مدل رگرسیون کلاس پنهان با متغیرهای کمکی مؤثر بر پاسخ پنهان و پاسخ‌های مشاهده شده

آوات فیضی، جمشید جمالی، سید محسن حسینی
گروه آمار زیستی و اپیدمیولوژی، دانشگاه علوم‌پزشکی اصفهان

در بسیاری از حوزه های پژوهشی مفاهیم نظری وجود دارند که مستقیم قابل مشاهده یا اندازه گیری نیستند که در ادبیات آماری موسوم به متغیرهای پنهان هستند. این مفاهیم براساس متغیرهای قابل مشاهده بعنوان نشانگرهایی از آنها استخراج می گردند. مدلهای متغیر پنهان ابزار توانمندی در مدلسازی رابطه بین متغیرهای مشاهده شده از طریق استخراج متغیرهای پنهان می باشند. در حوزه متغیرهای پنهان ، تحلیل کلاس پنهان با فرض وجود رابطه بین مجموعه ای از متغیرهای مشاهده شده گسسته اقدام به مدلبندی رابطه بین آنها در قالب استخراج یک متغیر پنهان گسسته می نماید. در سالهای اخیر ، مدلهای کلاس پنهان کاربرد فراوانی در تجزیه و تحلیل ارتباط بین متغیرهای گسسته و متغیرهای کمکی یافته اند که در قالب مدلهای رگرسیون کلاس پنهان مورد استفاده قرار گرفته اند. در این مقاله فرم گسترش یافته ای از رگرسیون کلاس پنهان معرفی می شود که امکان ارزیابی اثرگذاری متعیرهای کمکی بر متغیر پاسخ پنهان و نیز متغیرهای پاسخ مشاهده شده را همزمان فراهم می کند. مبانی نظری مدل بیان می شود، روشهای برآورد پارامترها و شناسایی پذیری مدل توضیح داده می شود. روش های نظری بسط مدل که موجب گسترش کارکرد عملی آن می شود مورد اشاره قرار می گیرد. کارکرد عملی مدل با ذکر مثالی در حوزه پزشکی به منظور سنجش میزان آگاهی افراد از علائم هشدار دهنده سرطان و عوامل مؤثر بر آن شرح داده می شود.

واژه‌های کلیدی: متغیر پنهان، رگرسیون کلاس پنهان، شناسایی پذیری، علائم هشدار دهنده سرطان.

۱ مقدمه

بسیاری از مفاهیم و معیارهای موجود در علوم اجتماعی ، رفتاری ، پزشکی و اپیدمولوژی بطور واضح و مستقیم قابل ارزیابی نیستند.نگرشها،استعداد ، خصوصیات اخلاقی ، رفتارهای اجتماعی ، ویژگیهای شخصیتی ، میزان آگاهی و بعضی از بیماریهای روان پزشگی مانند

استرس ، افسردگی ، اسکیزوفرنی و ... از این مقوله هستند که در ادبیات آماری به متغیرهای پنهان موسوم هستند. رهیافتی که به استخراج و مطالعه ماهیت واقعی مفاهیم پنهان کمک می کند مدلهای متغیر پنهان می باشد ، زیرمجموعه ای از این دسته کلی شامل آنالیز کلاس پنهان و رگرسیون کلاس پنهان می باشد. ساده ترین مدل در این زیر مجموعه ، آنالیز کلاس پنهان می باشد. آنالیز کلاس پنهان یک روش آماری برای استخراج زیر جامعه های همگن از یک جامعه براساس الگوی پاسخ آزمودنی ها به داده های چندمتغیره گسسته مشاهده شده می باشد. ایده اصلی آنالیز کلاس این است که آزمودنی های مورد مطالعه با هم همبسته هستند به گونه ای که جوامع مورد مطالعه را می توان ترکیبی از چندین زیرجامعه همگن در نظر گرفت. هدف آنالیز کلاس پنهان یافتن این زیرجوامع می باشد که نقش کلاسهای پنهان را ایفا می نمایند. در آنالیز کلاس پنهان از میان آزمودنی ها آنهایی که از لحاظ میزان پاسخ دهی به متغیرهای مشاهده شده شباهت ملموسی دارند در یک کلاس قرار می گیرند. آزمودنی های واقع در هر کلاس کاملا مشابه و آزمودنی های واقع در کلاسهای مختلف متفاوت از همدیگر هستند و همین خاصیت ، امکان انجام مقایسه در بین کلاسها را از لحاظ الگوی پاسخ به متغیرهای مشاهده شده فراهم می کند. براین اساس تحلیل کلاس پنهان در داده های گسسته را می توان متناظر با تحلیل عاملی در نظر گرفت. تحلیل عاملی با ساختار متغیرها مرتبط است یعنی متغیرها بر مبنای ملاکهایی همچون ضریب همبستگی در گروهها دسته بندی می شوند اما آنالیز کلاس پنهان با ساختار آزمودنی ها مرتبط است.

