مجله دانشگاه علوم پزشکی خراسان شمالی
۱۳۹۰؛ ۳ (ویژه نامه آمار زیستی و اپیدمیولوژی):۱۳-۷

Journal of North Khorasan University of Medical Sciences
2011;3(Biostatistics and Epidemiology Supplement):7-13

مقایسه برآورد نسبت شانس حاصل از برازش مدل رگرسیون لجستیک با سه روش استقلال ، حاشیه ای و اثرات تصادفی در مطالعات موردی–شاهدی همسان‌سازی شده با استفاده از شبیه سازی

حبیب اله اسماعیلی[۱]، مریم سالاری[۲]، آزاده ساکی[۳]، بهزاد قلی زاده[۴]، مصطفی بسکابادی[۴]، حسین لشکردوست[۵]*

۱ دانشیار گروه آمار زیستی و اپیدمیولوژی و عضو مرکز تحقیقات علوم بهداشتی دانشکده بهداشت ، دانشگاه علوم پزشکی مشهد
۲ دانشجوی کارشناسی ارشد آمار زیستی دانشکده بهداشت دانشگاه علوم پزشکی مشهد
۳ استادیار گروه آمار زیستی و اپیدمیولوژی دانشکده بهداشت، دانشگاه علوم پزشکی جندی شاپور اهواز
۴ کارشناس ارشد آمار ریاضی دانشگاه فردوسی مشهد
۵ کارشناس ارشد اپیدمیولوژی، دانشگاه علوم پزشکی خراسان شمالی
***نویسنده مسئول:** اهواز، دانشگاه علوم پزشکی جندی شاپور اهواز، دانشکده بهداشت، گروه آمارزیستی واپیدمیولوژی
پست الکترونیک:saki-a@ajums.ac.ir

چکیده

زمینه و هدف: *یکی از متداول‌ترین مطالعات در حیطه علوم پزشکی جهت یافتن ریسک فاکتورها و عوامل مرتبط با بیماری‌ها مطالعات موردی – شاهدی هستند که شاخص مهم قابل محاسبه درآن OR یا خطرنسبی است. اما در این بین بعضی عوامل مخدوش‌گرکه بر پاسخ موثرند اعتبار OR به دست آمده را زیر سوال می‌برند و OR را کمتر یا بیشتر نشان می‌دهد یکی از روش‌های حذف اثر مخدوش‌گر، طراحی مطالعات matching است. رگرسیون لجستیک یکی از روش‌های متداول مدل‌سازی برای این نوع مطالعات است که در مطالعه حاضر سه روش رگرسیون لجستیک در حالت استقلال ،حاشیه ای و شرطی با هم مقایسه می شوند.*

مواد و روش کار: *این مطالعه بر روی داده های شبیه سازی شده وابسته به هم انجام پذیرفته است. بدین ترتیب که داده ها از توزیع نرمال دو متغیره با ضریب همبستگی‌های (۰.۸،۰.۶،۰.۴،۰.۲،۰) تولید می‌شوند. سپس با انتخاب نقاط برش (۰.۲۵،۰.۲۵) ، (۰.۲۵،۰.۱۵) ، (۰.۲۵،۰.۱) ، (۰.۲۵، ۰.۰۵) برای تابع احتمال تجمعی آنها این داده‌ها که از توزیع پیوسته هستند به توزیع گسسته صفر و یک که به هم وابسته هستند تبدیل می‌شوند. سپس سه مدل رگرسیون لجستیک در حالت استقلال ، حاشیه ای و شرطی به داده‌ها برازش داده می شود و OR محاسبه می شود. با۱۰۰۰۰بار تکرار مقدار صدک ۲.۵ و ۹۷.۵ و همچنین میانه OR سه مدل در نقاط برش ذکرشده با هم مقایسه می شوند .*

یافته ها: *در همبستگی صفر هر سه مدل OR مشابه دارند و تغییر درنقاط باز هم ضرایب مشابه دارد. اما با افزایش میزان همبستگی بین مشاهدات OR بین مدل حاشیه ای و استقلال متفاوت نیست ولی مقدار آن با مدل شرطی متفاوت خواهد بود. به عنوان مثال در نقطه برش (۰.۲۵،۰.۱) و ضریب همبستگی ۰.۶ میانه OR به دست آمده در مدل استقلال و حاشیه ای ۲.۸ است ولی در مدل شرطی این مقدار ۵ یعنی دوبرابر مقدار برازش شده است.*

