجه می شویم که قسمتی از اطلاعاتشان به دلایل مختلفی چون عدم پاسخ، ناکامل بودن چارچوب بررسی، ازدست رفتن اطلاعات موجود در پرونده ها و ... از دست رفته اند. به این گونه داده ها در آمار داده گمشده گویند و در صورت نادیده گرفتن آنها استنباط آماری را دچار مشکل می کنند و بخصوص در برآورد به روش درستنمایی، ایجاد اریبی کرده و کارایی را کاهش می دهند(۳,۲,۱). روشهای گوناگونی برای مقابله با مشکل داده های گمشده ابداع شده که بر اساس آنها گمشده ها جانهی می شوند. یکی از این روشها استفاده از الگوریتم EM است(۴). الگوریتم EM یک روش تکرار شونده است که در هر تکرار دو گام را شامل می شود. گام E(گام امید ریاضی) که داده گمشده به

شرط داده های مشاهده شده محاسبه می شوند این امیدهای ریاضی را به جای داده های گمشده قرار می دهند و پارامترهای مورد نظر برآورد می شوند. در گام بعدی یعنی گام M (گام ماکزیمم کردن) بعد از جایگذاری اعداد اولیه بجای داده های گمشده به شرط داده های مشاهده شده لگاریتم تابع درستنمایی را حداکثر می کنیم. این مکانیسم آنقدر تکرار می شود تا به همگرایی میان پارامترهای برآورد شده در تکرارها برسیم.

با این حال بیشترین روشهای جانهی متوجه داده های گمشده در متغیر پاسخ بوده است حال آنکه متغیرهای پیشگو نیز می توانند دستخوش تغییر و گمشدگی شوند. از اینرو داده های گمشده در متغیرهای پیشگو نیز مورد توجه برخی آمار دانان قرار گرفته است.

ابراهیم روش درستنمایی ماکزیمم را برای مدلهای عمومی رگرسیونی پیشنهاد می کندکه با بهره گیری از الگوریتم EM گمشده های متغیرهای مستقل گسسته برآورد می شوند(۵).

در اکثرمطالعات آماری معمولا متغیرهایی بیش از آنچه برای طراحی مدل مورد نیاز است در مورد آزمودنی ها جمع آوری می شوند. به این متغیرها که مقادیر آنها جمع آوری و ثبت شده ولی به عنوان یک متغیر پیشگو در مدل استفاده نشده اند متغیرهای کمکی گویند. در حالی که متغیر پیشگو کامل نیست این متغیرها ممکن است کاملا مشاهده شده باشند. استفاده از متغیرهای کمکی که داده هایشان کامل است در حالتی که متتغیرهای پیشگو کامل نیستند می تواند موجب افزایش کارایی آنالیز مدل شود.

در این تحقیق، برآورد درستنمایی را در مدل رگرسیون لجستیک بررسی می کنیم که متغیرهای مستقل گسسته آن ناکامل اند ولی متغیرکمکی گسسته آن کامل است(۶)و با استفاده از داده های تحقیقی که درباره علل تمایل زنان باردار به نوع زایمان است (۷) کارایی این روش را براساس انحراف معیار برآوردها با روش تحلیل مورد کامل مقایسه می کنیم.

۱- تحلیل مورد کامل: در این روش همه واحدهایی که دارای مقادیر گمشده اند کنار گذاشته می شوند. شاید این متداول ترین روش برای حل مشکل داده های گمشده باشد ولی در کل شیوه خوبی تلقی نمی شود زیرا با حذف همه واحدهای دارایِ داده هایِ گمشده، اندازه نمونه به صورتی قابل ملاحظه کوچک شده و منجر به کاهش دقت می شود. همچنین اگر واحدهایی که از تحلیل حذف می شوند با آنهایی که باقی می مانند تفاوت زیادِی داشته باشند، ممکن است برآوردهای حاصل به شدت اریب شوند(۸,۲).