رگرسیون کلاس پنهان حالت تعمیم یافته آنالیز کلاس پنهان می باشد که در آن متغیرهای کمکی ، برای بهینه کردن تشکیل کلاسهای پنهان و ارزیابی نحوه اثرگذاری بر قرار گرفتن آزمودنی ها در کلاسهای تشکیل شده (نقش متغیر پاسخ را در این مدلها ایفا می کنند) وارد مدل می شوند. بر این مبنا این مدل رگرسیونی کاملاً شبیه مدل رگرسیون لجستیک می باشد با این تفاوت که در رگرسیون لجستیک متغیر پاسخ یک متغیر گسسته است که به طور مستقیم قابل مشاهده و اندازه گیری است ، در حالیکه در رگرسیون کلاس پنهان متغیر پاسخ یک متغیر گسسته (کلاسهای پنهان) است که به طور مستقیم قابل مشاهده یا اندازه گیری نیست. در رگرسیون لجستیک امکان مدلبندی رابطه فقط یک متغیر پاسخ با متغیرهای مستقل وجود دارد در حالیکه مدل رگرسیون کلاس پنهان امکان مدلبندی رابطه چندین متغیر پاسخ را با متغیرهای مستقل فراهم می سازد.

آغاز استفاده از آنالیز کلاس پنهان به اوایل دهه ۱۹۵۰ میلادی توسط پژوهشگران حوزه روانشناسی از جمله لازارسفلد و هنری برمی گردد، که در مطالعه خویش اقدام به ساختن الگوهایی براساس متغیرهای دوحالتی کردند (لازارسفلد ، ۱۹۵۰ و لازارسفلد و هنری ، ۱۹۶۸). در سال ۱۹۷۴ گودمن با معرفی یک روش نسبتاً ساده برای بدست آوردن

برآوردگرهای ماکزیمم درستنمایی برای پارامترهای مدل کلاس پنهان ، کاربرد این مدلها را در رشته های مختلف تحقیقی عمومی تر کرد(گودمن ، ۱۹۷۴). فورمن در سال ۱۹۸۲ با استفاده از الگوریتم EM روش پیشنهادی گودمن را توسعه داد(فورمن ، ۱۹۸۲). در سال ۱۹۸۸ دایتون و مک ریدی متغیرهای کمکی را وارد آنالیز کلاس پنهان کردند و تأثیر این متغیرها را بر کلاسهای تشکیل شده بررسی نمودند و در واقع رگرسیون کلاس پنهان را بنیانگذاری کردند(دایتون و مک ریدی ، ۱۹۸۸). فورمن در سال ۱۹۹۲ چارچوب کلی از مدلهای رگرسیون کلاس پنهان را پیشنهاد کرد که انواع مختلفی از این مدلها را پوشش می داد(فورمن ، ۱۹۹۲). بندن-روچ در سال ۱۹۹۷ مدل رگرسیون کلاس پنهان را برای متغیرهای گسسته چند رده ای بسط داد(بندن-روچ و همکاران ، ۱۹۹۷). هوانگ در سال ۲۰۰۴ فرم گسترش یافته ای از مدلهای رگرسیون کلاس پنهان را که یکی از اجزای مدل کلی پیشنهادی توسط فورمن بود را با جزئیات نظری کامل تر ارائه کرد ، در مدل پیشنهادی هوانگ تأثیر متغیرهای کمکی علاوه بر متغیرهای پنهان بر متغیرهای مشاهده شده نیز سنجیده می شد(هوانگ ، ۲۰۰۴). مدل یاد شده بطور همزمان در قالب یک مدل واحد کارکرد مدل رگرسیون کلاس پنهان معمولی و نیز نتایجی را که از اجرای همزمان چندین رگرسیون لجستیک به آن خواهیم رسید در خود دارد. بنابراین چنین مدلی می تواند به لحاظ کاربردی در بسیاری از حوزه های پژوهشی که باساختارهای نظری پنهان مواجه هستند بویژه در حوزه پزشکی و بهداشت در مواردی که پژوهشگران باسنجش آگاهی و بویژه با تشخیص های پزشکی سروکار دارند ابزار توانمندی در استخراج ساختار پنهان مورد مطالعه (برای مثال آگاهی ، تشخیص) و ارزیابی عوامل اثرگذار بر آن و بر اجزای سازنده آن (نشانگرهایی که ساختار پنهان براساس آنها ساخته می شود)باشد. در مقاله حاضر مبانی نظری روش پیشنهادی هوانگ معرفی و با توجه به اهمیت کارکرد عملی که در حوزه پزشکی و بهداشت می تواند داشته باشد مثالی در این حوزه شرح داده می شود و چگونگی گسترش نظری این مدل که موجب بسط کارکرد عملی آن خواهد شد مورد اشاره قرار خواهد گرفت.

۲ مبانی نظری مدل

در راستای معرفی ساختار نظری مدلهای رگرسیون کلاس پنهان ، فرض کنید$Y_i = (Y_{i1}, ..., Y_{iM})'$ نشان دهنده یک بردار با M متغیر مشاهده شده برای آزمودنی iام از یک نمونه n عضوی باشد.فرض می شود که Y_iها با هم وابسته آماری هستند. ایده اصلی

مدل کلاس پنهان این است که آزمودنی های مورد مطالعه با هم همبسته هستند و جوامع مورد مطالعه ترکیبی از J زیرجامعه می باشند که یکی از اهداف رگرسیون کلاس پنهان یافتن این زیر جوامع و برآورد حجم هر زیر جامعه می باشد.

