مجله تخصصي اپيدميولوژي ايران؛ ۱۳۸۴، دوره ۱، شماره ۴ و ۳: صفحات ۴۵-۴۱.

مقاله پژوهشي

به‌كارگيري متغيرهاي پنهان در مدل رگرسيون لجستيك براي حذف اثر هم‌خطي چندگانه در تحليل برخي عوامل مرتبط با سرطان پستان

محمد امين پورحسينقلي[۱]، يدا... محرابي[۲]، حميد علوي مجد[۳]، پروين ياوري[۴]

[۱]كارشناس ارشد آمار زيستي، دانشگاه علوم پزشكي شهيد بهشتي، تهران.
[۲]دانشيار آمار زيستي، دانشگاه علوم پزشكي شهيد بهشتي، تهران.
[۳]استاديار آمار زيستي، دانشگاه علوم پزشكي شهيد بهشتي، تهران.
[۴]استاد اپيدميولوژي، دانشگاه علوم پزشكي شهيد بهشتي، تهران.

نويسنده‌ي رابط: دكتر يدا... محرابي، گروه پزشكي اجتماعي و بهداشت دانشكده پزشكي، دانشگاه علوم پزشكي شهيد بهشتي، تهران، اوين، كد پستي ۱۹۳۹۵، تلفن: ۸-۲۳۸۷۲۵۶۷-۰۲۱، نمابر: ۲۲۴۱۴۱۰۸-۰۲۱، پست الكترونيك: mehrabi@sbmu.ac.ir ; ymehrabi@gmail.com
تاريخ دريافت: ۸۴/۱۰/۲۶، پذيرش: ۸۵/۲/۳

مقدمه و اهداف: رگرسيون لجستيك يكي از كاربردي‌ترين مدل‌هاي خطي تعميم‌يافته براي تحليل رابطه‌ي يك يا چند متغير توضيحي بر متغير پاسخ رسته‌اي است. زماني كه بين متغيرهاي توضيحي همبستگي‌هاي نسبتا قوي وجود داشته باشد هم‌خطي چندگانه ايجاد شده، ممكن است به كاهش كارآيي مدل منجر شود. هدف اين تحقيق استفاده از متغيرهاي پنهان براي كاهش اثر هم‌خطي چندگانه در تحليل يك مطالعه مورد ـ شاهدي است.
روش كار: داده‌هاي مورد استفاده در اين تحقيق متعلق به يك مطالعه مورد ـ شاهدي است كه در آن ۳۰۰ نفر زن مبتلا به سرطان پستان با ۳۰۰ زن شاهد از نظر عوامل خطر مورد مقايسه قرار گرفتند. براي بررسي اثر هم‌خطي، پنج متغير كمي كه بين آن‌ها همبستگي بالايي وجود داشت، در نظر گرفته شدند. ابتدا مدل لجستيك به متغيرهاي فوق برازش داده شد. سپس به منظور حذف اثر هم‌خطي، دو متغير پنهان با استفاده از هركدام از دو روش تحليل عاملي و تحليل مؤلفه‌هاي اصلي به‌دست آورده، بر مبناي آن‌ها پارامترهاي مدل‌هاي لجستيك مجدداً محاسبه شدند. كارآيي مدل‌ها، با استفاده از خطاي استاندارد پارامترها مقايسه گرديد.
نتايج: مدل رگرسيون لجستيك براساس متغيرهاي اوليه حاكي از مقادير غيرعادي نسبت شانس براي سن در اولين زايمان زنده (۴۵۳۵۰۳ و ۱۰۱۸۴==CI۹۵% و OR=۶۷۹۶۰) و سن در اولين حاملگي (OR=۰/۰۰۰۰۲۹) بود. درحالي كه پارامترهاي مدل‌هاي لجستيك حاصل از متغيرهاي پنهان به دست آمده از هر دو روش تحليل عامل و تحليل مؤلفه‌هاي اصلي، از نظر آماري معني دار (p<۰/۰۰۳) و خطاي استاندارد همه‌ي آن‌ها كوچك‌تر از خطاي استاندارد مربوط به رگرسيون لجستيك معمولي بود. فاكتورها و مولفه‌هاي اصلي توليد شده توسط دو روش حداقل ۸۵ درصد كل واريانس را تبيين كردند.
نتيجه‌گيري: تحقيق نشان داد انحراف استاندارد پارامترهاي برآورد شده در رگرسيون لجستيك براساس متغيرهاي پنهان از رگرسيون لجستيك براساس مشاهدات اوليه كوچكتر بوده و در نتيجه اين‌گونه مدل‌بندي در تحليل برخي عوامل خطر سرطان پستان كه هم‌خطي دارند كارآتر است.
واژگان كليدي: هم‌خطي چندگانه، متغير پنهان، تحليل عاملي، تحليل مؤلفه‌هاي اصلي، رگرسيون لجستيك، سرطان پستان.