نتیجه گیری: *استفاده ازمدل های شرطی زمانی که همبستگی بین مشاهدات زیاد است قطعا روش صحیح‌تری است و هرچه میزان این همبستگی بالا برود میزان خطای ما در استفاده از مدل استقلال یا حاشیه ای بالا می رود. اما زمانی که همبستگی بین مشاهدات ناچیز است استفاده از سه مدل برآوردهای یکسانی می دهد.*

واژه های کلیدی: *مطالعات موردی شاهدی ، مدل لجستیک حاشیه ای، مدل لجستیک شرطی، شبیه سازی ، مخدوش‌گر*

مقدمه

اساسا برای آزمون فرضیه ها درباره سبب شناسی بیماری ها دو روش وجود دارد: تجربی و مشاهده ای. در روش تجربی پژوهشگران اثر تغییرات عاملی را که در اختیار دارند مطالعه می کنند. برای مثال ممکن است پژوهشگری یک گروه موش همزاد را به طور تصادفی انتخاب و نیمی ازآنها را در معرض ماده ای که تصورمی رود سرطان زاست قرار می دهند و سپس فراوانی سرطان را در هر دو گروه ثبت می کنند. روش معمول‌تر این است که پژوهشگر تنها می تواند رویداد بیماری را در گروه های مردم که بر اساس تجربه مواجهه (مانند متاهل ها در مقابل مجردها یا سیگاری ها در مقابل غیرسیگاریها) از یکدیگر جداشده اند را مشاهده می‌کند .مشکل اساسی در مطالعات مشاهده ای این است که معمولا گروه های مشاهده شده علاوه برعامل خاص تحت مطالعه از جانب پاره ای خصوصیات دیگر با هم متفاوتند. به دلیل این عوامل مخدوشگر که اغلب غیرقابل اندازه گیری هستند نمایش نقش عامل تحت بررسی مشکل‌تر می شود. یکی از انواع مطالعات علت شناسی در پزشکی مطالعات موردی شاهدی است[۱].

مطالعات موردی شاهدی یکی از انواع مطالعات گذشته نگر هستند. در این مطالعات افرادی که تشخیص داده شده اند بیمار هستند (موارد) باافرادی که بیمار نیستند (شاهدها) مقایسه می شوند. درطراحی این مطالعات روشی که می توان اثرمخدوشگر را ازبین برد بسیارمهم است . جورکردن یا همسان سازی موارد با شاهدها از نظر عوامل مخدوشگر معمول ترین روش است. جورکردن عبارت است از انتخاب شاهدها به نحوی که از لحاظ پاره ای مشخصات اختصاصی با موارد مشابه باشند. موارد و شاهدها را ممکن است تک تک با هم جورکرد. سن ،جنس و نژاد معمول ترین متغیرهایی هستند که در جورکردن به کارگرفته می شود.یکی از شاخص های مهم برای تعیین میزان اهمیت عوامل خطر در بیماری ها محاسبه می شود خطرنسبی (OR) است. [۲،۳]. خطرنسبی (OR) یک اندازه خطر است که نشان دهنده آن است که چقدر شانس کسی که با یک عامل خطر مواجهه داشته ازکسی که مواجهه نداشته بیشتر است. خطرنسبی برای متغیر پاسخ گسسته هم از طریق جداول توافقی و هم از روش مدل سازی محاسبه کرد. در این مقاله روش مدل سازی موردنظر است[۴،۵].

روش کار

۱- داده ها: در این مقاله بجای استفاده از داده های واقعی از داده های شبیه سازی شده استفاده شده است. شبیه‌سازی فرآیندی است برای پاسخ دادن به سؤالات مربوط به دنیای واقعی، به کمک انجام آزمایش‌هایی که خیلی شبیه وضعیت‌های واقعی هستند. برای حل مسائل دو راه استفاده از روشهای ریاضی و احتمالی و همچنین روش شبیه سازی استفاده می شود.