۲- تحلیل درستنمایی با استفاده از الگوریتم EM و متغیر کمکی

۲-۱- الگوریتم EM : یکِی از روش هایی که برای برآورد داده های گمشده ابداع شده الگوریتم EM است.

الگوریتم EM یک روش محاسباتی عمومی برای برآوردهای حداکثر درستنمایی تحت داده های ناکامل است. تاریخچهٔ EM به سال ۱۹۲۶ و مقاله ایی از مک کندریک باز می گردد ولی اولین بار بوسیله دمپستر و همکاران ابداع شد(۴). نام این الگوریتم بر گرفته از دو گام این الگوریتم است: گام (E) که محاسبهٔ مقادیر مورد انتظار برای گمشده هاست و گام (M) که محاسبهٔ برآوردهای حداکثر درستنمایی پارامترها با فرض کامل بودن داده هاست که براساس دو مرحلهٔ پایه گذاری شد.

الف) اگر ما مقادیر گمشده را بدانیم می توانیم پارامترها را برآورد کنیم.

ب) اگر پارامترها را بدانیم می توانیم مقادیر گمشده را با مقادیر مورد انتظار جایگذاری کنیم.

۲-۲- مدل رگرسیون لجستیک با داده های کمکی

برای مشاهدات دو حالتی $y_1, y_2, \ldots, y_n$ و ماتریس x که مجموعه ایی از متغیرهای گسسته مستقل است مدل لجستیک زیر را داریم.

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

که بعد β ، $P \times 1$ است. حداکثر درستنمایی براساس احتمال شرطی $f\left(Y \middle| X, \beta\right)$ بدست می آید ولی اگر x دچار داده های گمشده باشد روش ML مؤثر نیست. ابراهیم روش درستنمایی ماکزیمم را پیشنهاد می کند که با استفاده از الگوریتم EM انجام شده و بجای مدل سازی احتمال Y به شرط X از احتمال توأم آنها نیز استفاده می کند(۵).

$$f(Y, X|\Omega) = f(Y|X, \beta) f(X|\gamma) \quad (۲)$$

که $\Omega = (\beta, \gamma)$

زیر مجموعه ایی از این متغیرها را وارد مدل می کنند و عموما تعدادی از متغیرها علی رغم گردآوری اطلاعاتشان، در مدل استفاده نمی شوند. به این گونه متغیرها، متغیرهای کمکی گویند(۶).

اگر متغیرکمکی A موجود باشد می توانیم تابع درستنمایی داده های کامل را به صورت زیر بنویسیم:

اگر $\boldsymbol{x}$ به طور کامل مشاهده شده باشد $f(X;Y)$ دردرستنمایی برای β شرکت نمی کند و اگر داده گمشده در $\boldsymbol{x}$ باشد برآورد β ممکن است دچار اریبی و کاهش کارایی شود. از سوی دیگر معمولا محققین مجموعه ایی از متغیرهای مستقل را در شروع تحقیق بررسی و اطلاعات آنها را ثبت می کنند ولی فقط

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

که $\Omega^* = (\beta^*, \gamma^*)$

و* را برای نشان دادن وابستگی به توزیع $\boldsymbol{A}$ بکار می بریم .

اگر فرض کنیم $f(Y|X, A, \beta^*) = f(Y|X, \beta)$ (یعنی استقلال شرطی y و A تحت x) روش ابراهیم را برای مدل می توان بکار برد.

واچ این برآوردگر را وقتی متغیر کمکی برای بهبود بخشیدن به مدل استفاده می شود پیشنهاد کرده است و زمانی بکار می رود که استقلال شرطی برقرار باشد(۹). اما استقلال شرطی همیشه برقرار نیست. اگر استقلال شرطی برقرار نباشد فاکتور بندی رابطهٔ (۳) طبیعی نیست زیرا ضرایب رگرسیونی برای $\mathcal{X}$ در $E(Y|X, A)$ عموماً برابر β در مدل (۱) نیستند .