مقدمه

رگرسيون لجستيك يكي از كاربردي‌ترين مدل‌هاي خطي تعميم‌يافته است كه براي تحليل رابطه‌ي يك يا چند متغير توضيحي بر متغير پاسخ رسته‌اي به كار مي‌رود (۱). زماني كه تعداد متغيرهاي توضيحي افزايش مي‌يابد، مدل‌سازي مشكل شده و كارآيي آن نيز كاهش مي‌يابد؛ به خصوص اگر برخي از متغيرها علي‌رغم فرض استقلال در مدل‌سازي، با يك‌ديگر هم‌بستگي قوي

داشته باشند و به عبارت دیگر هم‌خطي چندگانه (Multicollinearity) ایجاد شده باشد (۲). هم‌خطي چندگانه یکي از دلایل افزایش خطاي استاندارد برآورد ضرایب رگرسیوني و درنتیجه کاهش کارآیي مدل بوده و ممکن است منجر به پیش‌بیني‌هایي خارج از دامنه‌ي مورد انتظار شود (۳).

مسئله‌ي هم‌خطي در مدل‌هاي رگرسیون خطي مورد توجه بسیاري از محققان قرار گرفته و روش‌هاي گوناگوني براي مقابله با اثرات نامطلوب آن ابداع شده است (۲). از جمله این روش‌ها، کاهش ابعاد مدل با استفاده از متغیرهاي پنهان (Latent Variables) است. این نوع متغیرها مستقیماً مشاهده نمي‌شوند؛ بلکه از ترکیب سایر متغیرهاي مشاهده‌شده قابل دستیابي بوده، به عنوان نماینده‌ي برخي از متغیرهاي همبسته در مدل به‌کار مي‌روند (۴).

اگر چه استفاده از متغیرهاي پنهان براي کاهش ابعاد مدل، در عمل بیشترین کاربرد را در مطالعات مربوط به علوم اجتماعي و روانشناسي داشته است، ولي به دلیل نوع مطالعات انجام شده در علوم پزشکي و بهداشت که مستلزم جمع‌آوري تعداد قابل توجهي متغیرهاي مرتبط با یک‌دیگر است، مشکل هم‌خطي در بسیاري از مدل‌هاي آماري این مطالعات قابل انتظار است (۵) و علي‌رغم این‌که هم‌خطي چندگانه در مدل رگرسیون لجستیک نیز ایجاد مشکل مي‌کند (۶،۷)، تاکنون توجه محققان بیشتر بر رگرسیون خطي با متغیر پاسخ داراي توزیع نرمال متمرکز بوده است.

تحلیل مؤلفه‌هاي اصلي (Principal Component Analysis) یکي از کاربردي‌ترین روش‌هاي کاهش ابعاد در روش‌هاي چند متغیري است. تاریخچه‌ي ابداع این روش به ابداعات پیرسن (Pearson) در برازش حداقل مربعات متعامد برمي‌گردد؛ ولي بسط عمده‌ي تئوري به‌وسیله‌ي هتلینگ (Hotteling) انجام شده است (۸). مؤلفه‌هاي اصلي با توجه به خصوصیاتي که دارند براي مقابله با مشکل هم‌خطي و کاهش ابعاد مدل در رگرسیون‌هاي خطي مورد استفاده قرار مي‌گیرند (۲،۹). در این روش با استفاده از ماتریس مقادیر ویژه (Eigen Values) (۱۰)، مؤلفه‌هاي اصلي به صورت ترکیب خطي از متغیرهاي اولیه و مستقل از یکدیگر ساخته مي‌شوند و در آنالیز داده‌ها، به جاي متغیرهاي اولیه مورد استفاده قرار مي‌گیرند (۱۱).