۲- تولید داده ها به روش شبیه سازی: ابتدا برای تولید داده های دودویی همبسته به هم به روش زیر اجرا شد. نکته قابل توجه این است که متغیر مستقل یا همان Xها شبیه سازی می شوند به دلیل اینکه مطالعه موردی شاهدی است و متغیر Y از قبل مشخص است که برای گروه مورد برابر یک و برای گروه شاهد برابر صفر است. نمونه هایی به حجم۵۰ از توزیع نرمال دو متغیره با ضریب همبستگی ها در ۵ حالت (۰،۰.۲،۰.۴،۰.۶،۰.۸) تولید می شود. یعنی ضریب همبستگی ۰.۲ افزایش پیدا می کند. از آنجایی که توزیع نرمال یک توزیع پیوسته است برای تبدیل آن به توزیع گسسته صفر و یک از تابع توزیع تجمعی استفاده می شود. یعنی نقاط برش مختلفی برای تبدیل این داده ها به صفر و یک استفاده می شود. دلیل دیگر این است که در مطالعات موردی شاهدی به عنوان مثال اهمیت یک ریسک فاکتور را باگذاشتن نقاط برش مختلف کاهش یا افزایش می دهیم حتی ممکن است متغیر مستقل به عنوان عامل حفاظتی درنظر گرفته شود. نقاط برش برای دوگروه مورد شاهد به ترتیب به صورت (۰.۲۵،۰.۲۵) ، (۰.۲۵،۰.۱۵) ، (۰.۲۵،۰.۱) ، (۰.۲۵، ۰.۰۵) گرفته می شود یعنی اگر تابع توزیع تجمعی درگروه مورد از۰.۲۵کمتر باشد عدد صفر و اگر از ۰.۲۵ بیشتر باشد عدد یک درنظرگرفته می شود به این ترتیب ۵۰ عدد صفر و یک تولید خواهد شد. انتخاب گروه مورد و شاهد هم از بین دو دسته داده برای بار اول به صورت تصادفی خواهد بود و تا انتهای شبیه سازی ثابت خواهد ماند.

مدل سازی آماری: دراین مرحله سه مدل رگرسیون لجستیک درحالت استقلال ،حاشیه ای(GEE) و شرطی برازش داده می شود. نکته بسیار مهم در این مرحله این است که درحالت استقلال هیچ شرطی در مورد ارتباط گروه مورد و شاهد نداریم ولی در دو حالت دیگر حتما هرکنترل که با هرمورد همسان سازی شده به صورت یک خوشه یا یک طبقه درنظر گرفته شود .برای این منظور با یک تغییر درنحوه چیدمان داده های شبیه سازی شده داده ها به صورت زیر خواهند بود:

جدول۱: نحوه خوشه بندی مشاهدات برای برازش مدل شرطی

ردیف	شماره خوشه	مورد=۱شاهد=۰	مواجهه=۱عدم مواجهه=۰
۱	۱	۱	۱
۲	۱	۰	۱
۳	۲	۱	۰
۴	۲	۰	۰
....	.....	...	....
۹۹	۵۰	۱	۰
۱۰۰	۵۰	۰	۱

زمانی که متغیر پاسخ در یک مطالعه دارای توزیعی غیر از نرمال باشد همانند توزیع برنولی، دوجمله ای و یا پواسن از مدلهای خطی تعمیم یافته که تعمیمی از مدلهای خطی همانند رگرسیون می باشند استفاده می شود.

مدل رگرسیون لجستیکزمانی که متغیر پاسخ دوجمله ای (پیروزی/شکست) باشد رایج ترین مدل خطی تعمیم یافته مدل رگرسیون لجستیک است. در مطالعات مورد- شاهدی متغیر پاسخ ما دوجمله ای است: بیماری/ عدم بیماری که آن را با صفر و یک نشان می دهیم.

درواقع Y یک متغیرتصادفی برنولی با $P(y_i=1) = \pi_i$ می باشد. که مدل رگرسیون لجستیک یک رابطه خطی بین X و لگاریتم شانس رخداد پیشامد بیماری $\left(\frac{\pi_i}{1-\pi_i}\right)$برای فرد iام برقرار میکند و به صورت زیر نوشته میشود.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta X_i$$

هرچند مدل های لجستیک برای مطالعات موردی شاهدی قابل استفاده هستند ولی یک محدودیت بزرگ دارند و آن هم پیش بینی خطر برای یک فرد را نمی توانند انجام دهند. درواقع فقط خطرنسبی برای این مطالعات قابل محاسبه است.که درآن e^{β} همان خطرنسبی OR است. برای برآورد ضرایب رگرسیونی از روش معادلات درستنمایی استفاده می شود که با فرض مستقل بودن مشاهدات معادلات درستنایی نوشته می شوند به عنوان

مثال در مدل لجیت لگاریتم تابع درستنمایی به صورت رابطه ۱خواهد بود.