فرض کنید هر متغیر مشاهده شده مورد بررسی K رسته داشته باشد.($K > ۱$) و S_i نشان دهنده iامین زیر جامعه (کلاس پنهان) باشد و هر متغیر مشاهده شده مورد بررسی K رسته داشته باشد؛ K_m تعداد رسته های متغیر mام می باشد و $m = ۱, ..., M$ ، $S_i = ۱, ..., J$ ، $Y_{im} = ۱, ..., K_m$.

حداکثر تعداد کلاسهای ممکن برابر است با $K_۱ \times K_۲ \times ... \times K_m \times ... \times K_M$ که در آن K_M تعداد رسته های متغیر mام می باشد.

تابع چگالی Y_i در آنالیز کلاس پنهان بصورت زیر تعریف می شود.

$$p[Y_{i۱} = y_۱, ..., Y_{iM} = y_M] = \sum_{j=۱}^{J} \eta_i(x_i'\beta) \prod_{m=۱}^{M} \prod_{k=۱}^{k_m} p_{mkj}^{y_{mk}}$$

که در آن $p(s_i = j) = \eta_j$ احتمال عضویت هر آزمودنی در کلاس jام و p_{mkj} احتمال انتخاب kامین رسته متغیر mام به شرط حضور در کلاس jام توسطآزمودنی می باشد. $y_{mk} = ۱$ است اگر $y_m = k$ و در غیر اینصورت مقدار آن صفر می باشد.

فورمن و هوانگ تعمیمی از رگرسیون کلاس پنهان را ارائه نمودند که متغیرهای کمکی علاوه بر اثرگذاری بر کلاسها ، متغیرهای پاسخ مشاهده شده را نیز تحت تأثیر می گذارند.(فورمن ، ۱۹۹۲ و هوانگ ، ۲۰۰۴)

معادله این نوع رگرسیون کلاس پنهان که همزمان کلاس بندی آزمودنی ها و متغیرهای مشاهده شده اثرپذیر از متغیرهای کمکی می باشد ، بصورت زیر می باشد:

$$p[Y_{i۱} = y_۱, ..., Y_{iM} = y_M | x_i, z_i] = \sum_{j=۱}^{J} \eta_i(x_i'\beta) \prod_{m=۱}^{M} \prod_{k=۱}^{k_m} p_{mkj}^{y_{mk}}(\gamma_m + z_{im}'\alpha_m)$$

که در آن X بردار متغیرهای کمکی مؤثر بر کلاس بندی (متغیر پاسخ پنهان) و Z بردار متغیرهای کمکی مؤثر بر متغیرهای پاسخ مشاهده شده می باشد. همچنین $\eta_i(x_i'\beta)$ و $p_{mkj}^{y_{mk}}(\gamma_m + z_{im}'\alpha_m)$ توابع پیوندی هستند که در چارچوب مدلهای خطی تعمیم یافته تعریف می شود (مَک کولاق و نلدر ، ۱۹۸۹). توابع پیوند گوناگونی می توانند به آسانی برای این منظور مورد استفاده قرار گیرند. عمده ترین پیشنهاد استفاده از تابع لجیت تعمیم یافته می

باشد (اگرستی ،۲۰۰۲).

$$log[\frac{\eta_j(x_i'\beta}{\eta_J(x_i'\beta}] = \beta_{\circ j}+\beta_{\backslash j}x_{i\backslash}+...+\beta_{pj}x_{p\backslash} \qquad for \qquad i = \backslash,...,n; \qquad j = \backslash,...,J-\backslash$$

و

$$log[\frac{p_{mkj}(\gamma_{mj} + z_{im}'\alpha_m)}{p_{mKj}(\gamma_{mj} + z_{im}'\alpha_m)}] = \gamma_{mkj} + \alpha_{\backslash mk}z_{im\backslash} + ... + \alpha_{Lmk}z_{imL}$$

برای

$$i = \backslash,...,n; \qquad m = \backslash,...,M; \qquad k = \backslash,...,(K_m - \backslash); \qquad j = \backslash,...,J$$

برای انجام رگرسیون کلاس پنهان سه پذیره زیربنایی وجود دارد(بندن-روچ و همکاران ، ۱۹۹۷ و هوانگ ، ۲۰۰۵).

۱ـ احتمال عضویت آزمودنی ها در کلاس پنهان تنها به متغیر کمکی x مربوط است و متغیر z تأثیری بر کلاس بندی آزمودنی ها ندارد. $p(S_i = j|x_i, z_i) = p(s_i = j|x_i)$

۲ـ در داخل کلاسها ، متغیرهای مشاهده شده مورد بررسی از متغیرهای کمکی مستقلند.

$$p[Y_{i\backslash} = y_\backslash,...,Y_{iM} = y_M|S_i, x_i, z_i] = p[Y_{i\backslash} = y_\backslash,...,Y_{iM} = y_M|S_i, z_i]$$

۳ـ در داخل کلاسها ، متغیرهای پاسخ مشاهده شده مستقل از یکدیگرند.(استقلال موضعی)

$$p[Y_{i\backslash} = y_\backslash,...,Y_{iM} = y_M|S_i, z_i] = \prod_{m=\backslash}^{M} p[Y_{im} = y_m|S_i, z_{im}]$$

۱-۲ شناسایی پذیری

برای بررسی یکتا بودن پارامترها، شناسایی پذیری[1] باید ارزیابی گردد.