تحلیل عاملي (Factor Analysis) از دیگر روش‌هاي کاهش ابعاد داده‌ها است، که نخستین بار توسط اسپیرمن (Spearman) معرفي شد. در این روش با فرض وجود یک مدل مبنایي مشخص براي کل داده‌ها، و براساس ماتریس واریانس ـ کوواریانس یا ماتریس ضرایب همبستگي، عامل‌هاي مستقل از یکدیگر، از روي متغیرهاي اولیه به‌دست مي‌آیند (۱۲).

در این تحقیق، براي کاهش ابعاد مدل رگرسیون لجستیک با متغیرهاي توضیحي هم‌خط، در تحلیل داده‌هاي یک مطالعه مورد ـ شاهدی پیرامون عوامل خطرسرطان پستان از دو روش تحلیل عاملي و تحلیل مؤلفه‌هاي اصلي استفاده شده است.

روش‌ها

براي بررسي نحوه‌ي کاربرد دو روش تحلیل مؤلفه‌هاي اصلي و تحلیل عاملي در کاهش ابعاد مدل و ایجاد متغیرهاي پنهان، از داده‌هاي مطالعه‌ي مورد ـ شاهدي مربوط به عوامل خطر سرطان پستان استفاده شد (۱۳). در مطالعه‌ي مذکور که در فاصله‌ي زماني بهمن ۸۲ تا آذر ۸۳ در مرکز پزشکي ـ آموزشي، درماني شهداي تجریش انجام شد گروه مورد، بیماراني بودند که بیماري سرطان پستان آن‌ها با استفاده از آزمایش‌هاي پاتولوژیک، تشخیص قطعي داده شده و یا براي درمان یا پي‌گیري به درمانگاه بیمارستان شهداي تجریش مراجعه کرده بودند. گروه شاهد زناني بودند که به دلایل دیگري غیر از سرطان پستان و به‌طور هم‌زمان در بخش‌هاي دیگرِ بیمارستان شهداء، مثل جراحي، پوست، داخلي و غیره بستري و یا براي پي‌گیري یا درمان به درمانگاه بیمارستان مراجعه کرده بودند و از نظر سني با گروه مورد با حداکثر ۲ سال اختلاف مشابه‌سازي شدند. با اطمینان ۹۵ درصد و توان

آزمون ۸۰ درصد، تعداد نمونه براي هر گروه ۳۰۰ نفر انتخاب شد (۱۴). متغيرهاي مختلفي به عنوان عوامل خطر يا متغير كنترل جمع‌آوري شدند كه در اين مقاله، پنج متغير كه بين آن‌ها هم‌بستگي بالايي وجود داشت در نظر گرفته شدند. اين متغيرها عبارتند از: تعداد حاملگي (NP: Number of Pregnancy)، تعداد فرزندان زنده به دنيا آورده (NLB: Number of Live Birth)، كل طول مدت شيردهي به فرزندان (TLBF: Total Length of Breast Feeding)، سن در اولين حاملگي (AFP: Age at First Pregnancy) و سن در اولين زايمان زنده (AFLB: Age at First Live Birth). براي بررسي ميزان بروز هم‌خطي در اين مشاهدات از ماتريس ضرايب هم‌بستگي استفاده است (۲).