دو مدل رگرسیون لجستیک حاشیه ای و مدل رگرسیون لجستیک با اثرات تصادفی (شرطی) به طور معمول برای رگرسیون لجستیک برای مطالعات موردی شاهدی همسان سازی شده یعنی زمانی که مشاهدات به هم وابسته هستند استفاده میشود.

۲- مدل رگرسیون لجستیک حاشیه ای: مدلهای حاشیه ای مدل هایی هستندکه تاثیر متغیرمستقل روی پاسخ ها را به طور مجزا از همبستگی بین پاسخ ها برای یک آزمودنی معین مدل بندی می کند. در این مدلها امید حاشیه ای $E(Yij)$ به صورت تابعی از متغیرهای توضیحی مدل بندی می شود.

یک مدل حاشیه ای دارای اجزا زیر است:

۱- امید حاشیه ای پاسخ – $\mu_{ij}= E(Y_{ij})$ – به متغیرهای توضیحی – x_{ij}– به صورت $h(\mu ij) = xij\beta$ وابسته است که درآن hتابع پیوند شناخته شده ای است. به عنوان مثال برای پاسخ های دودویی تابع لگاریتم است.

۲- واریانس حاشیه ای به میانگین حاشیه ای به صورت Φ $var(y_{ij}) = v(\mu_{ij})$ وابسته است که v تابع واریانس شناخته شده ای است و Φ پارامتر مقیاسی است که ممکن است علاوه بر سایر برآوردها نیاز به برآورد داشته باشد.

$$L(\beta) = \exp\{L(\beta; y, x)\} = \sum_j(\sum_i x_{ij} y_i)\beta_j - \sum_{i=1}^{N} \log\{1 + \exp(\sum_j \beta_j x_{ij})\}$$ رابطه۱

۳- همبستگی yij و yik به صورت تابعی از میانگین حاشیه ای آنها و پارامترهای اضافی α است یعنی

$$\mathrm{corr}(y_{ij}, y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \alpha)$$

که درآن ρ تابع شناخته شده ای از μi هاست.

می توان گفت مدلهای حاشیه ای برای پاسخ های وابسته تعمیمی از GLM ها برای پاسخ های مستقل هستند. یکی ازمدلهای معروف حاشیه ای معادلات برآورد تعمیم یافته (GEE) است که بوسیله Liang و zeger درسال ۱۹۸۶ معرفی شد.

رهیافت معادلات برآورد تعمیم یافته (GEE) تابعی از امید حاشیه ای متغیر وابسته به صورت تابعی از متغیرهای کمکی مشخص می شود همچنین فرض می شود که واریانس تابعی معلوم از میانگین حاشیه ای است. به علاوه ماتریس همبستگی عملی برای مشاهدات هر آزمودنی معین می شود.

ب- مدل اثرات تصادفی (شرطی)

در این مدل فرض می شود ، مدل منحصر به یک آزمودنی خاص است. در این صورت ضرایب نشان دهنده تغییرات پاسخ در کلاس مورد نظر به ازاء یک واحد تغییر در متغیر پیشگوست. یا چنین تصور کنیم که ضرایب نشان دهنده میزان تغییر در داخل آزمودنی است وقتی متغیر پیشگو یک واحد تغییر می کند(وقتی بقیه ثابت هستند). در محاسبه تابع درستنمایی این مدلها فرض میشود که با در نظر گرفتن اثر تصادفی ui مقادیر yij در گروه مورد و شاهد از هم مستقل هستند. در روش مدل سازی شرطی، اطلاعات حاشیه ای را می توان با میانگین‌گیری بر روی همه کلاس ها به دست آورد. مدل GEE فقط مدل حاشیه ای را برازش می دهد در حالی که مدل شرطی قابلیت برازش مدل شرطی را هم دارد.