روش پیشنهاد شده دیگری که به وسیله واچ برای داده های کامل بکار رفته بصورت زیر است(۱۰):

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

مجموعه این از پارامترهای درستنمایی لگاریتمی تابع $\theta = (\alpha, \beta, \gamma)$

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

است .

۲-۳- برآورد درستنمایی ماکزیمم تحت متغیرهای مستقل ناکامل و اطلاعات کمکی

اگر داده ها کامل باشند $\prod f(Y_i|X_i, \beta)$ برای برآورد β براحتی ماکزیمم می شود. در این حالت β جدای از α و γ با رگرسیون لجستیک برآورد می شود. اما اگر Xi های گمشده باشند هر سه قسمت درستنمایی باید برآورد شوند. با فرض اینکه گمشده ها به صورت تصادفی حادث شده اند درستنمایی لگاریتمی زیر را داریم(۶):

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

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

بوسیلهٔ این تابع برآوردهای برای β بدست می آوریم که تحت متغیر کمکی و بدون فرض استقلال شرطی ماکزیمم شده است.

بر اساس روش ابراهیم (۵) متغیرهای پیشگو $X = (x_1, x_2, ..., x_p)$ متغیرهای گسسته تصادفی با توزیع چندگانه تحت پارامتر

$\gamma = (\gamma_1, \gamma_2, ..., \gamma_r)$ هستند. مثلاً اگر سه متغیر دو حالتی وجود داشته باشد بعد γ برابر است با $r = 2^3 - 1 = 7$. در حالت کلی فرض کنید $c_1,...,c_p$ نشان دهنده تعداد رده ها برای هر متغیر باشد. در نتیجه γ دارای بعد $r = c_1 \times ... \times c_p - 1$ است.

فرض کنید $f(Y|X,\beta)$ چگالی متغیر پاسخ و $f(A|X,Y,\alpha)$ و $f(X|Y)$ توزیع های چند جمله ایی با پارامترهای γ, α (مجزا از β) باشند.

اگر هیچ محدودیتی به توزیع X تحمیل نشود تحت داده های کامل، γ بوسیلهٔ $r+1$ خانه شمارشی مشاهده شده از رده های x در یک جدول توافقی p بعدی بدست می آید. به طور مشابه مدل چند جمله ایی برای برآورد $f(A|X,Y,\alpha)$ می تواند مورد استفاده قرار گیرد و رگرسیون لجستیک را وقتی X کاملا مشاهده شده است برای برآورد β بکار رود.

اما در حالتی که X ناکامل است از الگوریتم EM (۴)برای برآورد پارامترها در درستنمایی داده های مشاهده شده استفاده می کنیم. این الگوریتم تکرار شونده شامل دو مرحله است. در مرحله E امید $L_{a,y,x}(\theta|a,Y,X)$ براساس پارامترهای $\theta^{(t)}$ و داده های مشاهده شده محاسبه می شوند.

در این وضعیت، داده های مشاهده شده شامل $(y_i, x_{obs,i}, a_i)$ اند که $X_{obs,i}$ نشان دهنده مجموعه ایی با بعد کوچکتر یا مساوی P متغیر مستقل در X_i است و اگر i امین عضو مشاهده شده در متغیرهای مستقل باشد معادل X_i خواهد بود.

این امید با $Q(\theta|\theta^{(t)})$ نشان داده می شود و داریم(۶):

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{r+1} w_{ij}^{(t)} L_{a,y,x}(\theta|a_i, y_i, x^j)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{r+1} w_{ij}^{(t)} \{ l_{a|y,x}(\alpha|a_i, y_i, x^j) + l_{y|x}(\beta|y_i, x^j) + L_x(\gamma|x^j) \} \quad (٦)$$

که x^j، j امین نمونه ممکن از متغیرهای مستقل، $L_{a,y,x}(\theta|a_i, y_i, x_i)$ درستنمایی لگاریتمی داده های کامل برای θ با i امین مشاهده تحت X_i ارزیابی شده با x^j است و $w_{ij}^{(t)} = p(x^j|a_i, y_i, x_{obs,i}, \theta^{(t)})$ میتواند وزنی از j امین نمونه برای i امین مشاهده در t امین تکرار باشد. در نتیجه $W_{it}^{(t)} = 1$. البته بیشتر این w_{ij} ها صفر خواهند شد.