ابتدا بدون در نظر گرفتن وجود هم‌خطي، مدل رگرسيون لجستيك به داده‌ها برازش داده شد. سپس با تركيب پنج متغير مورد بررسي، يك بار با روش تحليل مؤلفه‌هاي اصلي و بار

جدول ۱- ماتريس ضرايب همبستگی متغيرهای توضيحی

	AFLB	AFP	TLBF	NLB	NP
NP	-۰/۴۲	-۰/۴۴	۰/۶۷	۰/۹۰	۱
NLB	-۰/۰۴۷	۰/۴۷	۰/۷۵	۱	
TLBF	-۰/۴۱	-۰/۴۲	۱		
AFP	۰/۹۷	۱			
AFLB	۱				

NP: تعداد حاملگي؛ NLB: تعداد فرزندان زنده به دنيا آورده؛ TLBF: كل طول مدت شيردهي به فرزندان؛ AFP: سن در اولين حاملگي؛ AFLB: سن در اولين زايمان زنده

ديگر به طريق تحليل عاملي، دو متغير پنهان به‌دست آمد و بر اساس آن‌ها، پارامترهاي مدل رگرسيون لجستيك برآورد شد. مدل‌هاي حاصل، براساس واريانس‌هاي تبيين شده به‌وسيله‌ي دو روش و خطاي استاندارد پارامترهاي برآورد شده، مورد مقايسه قرار گرفتند.

يافته‌ها

جدول يك نشان مي‌دهد بين سه متغير تعداد حاملگي، تعداد فرزندان زنده به دنيا آورده و كل طول مدت شيردهي به فرزندان و نيز بين دو متغير سن در اولين حاملگي و سن در اولين زايمان زنده هم‌بستگي خطي بالايي مشاهده مي‌شود. هم‌چنين مقادير عامل تورم واريانس (VIF: Variance Inflation Factor) نيز كه ميزان بروز هم‌خطي را نشان مي‌دهد، محاسبه شد كه همه‌ي مقادير بيشتر از يك و نشان دهنده‌ي وجود هم‌خطي در متغيرهاست. به ويژه اين‌كه دو متغير سن دراولين حاملگي و سن در اولين زايمان زنده داراى VIF بيشتر از ۱۵ بوده و بنابراين براساس اين معيار هم‌خطي شديدي دارند. آزمون بارتلت نيز نشان داد ماتريس ضرايب هم‌بستگي متغيرهاي توضيحي با صفر اختلاف معنی‌داری دارد (P<۰/۰۰۱). بنابراين بين متغيرهاي توضيحي مورد بررسي هم‌خطي چندگانه وجود دارد.

نتايج حاصل از رگرسيون لجستيك بدون در نظر گرفتن وجود اين هم‌خطي (جدول ۲)، نشان مي‌دهد كه فقط دو متغير سن در اولين حاملگي و سن در اولين زايمان زنده معني‌دار شده اند (P<۰.۰۰۱). هم‌چنين نسبت شانس به‌دست آمده براي سن در اولين زايمان زنده بسيار بزرگ (CI۹۵%=۱۰۱۸۴-۴۵۳۵۰۳ و OR=۶۷۹۶۰) و برعكس براى سن در اولين حاملگی بسيار كوچك (OR=۰/۰۰۰۰۲۹) بود كه هردو غيرعادی هستند. در مرحله‌ي بعد با روش تحليل عاملی دو عامل به صورت تركيب خطي زير از متغيرهاي اوليه بدست آمد:

Factor1= ۰/۸۵NP+۰/۹۷NLB+۰/۶۸TLBF-۰/۲۶ AFP-۰/۲۶AFLB

Factor2= ۰/۲۱NP-۰/۲۳NLB-۰/۲۵ TLBF+۰/۹۴AFP+۰/۹۵AFLB

در مجموع ۸۴/۷۹% واريانس توسط عامل اول (۴۵/۶۷%) و عامل دوم (۳۹/۱۲%) تبيين شده است. وارد كردن عوامل فوق به عنوان متغيرهاي توضيحي در مدل رگرسيون لجستيك، نشان داد كه عامل اول (P<۰/۰۰۲) و عامل دوم (P<۰/۰۰۱) هر دو تاثير معني‌داري بر متغير وابسته دارند (جدول ۳). در مرحله‌ي سوم، با روش تحليل مؤلفه‌هاي اصلي دو مؤلفه به صورت تركيب خطي زير از متغيرهاي اوليه بدست آمد:

Component1=۰/۹۰NP+۰/۹۲NLB+۰/۸۲TLBF -۰/۲۵AFP -۰/۲۴AFLB

جدول ٣- برآورد پارامترهاي رگرسيون لجستيك بر اساس متغيرهاي پنهان ايجاد شده به‌وسيله روش تحليل عاملي و مؤلفه‌هاي اصلي

متغيرهاي پنهان	ضرايب رگرسيوني	خطاي استاندارد	P-value	نسبت شانس (فاصله اطمينان ٩٥%)
روش تحليل عاملي				
عرض از مبداء	-٠/٠٤	٠/٠٨	٠/٦٤	٠/٩٦ (-٠/٩١ ٠/٧٦)
عامل ١	-٠/٢٧	٠/٠٩	٠/٠٠٢	٠/٧٦ (-١/٥٩ ١/٣٣)
عامل ٢	٠/٢٩	٠/٠٩	٠/٠٠١	١/٣٣ (١/١٢ ١/٣٣)
روش تحليل مؤلفه هاي اصلي				
عرض از				

Component2=-٠/٢١NP-٠/٢٥NLP-٠/٢٥TLBF+٠/٩٦AFP+٠/٩٦AFLB

مؤلفه‌ي اول ٤٩/٢٦% و مؤلفه‌ي دوم ٤٠% و در مجموع ٨٩/٢٦% كل واريانس را تبيين مي‌كنند. در برآورد پارامترهاي رگرسيون لجستيك براساس اين مؤلفه‌هاي اصلي، نتايج مشابهي حاصل شد. به عبارت ديگر مؤلفه اول با (P<٠/٠٠٢) و مؤلفه دوم با (P<٠/٠٠٣) معني‌دار شدند و نسبت شانس‌ها نيز تقريباً مشابه روش تحليل عاملي به‌دست آمدند (جدول ٣).

بحث

هدف پژوهش حاضر استفاده از متغيرهاي پنهان در مدل رگرسيون لجستيك به منظور كاهش اثر در حالت بروز هم‌خطي چندگانه بود. نتايج حاصل نشان داد برآورد پارامترهاي مدل در هر دو روش تحليل مؤلفه‌هاي اصلي و تحليل عامل، مشابه هستند و خطاي استاندارد آن‌ها نسبت به مدل اصلي بسيار كوچك‌تر است. بنابراين دو روش به‌كار رفته براي توليد متغيرهاي پنهان نسبت به مدل رگرسيون لجستيك براساس مشاهدات اوليه، در تحليل برخي عوامل مرتبط با سرطان پستان ازكارايي بالاتري برخوردار هستند. در مدل اوليه به دليل وجود هم‌خطي و بالا بودن خطاي استاندارد برآوردها، تعدادي از متغيرها معني‌دار نشده‌اند و از بين دو متغير معني‌دار، متغير سن در اولين زايمان زنده، نسبت شانس بسيار بزرگ و غير معمول و متغير سن در اولين حاملگي، نسبت شانسي نزديك به صفر را نشان داد.

يافته‌هاي اين تحقيق با نظر آگويلرا و اسكابياس (Aguilera & Escabias) كه در مقاله‌ي خود نشان داده‌اند استفاده از تحليل مؤلفه‌هاي اصلي در رگرسيون لجستيك با داده‌هاي هم‌خط مي‌تواند برآورد پارامترها را بهبود بخشد، مطابق است (١٥). اين روش در مقاله‌ي كاربردي اسكابياس، آگويلرا و والدراما (Valderrama) كه در مورد مدل‌سازي داده‌هاي هواشناسي بود نيز مورد استفاده قرار گرفته است. آنان در اين مطالعه بر نحوه‌ي انتخاب مؤلفه‌هاي اصلي براي بهبود برآورد پارامترها تأكيد كرده‌اند (١٦).