یافته ها

بااستفاده ازنرم افزار R ورژن۱۵ فرایند تولید داده تصادفی ۱۰۰۰۰ بار با تغییر ضریب همبستگی و نقاط برشی که در بالا ذکرشد تکرار شد و در هر تکرار، پس از برازش مدل ها مقادیر OR از رابطه $\mathrm{EXP}(\beta)$ برای هر مدل به دست آمد. سپس میانه مقادیر OR در ۱۰۰۰۰ بار تکرار برای هر مدل به همراه صدک ۲.۵ و صدک ۹۷.۵ که یک فاصله ۹۵ درصدی را شامل می‌شود محاسبه گردید. زمانی که همبستگی بین مشاهدات صفر است همانطورکه درجدول۴ ملاحظه می شود هر سه مدل شرطی ، حاشیه ای و استقلال برای OR مقادیرت مشابهی را نشان می دهند. یعنی اگرمشاهدات مستقل باشند تفاوتی در ضرایب مدل رگرسیونی و به همین ترتیب در خطرنسبی مشاهده درسه مدل نخواهیم کرد. بازه اطمینان OR در نقطه برش $(0.25,0.25)$ یعنی زمانی که مواجهه درگروه مورد و شاهد یکسان است عدد یک را شامل می شود که همانطورکه انتظارمی رود نشان دهنده این است که متغیر مورد بررسی یک عامل خطر نیست. اما در نقطه برش $(0.25,0.15)$ هرچند بازه شامل یک می شود ولی میانه فاصله عدد ۱.۹ در مدل استقلال و حاشیه ای و ۲ در مدل شرطی است . همینطورکه احتمال مواجهه درگروه مورد بالا می رود میانه OR از عدد یک در هر سه مدل دور می شود .در همبستگی ۰.۲ مقادیر OR در سه مدل کمی با هم متفاوت می شوند . درنقطه برش $(0.25,0.25)$ هرچند هر سه مدل میانه یک را نشان می دهند. ولی صدک ۲.۵ و ۹۷.۵. در مدل استقلال و حاشیه ای شبیه هم بین (۲.۴،۰.۴) ولی در مدل شرطی بین (۳،۰.۳۳) است. همینطورکه نقاط برش را تغییرمی دهیم OR از یک دور می شود و فاصله اطمینان هم در سه مدل متفاوت تر می شود.با بالا رفتن ضریب همبستگی به عنوان مثال در عدد۰.۶ تفاوت در برازش سه مدل آشکارتر می شود. به عنوان مثال در نقطه برش (۰.۲۵ ، ۰.۱) میانه OR برای دو مدل استقلال و حاشیه ای عدد ۲.۸۴ ولی در مدل شرطی ۵ است. یعنی زمانی که همبستگی بین مشاهدات اضافه می شود OR درمدل حاشیه ای و استقلال شبیه هم اما با مدل شرطی بسیار متفاوت خواهد بود.

بحث

هدف این مطالعه آشکارکردن میزان خطایی بود که زمانی که مشاهدات به هم وابسته هستند در انتخاب مدل استقلال یا حاشیه ای خواهیم داشت بود. همانطورکه در بالا ذکر شد استفاده از مدل های شرطی زمانی که همبستگی بین مشاهدات زیاد است قطعا روش صحیح تری است و هرچه میزان این همبستگی بالا برود میزان خطای ما در استفاده از مدل استقلال یا حاشیه ای بالا می رود. اما زمانی که همبستگی بین مشاهدات ناچیز است استفاده از سه مدل برآوردهای یکسانی می دهد. تغییر در نقاط برش زمانی نیز تاثیر زیادی در مقدار OR در هر سه مدل دارند. زمانی که ضریب همبستگی زیاد می شود تغییر در میانه‌ها و فاصله اطمینان را نشان میدهد. هرچند مدل حاشیه ای درگروه مدلهایی قرار می گیردکه همبستگی بین مشاهدات را لحاظ می کند ولی چون روی میانیگن حاشیه ها مدل بندی می‌کند ، ضرایب را بهبود نمی بخشد.