پس هر مقدار $X_{obs,i}$ تحت x^j مناسب تشخیص داده نشود مورد استفاده قرار نمی گیرد. (ستون x^j و $X_{obs,i}$ زمانی متناسب اند که اعضای متناظر با ستون x^j برابر داده های مشاهده شده $X_{obs,i}$ باشد).

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

وزنها را می توان از قانون بیز به صورت زیر محاسبه کرد(۶):

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

$$= \begin{cases} 0 & if x^j \text{ is not compatibel with } x_i \\ \dfrac{p(y_i|x_i^j)P(a_i|x_i^j, y_i)p(x_i^j)}{\sum_{k \in obs,i} p(y_i|x_i^k)p(a_i|x_i^k, y_i)p(x_i^k)} & \\ & if x^j \text{ is compatibel with } x_i \end{cases} \quad (۷)$$

که x^j نشان دهنده یک بردار p تایی است که اجزاء مشاهده شده اش $X_{obs,i}$ و اجزاء باقی مانده آن بر اساس j امین نمونه برای متغیرهای گمشده است.

برای مرحله M معادلهٔ (۶) را به عنوان یک تابع از θ و بوسیله یافتن راه حل معادله درستنمایی لگاریتمی داده های کامل با استفاده از این وزنها ماکزیمم می کنیم. اکنون می توانیم از رگرسیون لجستیک وزن داده شده برای برآورد β استفاده کنیم و برای برآورد γ, α نیز از خانه های شمارش شده مورد انتظار بهره می گیریم.

برای اولین تکرار، وزنها را می توان از رابطه (۷) و نخستین برآوردها برای θ را با فرض کامل بودن داده ها بدست آورد. این مراحل آنقدر تکرار می شوند که به برآوردهای درستنمایی ماکزیمم همگرا شوند و اختلاف میان پارامترها در تکرارهای بعد معنی دار نباشد.

روش کار

برای مقایسه کارایی دو روش تحلیل مورد کامل و تحلیل درستنمایی با الگوریتم EM از اطلاعات تحقیقی استفاده کرده ایم که در مورد عوامل مرتبط با انتخاب نوع زایمان در زنان مراجعه کننده به مراکز بهداشتی و درمانی شهر تهران (۷) است. این تحقیق یک مطالعه توصیفی بوده است و در آن نمونه ای به حجم ۳۸۵ نفر از مادران باردار با سن حاملگی ۲۸ هفته و بالاتر از مراکز بهداشتی و درمانی شهر تهران انتخاب شدند. روش نمونه گیری از نوع چند مرحله ای بوده و مشخصات فردی، سوابق مامایی، نوع نگرش و عوامل اجتماعی نمونه ها از طریق پرسشنامه ثبت شدند. متغیر پاسخ در این تحقیق نوع زایمان انتخاب شده بوسیله زنان باردار است که به دو سطح زایمان طبیعی و سزارین تقسیم شده است. متغیرهای پیشگویی که برای ساختن مدل انتخاب شده اند شش متغیر دوحالتی هستند شامل شغل آزمودنی، شرکت در کلاسهای آموزشی دوران بارداری، نوع حاملگی، نگرش زنان باردار به زایمان طبیعی ، داشتن سابقه تولد نوزاد باوزن بیشتر از ۴۰۰۰ گرم و داشتن سابقه تولد نوزاد باوزن کمتر از ۲۵۰۰ گرم (جدول یک).