بنا به تعریف ، توزیع FY را برای پارامتر $\phi_\circ$ شناسایی پذیر گوئیم هرگاه یک X در همسایگی $\phi_\circ$ وجود داشته باشد بطوریکه: $\phi \in$ $y \in U_Y \leftrightarrow \phi = \phi_\circ$ $F_Y(y;\phi_\circ) = F_Y(y;\phi)$

[1] Identifiability

$$X \subset \phi$$

که در آن ϕ فضای پارامتر و U_Y تکیه گاه y می باشد.
به زبان ساده ، شناسایی پذیری با این موضوع سروکار دارد که آیا برای هر یک از پارامترهای آزاد مدل یک مقدار یگانه از روی داده های مشاهده شده به دست می آید یا خیر؟

در رگرسیون کلاس پنهان و تعمیم های ارائه شده بر آن ، بیشتر بحث شناسایی پذیری بر روی شناسایی پذیری موضعی متمرکز شده است. شرایط کافی برای شناسایی پذیری مدل رگرسیون کلاس پنهان بصورت زیر می باشد.

قضیه: مدل رگرسیون کلاس پنهان شناسایی پذیر است هرگاه:
۱) تعداد پارامترهای مورد برآورد از تعداد پارامترهای مدل اشباع کمتر باشد.
۲) مقادیر پارامترهای مدل $(\gamma_{mkj}, \alpha_{qmk}, \beta_{pj})$ و متغیرهای کمکی متناهی باشند.
۳) ماتریس اطلاع فیشیر مدل ، پررتبه ستونی باشد یعنی مقادیر ویژه ماتریس اطلاع فیشیر بزرگتر از صفر باشد.
۴) ماتریس طرح متغیرهای کمکی و پیشگو پررتبه ستونی باشد.

در عمل برای بررسی شناسایی پذیری مدل از چندین مجموعه مقادیر اولیه برای برآورد پارامترها استفاده می کنیم ، چنانچه مقادیر متفاوت اولیه در تابع درستنمایی ماکزیمم ، منجر به برآورد مشابهی برای پارامترها شود ، مدل شناسایی پذیر خواهد بود (هوانگ ،۲۰۰۴).

۲-۲ برآورد پارامترها

برآورد پارامترها در مدلهای کلاس پنهان معمولاً از طرق روش حداکثر درستنمایی انجام می شود. از آنجائیکه‌در مدلهای رگرسیون کلاس پنهان با متغیرهای گسسته و حجم وسیع داده ها روبرو هستیم برای برآورد پارامترهای معمولا از روشهای پیشرفته ریاضی و الگوریتم های خاص از جمله الگوریتم EM استفاده می شود.

الگوریتم EM تابع درستنمایی داده های ناکامل را در یک فرآیند تکراری بین مرحله ماکزیمم سازی تابع درستنمایی داده های کامل(مرحله M) و جانهی داده های ناکامل از طریق مدلی که پارامترهای آن در آخرین تکرار برآورد شده بودند(مرحله E)، ماکزیمم می کند. عضویت در کلاس های پنهان S_i به عنوان یک متغیر غیر قابل مشاهده ، ساختار داده های ناکامل را در مدل رگرسیونی مدنظر شکل می دهد؛ در زیر به اختصار عملکرد این الگوریتم را در برآورد پارامترهای مدل شرح می دهیم.

با فرض قابل مشاهده بودن S_i فرم لگاریتم تابع درستنمایی برای داده های کامل بصورت زیر خواهد بود.

$$logL_c(\phi;Y,S) = \sum_{i=1}^{n}\sum_{j=1}^{J} S_{ij}[log\eta_j(x_i'\beta)] + \sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{k=1}^{K_m} S_{ij}Y_{imk}[logp_{mkj}(\gamma_{mj}+z_{im}'\alpha_m)]$$

که در آن S_{ij} نشان دهنده عضویت iامین آزمودنی در jامین کلاس و Y_{imk} نشان دهنده kامین سطح متغیر mام که توسط iامین آزمودنی پاسخ داده شده است. همچنین $\phi=(\gamma_{mj},\alpha_m,\beta)$ پارامترهای مدل می باشند.
تابع Q را به صورت زیر تعریف می کنیم.

$$Q(\phi|\phi') = E[logL_c(\phi,Y,S|Y=y,\phi',x,z]$$

این تابع، امید لگاریتم تابع درستنمایی به شرط داده های مشاهده شده y,x,z و برآوردهای بدست آمده پارامترهادر مراحل قبل یعنی (ϕ') می باشد. الگوریتم EM از ϕ^p به ϕ^{p+1} بصورت زیر است.
مرحله E: محاسبه $Q(\phi|\phi^p)$

$$Q(\phi|\phi^p) = \sum_{i=1}^{n}\sum_{j=1}^{J}\theta_{ij}(\phi^p)[log\eta_j(x_i'\beta)] + \sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{k=1}^{K_m}\theta_{ij}(\phi^p)y_{imk}[logp_{mkj}(\gamma_{mj}+z_{im}'\alpha_m)]$$

که در آن

$$\theta_{ij}(\phi^p) = E(S_{ij}|Y_i,\phi^p,x_i,z_i)$$

$$= \frac{\eta_j(x_i'\beta)\prod_{m=1}^{M}\prod_{k=1}^{k_m}p_{mkj}^{y_{mk}}(\gamma_{mj}^p+z_{im}'\alpha_m^p)}{\sum_{l=1}^{J}\eta_l(x_i'\beta)\prod_{m=1}^{M}\prod_{k=1}^{k_m}p_{mkl}^{y_{mk}}(\gamma_{ml}^p+z_{im}'\alpha_m^p)}$$

احتمال پسین عضویت کلاسی به ازای ϕ^p می باشد.
مرحله M: یافتن مقادیری از پارامترهای مدل (ϕ) که $Q(\phi|\phi^p)$ را ماکزیمم می نمایند.