وال و لي (Wall & Li) در مطالعه اي كه براساس متغيرهاي توصيفي ميزان مرگ و مير بيماري‌هاي تنفسي بود، نتايج

جدول ٢- برآورد پارامترهاي مدل رگرسيون لجستيك بدون درنظر گرفتن هم‌خطي بين متغيرهاي توضيحي

متغيرهاي اصلي	ضرايب رگرسيوني	خطاي استاندارد	P-value	نسبت شانس (فاصله اطمينان ٩٥%)	عامل تورم واريانس (VIF)*
عرض از مبداء	٠/١٠	٠/٢٠	٠/٦٢	٦٧٩٦٠ (-٤٥٣٥٠٣ ١٠١٨٤)	
تعداد حاملگي (NP)	٠/٠٦	٠/٠٩	٠/٥٠	١/٠٦ (٠/٨٩-١/٢٧)	٤/٥٧
تعداد فرزندان زنده به دنيا آورده (NLB)	-٠/١٧	٠/٢٦	٠/٥١	٠/٨٤ (٠/٥٠-١/٤٠)	٥/٨٣
كل طول مدت شيردهي به فرزندان (TLBF)	-٠/٢٢	٠/١٥	٠/١٤	٠/٨٠ (٠/٥٩-١/٠٨)	٢/١٣
سن در اولين حاملگي (AFP)	-١٠/٤٦	٠/٩٢	<٠/٠٠١	٠/٠٠٠٠٢٩	١٥/٤١
سن در اولين زايمان زنده (AFLB)	١١/١٣	٠/٩٧	<٠/٠٠١	٦٧٩٦٠ (-٤٥٣٥٠٣ ١٠١٨٤)	١٥/٣٤

حاصل از دو روش تحليل عامل و معادله‌ي مدل‌سازي ساختاري (Structural Equation Modeling) را با رگرسيون كلاسيك براساس متغيرهاي اوليه هم‌خط مقايسه كرده نشان دادند كه متغيرهاي پنهان، پارامترهايي با خطاهاي استاندارد كوچكتر توليد مي‌كنند (١٧). نتايج تحقيق حاضر از اين لحاظ با مطالعه‌ي آنان هم‌خواني دارد.

ايده‌ي استفاده از متغيرهاي پنهان به‌جاي متغيرهاي اصلي، با هدف كاهش ابعاد داده‌ها از اين حقيقت ناشي مي‌شود كه اين متغيرها مي‌توانند بازتاب‌دهنده‌ي ارتباط بين مشاهدات باشند (١٨). با اين حال استفاده از مدل‌هاي دربرگيرنده‌ي متغيرهاي پنهان تبعاً مزايا و محدوديت‌هايي دارد. يكي از اهداف اصلي در ساختن مدل‌هاي آماري تفسير مدل با توجه به پارامترهاي برآوردشده مي‌باشد؛ ولي تفسير مدل‌هايي كه براساس عامل‌ها يا تحليل مؤلفه‌هاي اصلي به‌دست مي‌آيند قدري پيچيده است (١٩،١١). براي اين كار استفاده از روش تحليل مؤلفه‌هاي اصلي بهتر از تحليل عاملي است؛ زيرا مؤلفه‌هاي اصلي صرفاً تركيبي خطي از متغيرهاي اوليه هستند و بر خلاف روش تحليل عامل، مدلي براي داده‌ها فرض نمي‌كنند (١٩،٢٠،١١). در نتيجه از طريق معكوس ماتريس دوران مي‌توان برآوردهاي تصحيح شده پارامترهاي متغيرهاي اوليه را به‌دست آورد (١١). هم‌چنين روش‌هايي نيز براي تفسير اين مؤلفه‌ها در مدل كاهش‌يافته پيشنهاد شده است (٢١). به هر حال سودمندي‌هاي حاصل از كاهش ابعاد مدل و كاستن تعداد متغيرها آن‌چنان قابل ملاحظه است كه علي‌رغم مشكلات حاصل در تفسير پارامتر، برخي تكنيك‌هاي جديد علاوه بر توليد متغيرهاي پنهان براي متغيرهاي توضيحي، اكنون بر توليد اين متغيرها براي متغيرهاي پاسخ توجه دارند (٢٢).