در متغیر داشتن سابقه تولد نوزاد باوزن کمتر از ۲۵۰۰، ۲۵ درصد از داده ها به عنوان داده گمشده به صورت تصادفی حذف شدند. همچنین داشتن سابقه زایمان پس از موعد نیز به همراه سایر متغیرها در پرسشنامه ثبت شده و اطلاعات آن کامل بود که در مدل وارد نشده و به عنوان یک متغیر کمکی مورد استفاده قرار گرفت (این متغیر نیز دو حالتی است).

ابتدا با روش تحلیل مورد کامل، قسمتهای گمشده از مطالعه حذف و پارامترها برآورد شدند. درمرحله بعد، از برآورد درستنمایی ماکزیمم تحت متغیر کمکی برای برآورد پارامترهای مدل استفاده کردیم. یک برنامه کامپیوتری تحت نرم افزارs-plus برای این منظور مورد استفاده قرار گرفته و برای مقایسه میزان کارایی دو روش برآورد انحراف معیار پارامترها مورد استناد قرار گرفت که انحراف معیارها در روش تحلیل درستنمایی با الگوریتم EM از روش لوئیس (۱۱) برآورد شدند.

یافته ها

نتایج حاصل از برآورد انحراف معیارهای دو روش تحلیل مورد کامل و جانهی با الگوریتم EM نشان می دهد که انحراف معیارهای برآورد شده برای

تمام پارامترهادر روش جانهی با الگوریتم EM کمتر از انحراف معیارهایی هستند که در روش تحلیل مورد کامل بدست آمده است که نشان مِیدهد کارایی برآورد پارامترها در اِین روش بهتر از روش تحلِیل مورد کامل است و علِی رغم اِینکه هِیچ قسمت از اطلاعات از مدل حذف نمِیشوند پارامترهایی با انحراف معِیارهای پایینتر برآورد مِیشوند و درنتِیجه فاصله اطمِینانهای کوتاهتری برای پارامترها بدست مِی آید(جدول دو). اگر چه اختلافات قابل توجهی نیز در برآورد پارامترهای لجستیک در دو روش مشاهده می شود که احتمالا ناشِی از اریبی حاصل از حذف قسمتِی از اطلاعات در روش تحلیل مورد کامل است(۲).

بحث

مشکل داده هایِ گمشده در بسیاری از مطالعات آماری وجود دارد و اگر بدون توجه به وجود گمشده ها، مدل رابرازش دهیم و ضرایب رگرسیونی را با روش درستنمایی ماکزیمم برآورد کنیم گمشده ها بر برآورد ها اثر منفی گذاشته و موجب اریبی و کاهش کارایی می شوند(۳,۲,۱). در این تحقیق روشی را بررسی کردیم که با بهره گیری از متغیری کمکی آنالیز درستنمایی ماکزیمم را در مدل لجستیک انجام داده و با استفاده از الگوریتم EM برآورد پارامترها را در حالتیکه متغیر پاسخ کامل نیست محاسبه می کند. نتایج حاصل از بررسِی این روش در مدل مربوط به عوامل موثر بر انتخاب نوع زایمان نشان داد که برآورد انحراف معیار پارامترها در این روش از تحلیل مورد کامل کمتر است که این مساله با یافته هایِ مدل کاربردِی هورتون و لایرد تطابق دارد(۶).

وجود متغیری کمکی که اطلاعات آن به طور کامل موجود است ولی مستقیما به عنوان یک متغیر پیشگو در مدل مورد استفاده قرار نگرفته میتواند برای افزایش کارایی مدل و برآورد دقیقتر پارامترها مفید باشد. انتخاب متغیر کمکی بستگی به اطلاعات موجود در تحقیق و نظر محقق دارد ولی هرچقدر ارتباط میان متغیر کمکی و متغیری که دچار گمشدگی شده بیشتر باشد این روش مؤثرتر خواهد بود(۶). از طرف دیگر برآورد انحراف معیار پارامترهایی که با EM بدست آمده اند مشکل است. لوئیس روشِی برای برآورد انحراف معیار پیشنهاد کرده که در این تحقیق مورد استفاده قرار گرفته است(۱۱). همچنین از روش بوت استراپ نیز مِی توان برای برآورد انحراف معیارها استفاده کرد.