جدول ۲: مقایسه OR حاصل از سه مدل استقلال ، حاشیه ای و شرطی در نقاط برش و ضریب همبستگی های متفاوت

نقاط برش	مدل حاشیه ای			مدل استقلال			مدل شرطی		
	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵
(۰/۲۵,۰/۲۵)	۰/۳۹	۱/۰۰	۲/۵۸	۰/۳۹	۱/۰۰	۲/۵۸	۰/۳۶	۱/۰۰	۲/۷۵
(۰/۲۵,۰/۱۵)	۰/۷۰	۱/۹۴	۶/۰۰	۰/۷۰	۱/۹۴	۶/۰۰	۰/۷۰	۲/۰۰	۷/۰۰
(۰/۲۵,۰/۱)	۱/۰۰	۳/۱۴	۱۱/۲۹	۱/۰۰	۳/۱۴	۱۱/۲۹	۱/۰۰	۳/۰۰	۱۳/۰۰
(۰/۲۵,۰/۰۵)	۱/۸۳	۶/۰۰	۲۳/۰۶	۱/۸۳	۶/۰۰	۲۳/۰۶	۱/۷۵	۶/۰۰	۱۷/۰۰

ρ =0.2

نقاط برش	مدل حاشیه ای			مدل استقلال			مدل شرطی		
	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵
(۰/۲۵,۰/۲۵)	۰/۴۰	۱/۰۰	۲/۴۵	۰/۴۰	۱/۰۰	۲/۴۵	۰/۳۳	۱/۰۰	۳/۰۰
(۰/۲۵,۰/۱۵)	۰/۷۲	۱/۸۸	۵/۵۰	۰/۷۲	۱/۸۸	۵/۵۰	۰/۷۰	۲/۰۰	۹/۰۰
(۰/۲۵,۰/۱)	۱/۰۰	۲/۸۹	۱۰/۷۶	۱/۰۰	۲/۸۹	۱۰/۷۶	۱/۰۰	۳/۳۳	۱۴/۰۰
(۰/۲۵,۰/۰۵)	۱/۸۷	۵/۵۰	۲۱/۰۰	۱/۸۷	۵/۵۰	۲۱/۰۰	۲/۰۰	۶/۵۰	۱۶/۰۰

ρ =0.4

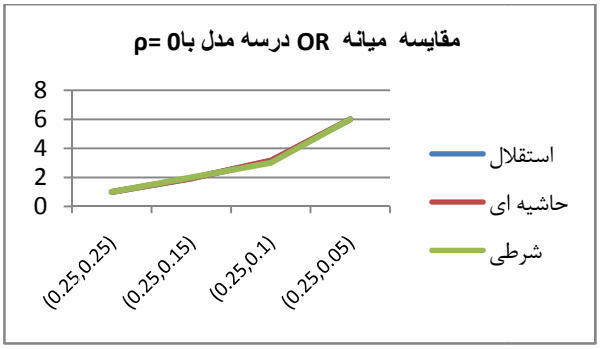
نقاط برش	مدل حاشیه ای			مدل استقلال			مدل شرطی		
	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵
(۰/۲۵,۰/۲۵)	۰/۴۴	۱/۰۰	۲/۲۵	۰/۴۴	۱/۰۰	۲/۲۵	۰/۳۰	۱/۰۰	۳/۰۰
(۰/۲۵,۰/۱۵)	۰/۷۹	۱/۸۸	۵/۰۶	۰/۷۹	۱/۸۸	۵/۰۶	۰/۷۵	۲/۳۳	۱۰/۰۰
(۰/۲۵,۰/۱)	۱/۱۷	۲/۸۸	۹/۳۳	۱/۱۷	۲/۸۸	۹/۳۳	۱/۲۰	۴/۰۰	۱۴/۰۰
(۰/۲۵,۰/۰۵)	۱/۹۴	۵/۲۷	۱۹/۰۶	۱/۹۴	۵/۲۷	۱۹/۰۶	۲/۲۵	۸/۰۰	۱۶/۰۰