۳-۲ تعیین تعداد کلاسهای پنهان (آزمونهای برازندگی مدل)

تعیین تعداد کلاسهای پنهان امری چالش برانگیز بین محققان می باشد. از همین رو ملاکهای گوناگونی برای انتخاب مدل مناسب مطرح شده اند. رایجترین روش این است که با نظر محقق تعداد کلاسها تعیین می شود سپس شاخصهای نیکویی برازش مانند $AIC۳ = -LnL + ۳p$ و $AIC = -۲LnL + ۲p$ و $BIC = -۲LnL + ۳pLn(n)$ را برای تعداد کلاسهای معین محاسبه می کنند و مدلی که مقادیر کمتر این شاخصها را داشته باشد مناسبتر می باشد.

تازه ترین مطالعات نشان می دهد که معیار AIC3 معیار بهتری نسبت به AIC , BIC برای تعیین تعداد کلاسهای پنهان در مدل رگرسیون کلاس پنهان می باشد(اندروز و کوریم ، ۲۰۰۳ و دیاس ،۲۰۰۴).

۳ مثال

برای توضیح رگرسیون کلاس پنهان از داده های که به منظور بررسی سطح آگاهی نسبت به علائم هشدار دهنده سرطان و عوامل مؤثر بر آن در ساکنین شهر تهران و حومه جمع آوری شده اند استفاده می کنیم. این مطالعه در سال ۱۳۸۶ و در بین ۲۵۰۰ شهروند ۱۸ ساله و بالاتر تهرانی صورت گرفته است. در این پژوهش ، آگاهی یا شناخت به عنوان یک مفهوم انتزاعی (پنهان) در نظر گرفته می شود که مستقیم قابل استنتاج نبوده و ارزیابی آن بر مبنای پاسخ به پرسشهای دوگزینه ای در خصوص شناسایی ۹ علامت هشدار دهنده سرطان (به عنوان متغیرهای پاسخ مشاهده شده) بعنوان یک متغیر پنهان گسسته استخراج می شود. این متغیر نقش متغیر پاسخ پنهان را در مدل رگرسیون ایفا می کند. در این مطالعه علائم تغییر در عادات دفع ادرار و مدفوع ، زخمی که بیش از سه هفته بهبود و ترمیم نیابد ، خونریزی یا ترشح غیر معمول ، سفتی و وجود توده ای در پستان و یا سایر ارگانها ، اشکال در بلع (اشکال در قورت دادن غذا) ، اشکال در هضم و جذب غذا(سوء هاضمه) ، تغییرات قابل ملاحظه در خالها و زگیل ها ، سرفه های مکرر و یا خشونت و تغییر در صدا، کاهش وزن ناگهانی بعنوان علائم احتمالی سرطان در نظر گرفته شده اند و تأثیر متغیرهای سن ، جنسیت ، وضعیت تاهل ، سطح تحصیلات ، سابقه خانوادگی ابتلا به سرطان ، مصرف دخانیات و مصرف الکل بر میزان آگاهی کلی و هر یک از علائم هشدار دهنده سنجیده می شود.

جدول ۱ نسبت افراد در کلاسهایی که بر مبنای جواب سئوالات مطرح شده در مورد

علائم هشدار دهنده سرطان توسط مدل رگرسیون کلاس پنهان ساخته شده است را نشان می دهد.تفسیر و نامگذاری هر کلاس بر مبنای درصدهای شناخت صحیح علائم صورت می گیرد. بر این اساس بررسی ساختار کلاس های تشکیل شده بر حسب سطح آگاهی افراد نشان می دهد کلاس اول که رده افراد آگاه را تشکیل می دهد ، شامل بخش کمی از پرسش شوندگان (۱۸/۸۳٪) می باشد و کلاسهای دوم و سوم در بردارنده قسمت اعظم افراد (در مجموع۸۲/۱۸٪) هستند که دارای اطلاع متوسط (۵۴/۴۹٪) و ضعیف (۲۶/۶۹٪) در مورد علائم هشدار دهنده سرطان هستند، بنابراین می توان نتیجه گیری کرد که در مجموع سطح آگاهی و شناخت نسبت به علائم هشدار دهنده سرطان در جامعه مورد مطالعه پایین است.