نتيجه‌گيري

براساس يافته‌هاي اين تحقيق مي‌توان نتيجه‌گيري كرد كه در بررسي برخي عوامل خطر سرطان پستان، دو روش تحليل عاملي و تحليل مؤلفه‌هاي اصلي نتايج مشابهي داشته، نسبت به مدل لجستيك با متغيرهاي هم‌خط اوليه از كارآيي بالاتري برخوردار هستند.

تشكر و قدرداني

در اين مقاله از داده‌هاي طرح تحقيقاتي عوامل خطر سرطان پستان، مصوب دانشگاه علوم پزشكي شهيد بهشتي استفاده شده است كه به اين وسيله از كليه همكاران طرح مذكور و نيز از معاونت پژوهشي دانشكده پزشكي سپاسگزاري به عمل مي آيد.

منابع

1. Myers R.H., Montgomery D.C. and Vining G.G., Generalized linear models with application in engineering and sciences, 2002, John Wiley & Sons.
2. Chattergee, S., Hadi, A.S. and Price, B. (2000). Regression analysis by example, 2002, John Wiley & Sons, USA. PP: 225-258.
3. Myers, R.H. (1990). Classical and modern regression with applications., 1990, Pws-Kent publishing company. PP: 123-129.
4. Van Eye, A., Clogg, C.C., Latent variables analysis; application for developing research. 1994, SAGE publication. PP: 3-35.
5. Hazard munro, B. Statistical methods for health care research., 2001, Philadelphia: Lippincott. PP: 287-288.
6. Kleinbaum, D. Logistic Regression., 1994, Springer, New York. PP: 168.
7. Hosmer, D.W., Lemeshow, S. Applied logistic regression., 1989, John Wiley & Sons.
8. Morrison, D. F. Multivariate statistical methods. 2002, John Wiley & Sons. PP: 312-398.
9. Rawlings, J. O. Applied regression analysis: A research tools., 1988, Belmont: Wadsworth. PP: 327-356.
10. Schott, J. R. Matrix analysis for statistics., 1997, John Wiley & Sons. PP: 84-131.
11. Jollife, I.T. PrinCl95%pal component analysis., 1986, Springer. PP: 129-141.
12. Srivastava, M. S. Methods of multivariate statistics, 2002, John Wiley & Sons. New York. PP: 397-450.
13. Yavari, P., Mousavizadeh, M., Sadrol-Hafezi, B. and Mehrabi, Y., Reproductive characteristics and the risk of breast cancer, A case-control study. Asian PaCl95%fic J Cancer Prev, 2005, 6, 370-375.
14. Lemeshow, S., Hosmer, D. W. and Klar, J. Adequacy of sample size in Health studies. World Health Organization, 1998, John Wiley & Sons. PP: 19.
15. Aguilera, A.M. and Escabias, M., PrinCl95%pal component logistic regression. Proceedings in computational statistics, 2000, 175-180. Physica-Verlag.
16. Escabias, M., Aguilera, A. M. and Valderrama, M. J., Modeling climatological data by functional logistic regression. The ISI International Conference on Environmental Statistics and Health, 2003.
17. Wall. M. M. and Li, R., Comparison of multiple regression to two latent variable techniques for estimation and prediction. Statistics in Medicine; 2003, 22:3671-3685.
18. Sobel, M. E. Causal inference in latent variable models. In Latent variables analysis; application for developing research. By Van Eye, A., Clogg, 1994, C.C SAGE publication. PP: 3-35.

2002, http://ace.acadiau.ca/math/chipmanh/publications.html.
22. Guo, J., Wall, M. M. and Amemiya Y. Latent class regression on latent factors to appear in Biostatistics .
19. Rencher, A. C. Methods of multivariate analysis, 2002, John Wiley & Sons.
20. Armitage, P. and Colton, T., Encyclopedia of Biostatistics. Volume 2. Chichester: 1998, John Wiley & Sons. PP: 1480-1481.
21. Chipman HA and Gu H. Interpretable dimension reduction.