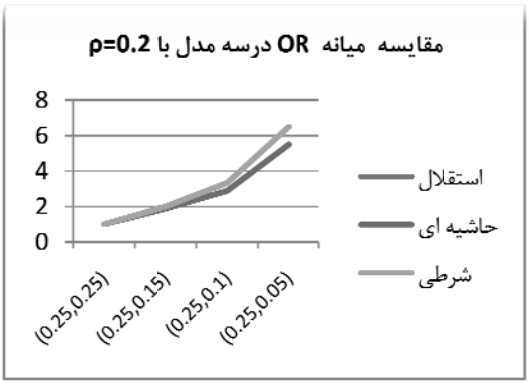
ρ =0.6

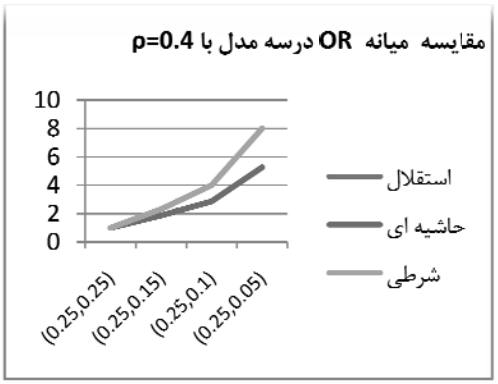
نقاط برش	مدل حاشیه ای			مدل استقلال			مدل شرطی		
	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵
(۰/۲۵,۰/۲۵)	۰/۴۷	۱/۰۰	۲/۰۶	۰/۴۷	۱/۰۰	۲/۰۶	۰/۲۵	۱/۰۰	۳/۶۷
(۰/۲۵,۰/۱۵)	۰/۸۵	۱/۸۴	۴/۵۷	۰/۸۵	۱/۸۴	۴/۵۷	۰/۷۵	۲/۶۷	۱۱/۰۰
(۰/۲۵,۰/۱)	۱/۲۳	۲/۸۴	۷/۹۸	۱/۲۳	۲/۸۴	۷/۹۸	۱/۴۰	۵/۰۰	۱۴/۰۰
(۰/۲۵,۰/۰۵)	۱/۹۸	۴/۹۵	۱۷/۲۲	۱/۹۸	۴/۹۵	۱۷/۲۲	۲/۶۷	۹/۰۰	۱۶/۰۰

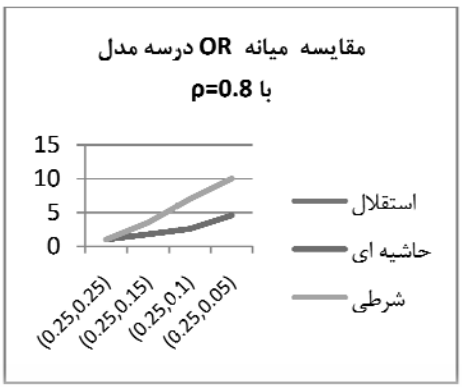
ρ =0.8

نقاط برش	مدل حاشیه ای			مدل استقلال			مدل شرطی		
	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵	OR۰/۰۲۵	OR۰/۵	OR۰/۹۷۵
(۰/۲۵,۰/۲۵)	۰/۵۴	۱/۰۰	۱/۸۳	۰/۵۴	۱/۰۰	۱/۸۳	۰/۱۷	۱/۰۰	۵/۰۰
(۰/۲۵,۰/۱۵)	۱/۰۰	۱/۷۷	۳/۸۶	۱/۰۰	۱/۷۷	۳/۸۶	۱/۰۰	۳/۵۰	۱۱/۰۰
(۰/۲۵,۰/۱)	۱/۳۱	۲/۵۸	۶/۷۷	۱/۳۱	۲/۵۸	۶/۷۷	۲/۰۰	۷/۰۰	۱۳/۰۰
(۰/۲۵,۰/۰۵)	۱/۹۸	۴/۵۷	۱۵/۴۷	۱/۹۸	۴/۵۷	۱۵/۴۷	۳/۵۰	۱۰/۰۰	۱۶/۰۰









نمودار۱: مقایسه OR حاصل از سه مدل استقلال ، حاشیه ای و شرطی در نقاط برش و ضریب

نتیجه گیری

استفاده ازمدل های شرطی زمانی که همبستگی بین مشاهدات زیاد است قطعا روش صحیح‌تری است و هرچه میزان این همبستگی بالا برود میزان خطای ما در استفاده از مدل استقلال یا حاشیه ای بالا می رود. اما زمانی که همبستگی بین مشاهدات ناچیز است استفاده از سه مدل برآوردهای یکسانی می دهد.