جدول۱ـ درصد شناخت صحیح علائم هشدار دهنده سرطان و نسبت افراد واقع در هر کلاس

علائم هشدار دهنده	پاسخ	کلاس۱	کلاس۲	کلاس۳
تغییر در عادات دفع ادرار و مدفوع	بله	۸۵/۸۳	۲۸/۰۶	۴/۷۳
	خیر	۱۴/۱۷	۷۱/۹۴	۹۵/۲۷
زخمی که بیش از سه هفته بهبود و ترمیم نیاید	بله	۹۵/۴۸	۴۴/۸۴	۲/۸۲
	خیر	۴/۵۲	۵۵/۱۶	۹۷/۱۸
خونریزی یا ترشح غیرمعمول	بله	۹۸/۲۲	۵۲/۱۸	۱/۹۵
	خیر	۱/۷۸	۴۷/۸۲	۹۸/۰۵
سفتی و وجود توده ای در پستان و یا سایر ارگانها	بله	۹۹/۸۶	۸۱/۷۴	۱۵/۳۶
	خیر	۰/۱۴	۱۸/۲۶	۸۴/۶۴
اشکال در بلع	بله	۹۵/۰۰	۲۸/۳۵	۰/۸۶
	خیر	۵/۰۰	۷۱/۶۵	۹۹/۱۴
اشکال در هضم و جذب غذا(سوء هاضمه)	بله	۹۴/۴۹	۳۴/۷۹	۳/۹۶
	خیر	۵/۵۱	۶۵/۲۱	۹۶/۰۴
تغییرات قابل ملاحظه در خال ها و زگیل ها	بله	۹۸/۲۰	۵۶/۰۹	۲/۹۸
	خیر	۱/۸۰	۴۳/۹۱	۹۷/۰۲
سرفه های مکرر و یا خشونت و تغییر در صدا	بله	۹۳/۰۰	۴۴/۰۸	۳/۲۷
	خیر	۷/۰۰	۵۵/۹۲	۹۶/۷۳
کاهش وزن ناگهانی	بله	۹۴/۶۶	۷۰/۶۹	۱۰/۶۱
	خیر	۵/۳۴	۲۹/۳۱	۸۹/۳۹
حجم هر کلاس		۱۸/۸۲	۵۴/۴۹	۲۶/۶۹

جدول ۲ تأثیر متغیرهای مستقل بر قرار گرفتن افراد در کلاس های تشکیل شده (سطوح مختلف آگاهی) در قالب ضرایب رگرسیونی را نشان می دهد. بررسی ضرایب متغیرها (نسبت بخت) در جدول نشان می دهد که مؤثرترین عامل بر سطح آگاهی میزان تحصیلات می باشد. به عنوان مثال افزایش یک واحدی در سطح تحصیلات ، شانس قرار گرفتن در رده سه نسبت به رده یک را ۷۰٪ کاهش می دهد. متغیر مستقل مهم دیگر که از نظر آماری معنادار گردیده است جنسیت می باشد. شانس قرار گرفتن مردان در کلاس سوم نسبت به کلاس اول ۱۴۷٪ بیشتر از زنان می باشد.از دیگر متغیرهای مستقلی که از نظر آماری معنادار شده اند وضعیت تاهل می باشد . شانس قرار گرفتن افراد متاهل در کلاس یک نسبت به کلاس سوم ، ۲ برابر افراد مجرد است. علاوه بر این با آنکه اثر سن معنادار گردیده است اما افزایش سن تأثیر چندانی در قرار گرفتن افراد در رده های مختلف ندارد.

بطور کلی افراد متاهل ، دارای تحصیلات بالاتر و زنان احتمال بیشتری دارند که در رده افراد دارای آگاهی بالا قرار گیرند.