توزیعهای $f(A|X,Y,\alpha), f(X;\delta)$ از طریق برآورد پارامترهای توزیع چند جمله ایی برآورد می شوندهرچند گمشده ها می توانند تاثیر نامطلوبی بر برآورد‌داشته باشند.

از سوی دیگر اگر Y متغیری پیوسته باشد بدست آوردن مدلی برای $f(A|Y,X)$ روشی متفاوت می طلبد یا می توان با تبدیل کردن Y به متغیری گسسته از روش قبل استفاده کرد.

توسعه این روش به سایر مدلهای رگرسیونی قابل بررسی است و برخی از این مدلها – بدون استفاده از متغیر کمکی – بوسیله ابراهیم و همکاران مورد توجه قرار گرفته است. همچنین توجه به نوع مکانیسم گمشدگِی (۱۲) در استفاده از روش جانهی با الگوریتم EM مهم است زیرا این روش مانند اکثر روشهای جانهی (۱۳,۲) صرفا در مکانیسمهای گمشدگِی تصادفِی مناسب عمل می کند.

جدول ۱: متغیرهای توضیحی مدل

متغیر	رده ها

شغل آزمودنی	شاغل=۱ ، خانه دار=۰
شرکت در کلاسهای آموزشی دوران بارداری	شرکت =۱ ، عدم شرکت= ۰
نوع حاملگی	خواسته=۱ ، ناخواسته=۰
نگرش زنان باردار به زایمان طبیعی	نگرش مثبت = ۰
	نگرش منفی = ۱
سابقه تولد نوزاد باوزن بیشتر از ۴۰۰۰ گرم	سابقه دارد=۱ ، سابقه ندارد=۰
سابقه تولد نوزاد باوزن کمتر از ۲۵۰۰ گرم	سابقه دارد=۱ ، سابقه ندارد=۰

جدول ۲- برآورد پارامترها و انحراف معیار آنها از دو روش درستنمایی الگوریتم EM و تحلیل مورد کامل

	روش درستنمایی با الگوریتم EM و متغیر کمکی			روش تحلیل مورد کامل
پارامترهای مدل	انحراف معیار	برآورد ضریب رگرسیونی	انحراف معیار	برآورد ضریب رگرسیونی
عرض از مبدا	۰/۲۷۲۸	-۱/۳۱۷۲	۰/۴۶۲	-۲/۲۴۲
شغل آزمودنی	۰/۴۰۰۶	-۱/۲۴۹۶	۰/۹۳۵	۰/۹۸۵
شرکت در کلاسهای آموزشی دوران بارداری	۰/۲۶۸۳	-۰/۵۳۳۹	۰/۴۵۵	۰/۱۶۶
نوع حاملگی	۰/۲۲۷	-۰/۳۲۲۱	۰/۳۲۱	-۰/۰۸۳
نگرش زنان باردار به زایمان طبیعی	۰/۲۸۶۴	-۰/۵۷۹۴	۰/۳۵۴	۰/۴۶۰
سابقه تولد نوزاد باوزن بیشتر از ۴۰۰۰ گرم	۰/۲۸۴۳	۲/۴۸۳۸	۰/۳۸۳	۲/۵۳۱
سابقه تولد نوزاد باوزن کمتر از ۲۵۰۰ گرم	۰/۳۸۶۷	-۰/۳۴۸۸	۰/۸۱۳	۰/۲۱۰

References

۱- پروانه وار سیمین. بررسی عوامل مرتبط با انتخاب نوع زایمان در زنان باردار مراجعه کننده به مراکز بهداشتی و درمانی شهر تهران سال ۱۳۸۲. پایان نامه کارشناسی ارشد رشته مامایی، دانشگاه علوم پزشکی شهید بهشتی 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