تشکر و قدر دانی

این مقاله قسمتی از پایان نامه کارشناسی ارشد رشته آمار زیستی می باشد و نویسندگان مقاله از معاونت محترم پژوهشی دانشگاه علوم پزشکی مشهد که حمایت مالی آن را برعهده داشته اند، تشکر و قدردانی می نمایند .

References

1- Agresti, A, Categorical Data Analysis, 2nd edn. NewYork:Wiley , 2002.
2 - Breslow, N. and N. E. Day. Statistical Methods in Cancer Research, Vol. I: The Analysis of Case–Control Studies. Lyon: IARC. 1980.
3 - Aviva P.,medical statistics at a glance,blackwell science , 2000
4 - Longholz B, Conditional logistic analysis of case-control studies with complex sampling,Biostatistics ,2001;63-84.
5 - James A. Hanley Olli S. Miettinen, An unconditional- like structure for the conditional estimator of odds ratio 2*2 tables ,Biometrical Journal 48 (2006) 1; 23–34.
6 - Jacqueline S. (2003) , The application of case-cohort method to data on pulp and paper mill workers in British Columbia, university of Victoria .
8 – Alice S. , logistic regression of family data from study designs,genetic epidemiology 2005:177-189
9 - Giovanni M a. ,Simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients , bioinformatics Vol. 21 no. 23 2005: 4309–4311
10-kazem nezhad A,Gilani N,Zayri F,comparison of marginal model with repeated measures and conditional logistic model in risk factors affecting hypertension,J Mazand Univ Med Sci 2011,82:27-35
11-mazner s,keramer s,An introduction text epidemiology,W.B.Sunders company,1983,

Comparing odds ratio (OR) from fitting Independence, marginal and conditional models in analyzing the individual matched case-control studies with simulation data

*Esmaeili H[1], Salari M[2], saki A[*3], Gholizadeh B[4], Boskabadi M[4], Lashkardost[5]*

[1]Associated professor, Department Of Environmental Statistic and Epidemiology, School Of Health, Mashhad University of Medical Sciences, Mashad, Iran
[2]M.Sc of Biostatistics. School of Health, Mashhad University of Medical Sciences, Mashad, Iran
[3] Assistant professor, Department Of Environmental Statistic and Epidemiology, Jondi Shapur University of Medical Sciences, Ahvaz, Iran
[4]M.Sc of Statistics, Ferdosi University, Mashad, Iran
[5] M.Sc of Epidemiology, North Khorasan University of Medical Sciences, Bojnurd, Iran

***Corresponding Author:** Department of Epidemiology, North Khorasan University of Medical Sciences, Bojnurd, Iran
Email: saki-a@ajums.ac.ir*

Abstract

Background & Objectives *One of the popular studies in medical sciences for finding risk factors and the reason of the disease, are case-control studies that the important index we can calculate is odds ratio. but some confounders which effect on response variable challenge the OR's validity and present OR more or less than the real value. One way of omitting the effect of confounder is designing matched studies .Logistic regression is one of the general methods for modeling these studies. This article compares 3 logistic regression models: independence, marginal and conditional.*

Materials & Methods: *This study has been conducted on correlated simulated data. Thus the data is simulated from bivariate normal distribution with the correlation coefficients (0, 0.2, 0.4, 0.6, and 0.8). Then with changing cut-off points at (0.05, 0.25), (0.25, 0.1), (0.25, 0.15), (0.25, 0.25) for their c.d.f ,we convert continues distribution to categorical binary distribution which data are related together. Then 3 logistic regression model in independence, marginal and conditional version fit to data and calculate OR. With 10000 times iteration, we compare 2.5 and 97.5 percentiles values and the median OR percentile value at the above cut-off points for all their models.*

Results: *When the correlation is zero, all three models have the same quantity for OR and also changing in point have the same coefficient. But with the increasing correlation between the observations, OR between marginal model and independence model is not different. But its value will vary with the conditional model. For example, the cut-off point (0.25,0.1) and when the correlation is 0.6, median of OR that obtained in marginal model and independence model is 2.8, but in conditional model this quantity is 5 and it is twice of fitted value.*

Conclusion: *When the correlation between observations is high, using of conditional model is more correctly method and with increasing this correlation, our error rate by using independence or marginal model rises. But when correlation between observations is negligible, using of the three models gives similar estimates.*

Key words: *case-control studies, marginal logistic model, conditional logistic model, simulation, confounder.*