جدول۲ـبرآورد اثرات متغیرهای مستقل بر متغیر پاسخ کلاس پنهان

متغیرها	ضرایب	بازه اطمینان ۹۵٪ برای ضرایب	آماره آزمون (p-مقدار)	نسبت بخت	بازه اطمینان ۹۵٪ برای نسبت بخت
کلاس ۲ (کلاس ۱ بعنوان مرجع)					
عرض از مبدا	۴٫۷۲	(۳٫۹۳ـ۵٫۵۰)	۱۱٫۸۴(۰٫۰۰۰)	۱۱۱٫۶۳	(۵۱٫۱۳ـ۲۴۳٫۷۳)
سن	ـ۰٫۰۳	(ـ۰٫۰۴ـ۰٫۰۲)	ـ۴٫۷۴(۰٫۰۰۰)	۰٫۹۷	(۰٫۹۶ـ۰٫۹۸)
جنس (زن)	۰٫۳۱	(۰٫۰۴ـ۰٫۵۷)	۲٫۲۴(۰٫۰۲۵)	۱٫۳۶	(۱٫۰۴ـ۱٫۷۷)
وضعیت تاهل (مجرد)	ـ۰٫۶۲	(ـ۰٫۹۴ـ۰٫۲۹)	ـ۳٫۷۳(۰٫۰۰۰)	۰٫۵۴	(۰٫۳۹ـ۰٫۷۵)
سطح تحصیلات	ـ۰٫۷۶	(ـ۰٫۹۴ـ۰٫۵۹)	ـ۸٫۵۶(۰٫۰۰۰)	۰٫۴۷	(۰٫۳۹ـ۰٫۵۵)
سابقه خانوادگی ابتلا (نداشتن)	ـ۰٫۲۸	(ـ۰٫۵۹ـ۰٫۰۴)	ـ۱٫۷۰(۰٫۰۸۹)	۰٫۷۶	(۰٫۵۵ـ۱٫۰۴)
استعمال دخانیات (نکردن)	۰٫۱۶	(ـ۰٫۱۶ـ۰٫۴۹)	۰٫۹۸(۰٫۳۳۰)	۱٫۱۸	(۰٫۸۵ـ۱٫۶۳)
مصرف الکل (نداشتن)	۰٫۰۱	(ـ۰٫۵۱ـ۰٫۵۴)	۰٫۰۵(۰٫۹۶۰)	۱٫۰۱	(۰٫۶۰ـ۱٫۷۱)
کلاس ۳ (کلاس ۱ بعنوان مرجع)					
عرض از مبدا	۴٫۷۳	(۳٫۸۳ـ۵٫۶۲)	۱۰٫۳۴(۰٫۰۰۰)	۱۱۳٫۰۲	(۴۶٫۱۲ـ۲۷۶٫۹۷)
سن	ـ۰٫۰۲	(ـ۰٫۰۳ـ۰٫۰۰)	ـ۲٫۵۰(۰٫۰۱۲)	۰٫۹۸	(۰٫۹۷ـ۱٫۰۰)
جنس (زن)	۰٫۹۱	(۰٫۵۹ـ۱٫۲۲)	۵٫۷۱(۰٫۰۰۰)	۲٫۴۷	(۱٫۸۱ـ۳٫۳۷)
وضعیت تاهل (مجرد)	ـ۰٫۶۹	(ـ۱٫۰۸ـ۰٫۲۹)	ـ۳٫۴۳(۰٫۰۰۱)	۰٫۵۰	(۰٫۳۴ـ۰٫۷۵)
سطح تحصیلات	ـ۱٫۲۱	(ـ۱٫۴۱ـ۱٫۰۰)	ـ۱۱٫۶۲(۰٫۰۰۰)	۰٫۳۰	(۰٫۲۴ـ۰٫۳۷)
سابقه خانوادگی ابتلا (نداشتن)	ـ۰٫۳۵	(ـ۰٫۸۱ـ۰٫۱۱)	ـ۱٫۴۹(۰٫۱۴۰)	۰٫۷۰	(۰٫۴۴ـ۱٫۱۲)
استعمال دخانیات (نکردن)	ـ۰٫۱۴	(ـ۰٫۵۲ـ۰٫۲۴)	ـ۰٫۷۲(۰٫۴۷۰)	۰٫۸۷	(۰٫۶۰ـ۱٫۲۷)
مصرف الکل (نداشتن)	۰٫۴۹	(ـ۰٫۱۰ـ۱٫۰۷)	۱٫۶۴(۰٫۱۰۰)	۱٫۶۳	(۰٫۹۱ـ۲٫۹۳)

مدل رگرسیون کلاس پنهان مورد استفاده در این پژوهش امکان ارزیابی نحوه اثرگذاری متغیرهای مستقل بر تک تک علائم هشدار دهنده سرطان را نیز فراهم می سازد. جدول ۳ ضرایب متغیرهای مستقلی که بر آگاهی از هر یک از علائم هشدار دهنده سرطان مؤثرند را نشان می دهد. در این جدول ، فقط متغیرهای مستقل مؤثر بر شناخت هر یک از علائم هشدار دهنده ذکر شده است و دیگر متغیرها در سطح ۵٪ معنی دار نشده اند.

جدول۳- برآورد اثر مستقیم متغیرهای مستقل بر تک تک علائم هشدار دهنده سرطان

علائم هشدار دهنده سرطان	متغیرهای کمکی	آماره آزمون(p-مقدار)	نسبت بخت
تغییر در عادات دفع ادرار و مدفوع	سن	۳٫۰۴(۰٫۰۰۲)	۱٫۰۱
	جنسیت	۱٫۹۷(۰٫۰۴۹)	۱٫۱۳
	تحصیلات	۲٫۳۳(۰٫۰۲۰)	۱٫۰۹
زخمی که بیش از سه هفته بهبود و ترمیم نیابد	...		
خونریزی یا ترشح غیرمعمول	تحصیلات	۲٫۴۵(۰٫۰۱۴)	۱٫۱۰
سفتی و وجود توده ای در پستان و یا سایر ارگانها	جنسیت	۶٫۴۱(۰٫۰۰۰)	۱٫۶۰
	وضعیت تاهل	۲٫۸۰(۰٫۰۰۵)	۱٫۲۹
	تحصیلات	۴٫۸۷(۰٫۰۰۰)	۱٫۲۵
اشکال در بلع	...		
اشکال در هضم و جذب غذا(سوء هاضمه)	تحصیلات	۲٫۹۸(۰٫۰۰۳)	۱٫۱۲
تغییرات قابل ملاحظه در خال ها و زگیل ها	وضعیت تاهل	۲٫۴۶(۰٫۰۱۴)	۱٫۲۱
	تحصیلات	۵٫۸۵(۰٫۰۰۰)	۱٫۲۷
سرفه های مکرر و یا خشونت و تغییر در صدا	...		
کاهش وزن ناگهانی	جنسیت	۱٫۹۷(۰٫۰۴۹)	۱٫۱۳
	وضعیت تاهل	۲٫۰۴(۰٫۰۴۱)	۱٫۱۶
	مصرف الکل	۲٫۳۳(۰٫۰۲۰)	۱٫۳۰

۴ بحث و نتیجه گیری

در این پژوهش فرم تعمیم یافته ای از مدل رگرسیون کلاس پنهان معرفی گردید و بیان شد که با توجه به ماهیت نظری آن که امکان ارزیابی متغیر پنهان ، نحوه اثرگذاری متغیرهای کمکی بر متغیر پنهان و نیز بر هر یک از متغیرهای پاسخ مشاهده شده که متغیرهای پنهان بر مبنای آنها تشکیل می شود را فراهم می کند می تواند در حوزه پزشکی و بهداشت در مقوله هایی چون بیماریهای روانی ، سنجش نگرشها و از همه مهمتر در تشخیص های پزشکی بسیار مفید واقع شود. در مثالی که برای توضیح کارکرد این روش بیان شد میزان آگاهی یک جمعیت عمومی از علائم هشدار دهنده بیماری سرطان و عوامل مؤثر بر سطح کلی آگاهی و شناخت هر یک از علائم هشدار دهنده ارزیابی گردید.

رگرسیون کلاس پنهان یک روش ناپارامتری است و تنها فرض عمده آن شناسایی پذیر بودن مدل می باشد که در این مطالعه با بیان یک قضیه شرایط کافی برای شناسایی پذیر بودن مدل بیان گردید. برای برآورد پارامترها در رگرسیون کلاس پنهان دو روش عمده ماکزیمم درستنمایی (ML) و روش ماکزیمم پسین[۲](MAP) وجود دارد. برآورد MAP بیزی از ماکزیمم کردن توزیع لگاریتم پسین بدست می آید. بطوریکه توزیع لگاریتم پسین از مجموع تابع لگاریتم درستنمایی و لگاریتم های پیشین برای پارامترها حاصل می شود. گرچه بطور کلی تفاوت بسیار زیادی بین ML و MAP وجود ندارد. مهمترین مزیت روش MAP نسبت به ML این است که از وقوع جوابهای مرزی جلوگیری می کند بعبارت دیگر احتمالات و واریانس صفر نخواهد شد. با کمی اطلاع از توزیع پیشین ، برآوردهای پارامترها در داخل فضای پارامتر قرار خواهند داشت. امروزه ، بیشتر نرم افزارها از الگوریتم EM یا تعدیل یافته آن برای برآورد پارامترهای مدل استفاده می کنند. استفاده از الگوریتم EM باعث یافتن جوابهای بهینه برای پارامترها حتی زمانیکه مقادیر اولیه از جواب بهینه دور باشند می گردد. مهمترین ویژگی ارزشمند این روش پایایی برآوردگرها می باشد.

مدل بیان شده در این مقاله از طریق وارد کردن متغیرهای پیش بین پنهان در جزء پیش بینی کننده مدل در حالت معمولی و نیز در حالت چندسطحی قابل تعمیم می باشند (فیضی و ورمانت ، ۲۰۰۹). با توجه به آنکه در عمل متغیرهای پیش بینی که امکان ارزیابی مستقیم آنها وجود ندارد (از طرق مدلهای عاملی استخراج می گردند) به وفور با آن مواجه می شویم و از طرفی با توجه به حوزه کاربری وسیع داده های چندسطحی ، بسط مدل معرفی شده در مقاله حاضر در چنین ساختاری موجب گسترش چشمگیر حوزه کاربری آن می شود.

مراجع

[1] Agresti A. (2002). *Analysis of Categorical Data*, 2nd Ed., New York:John Wiley and Sons.

[2] Andrews RL, Currim IS. (2003). *A Comparison of segment retention criteria for finite mixture logit models* , journal of Marketing Research, 40, 235-43.

[2] MAximum Posterior method

[3] Banden-Roche K, Miglioretti D, Zeger S, Rathouz P. (1997). *Latent variable regression for multiple discrete outcomes*, journal of the American Statistical Association, 92, 1375-86.

[4] Dayton CM, Macready GB. (1988). *Concomiatant-variable latent class models*, journal of the American Statistical Association, 83, 173-78.

[5] Dias JG. (2004). *Finite Mixture Models: Review, Application, and computer intensive Methods*, The Netherlands: Groningen of University.

[6] Feizi A, Vermunt JK. (2009). *Mixed effects latent class regression with latent factor predictors. Multivariate behavioral research*,(under review).

[7] Formann AK. (1982). *Linear logistic latent class analysis*, Biometrikam, 24(2), 171-90.

[8] Formann AK. (1992). *Linear logistic latent class analysis for polytomous data*, journal of the American Statistical Association, 87, 476-86.

[9] Goodman LA. (1974). *Exploratory latent structure analysis using both identifiable and unidentifiable models*, Biometrikam, 61, 215-31.

[10] Huang GH, Bandeen-Roche K. (2004). *building identifiable latent variable model with covariate effects on underlying and measured variables*, Psychometrika, 69, 5-32.

[11] Huang GH. (2005). *Selection the number of classes under latent class regression: a factor analytic analogue*, Psychometrika, 70, 325-45.

[12] Lazarsfeld PF. (1950). *The logical and mathmatical foundations of latent structure analysis. In: Stouffer SS and et al, editors. Measurement and prediction*, Princeton:Princeton Unversity Press, 362-412.

[13] Lazarsfeld PF, Henry NW. (1968). *Latent Structure Analysis*, Boston:Houghton Mill.

[14] McCullagh P, Nelder JA. (1989). *Generalized Linear Models*, 2nd Ed., London:Chapman and Hall.