

مقایسه برآورده نسبت شانس حاصل از برازش مدل رگرسیون لجستیک با سه روش استقلال، حاشیه‌ای و اثرات تصادفی در مطالعات موردی-شاهدی همسان‌سازی شده با استفاده از شبیه‌سازی

حبيب الله اسماعيلي^۱، مریم سلاذری^۲، آزاده ساکی^{۳*}، بهزاد قلی‌زاده^۴، مصطفی بسکابادی^۵، حسین لشکردوست^۶

۱ دانشیار گروه آمار زیستی و اپیدمیولوژی و عضو مرکز تحقیقات علوم بهداشتی دانشگاه بهداشت، دانشگاه علوم پزشکی مشهد

۲ دانشجوی کارشناسی ارشد آمار زیستی دانشگاه بهداشت دانشگاه علوم پزشکی مشهد

۳ استادیار گروه آمار زیستی و اپیدمیولوژی دانشگاه 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، دانشگاه بهداشت، گروه آمار زیستی و اپیدمیولوژی

پست الکترونیک: saki-a@ajums.ac.ir

چکیده

زمینه و هدف: یکی از متداول‌ترین مطالعات در حیطه علوم پزشکی جهت یافتن ریسک فاكتورها و عوامل مرتبط با بیماری‌ها مطالعات موردی - شاهدی هستند که شاخص مهم قابل محاسبه در آن OR یا خط‌نسبی است. اما در این بین بعضی عوامل مخدوش‌گرکه بر پاسخ 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 متدالوی مدل‌سازی برای این نوع مطالعات است که در مطالعه حاضر سه روش رگرسیون لجستیک در حالت استقلال، حاشیه‌ای و شرطی با هم مقایسه می‌شوند.

مواد و روش کار: این مطالعه بر روی داده‌های شبیه‌سازی شده وابسته به هم انجام پذیرفته است. بدین ترتیب که داده‌ها از توزیع نرمال دو متغیره با ضریب همبستگی‌های ($0.25, 0.25, 0.25, 0.25, 0.25$) تولید می‌شوند. سپس با انتخاب نقاط برش ($0.25, 0.25, 0.25, 0.25, 0.25$)، ($0.1, 0.1, 0.1, 0.1, 0.1$) برای تابع احتمال تجمعی آنها این داده‌ها که از توزیع پیوسته هستند به توزیع گستته صفر و یک که به هم وابسته هستند تبدیل می‌شوند. سپس سه مدل رگرسیون لجستیک در حالت استقلال، حاشیه‌ای و شرطی به داده‌ها برازش داده می‌شود و OR محاسبه می‌شود. با 10000 بار تکرار مقدار صدک 2.5 و 97.5 و همچنین میانه OR سه مدل در نقاط برش ذکر شده با هم مقایسه می‌شوند.

یافته‌ها: در همبستگی صفر هر سه مدل OR مشابه دارند و تغییر در نقاط باز هم ضرایب مشابه دارد. اما با افزایش میزان همبستگی بین مشاهدات OR بین مدل حاشیه‌ای و استقلال متفاوت نیست ولی مقدار آن با مدل شرطی متفاوت خواهد بود. به عنوان مثال در نقطه برش ($0.1, 0.1, 0.1, 0.1, 0.1$) و ضریب همبستگی 0.6 میانه OR به دست آمده در مدل استقلال و حاشیه‌ای 2.1 است ولی در مدل شرطی این مقدار 5 یعنی دو برابر مقدار برازش شده است.

نتیجه گیری: استفاده از مدل‌های شرطی زمانی که همبستگی بین مشاهدات زیاد است قطعاً روش صحیح‌تری است و هرچه میزان این همبستگی بالا بود میزان خطای ما در استفاده از مدل استقلال یا حاشیه‌ای بالا می‌رود. اما زمانی که همبستگی بین مشاهدات ناچیز است استفاده از سه مدل برآوردهای یکسانی می‌دهد.

واژه‌های کلیدی: مطالعات موردی شاهدی، مدل لجستیک حاشیه‌ای، مدل لجستیک شرطی، شبیه‌سازی، مخدوش‌گر

مقدمه

اساساً برای آزمون فرضیه‌ها درباره سبب شناسی بیماری‌ها دو روش وجود دارد: تجربی و مشاهده‌ای. در روش تجربی پژوهشگران اثر تغییرات عاملی را که در اختیار دارند مطالعه می‌کنند. برای مثال ممکن است پژوهشگری یک گروه موش همزاد را به طور تصادفی انتخاب و نیمی از آنها را در معرض ماده‌ای که تصور می‌رود سرطان زاست قرار می‌دهند و سپس فراوانی سرطان را در هر دو گروه ثبت می‌کنند. روش معمول تر این است که پژوهشگر تنها می‌تواند رویداد بیماری را در گروه‌های مردم که بر اساس تجربه مواجهه (مانند متاحل‌ها در مقابل مجردان) یا سیگاری‌ها در مقابل غیرسیگاریها) از یکدیگر جداشده اند را مشاهده می‌کند. مشکل اساسی در مطالعات مشاهده‌ای این است که عموماً گروه‌های مشاهده شده علاوه بر عامل خاص تحت مطالعه از جانب پاره‌ای خصوصیات دیگر با هم متفاوتند. به دلیل این عوامل مخدوشگر که اغلب غیرقابل اندازه گیری هستند نمایش نقش عامل تحت بررسی مشکل‌تر می‌شود. یکی از انواع مطالعات علت شناسی در پژوهشکی مطالعات موردي شاهدی است [۱] .

مطالعات موردي شاهدی یکی از انواع مطالعات گذشته نگر هستند. در این مطالعات افرادی که تشخیص داده شده اند بیمار هستند (موارد) بالفرادی که بیمار نیستند (شاهدها) مقایسه می‌شوند. در طراحی این مطالعات روشنی که می‌توان اثر مخدوشگر را از بین برد بسیار مهم است. جور کردن یا همسان سازی موارد با شاهدها از نظر عوامل مخدوشگر معمول ترین روش است. جور کردن عبارت است از انتخاب شاهدها به نحوی که از لحاظ پاره‌ای مشخصات اختصاصی با موارد مشابه باشند. موارد و شاهدها را ممکن است تک تک با هم جور کرد. سن، جنس و نژاد معمول ترین متغیرهایی هستند که در جور کردن به کار گرفته می‌شود. یکی از شاخص‌های مهم برای تعیین میزان اهمیت عوامل خطر در بیماری‌ها محاسبه می‌شود خطرنسی (OR) است. [۲,۳]. خطرنسی (OR) یک اندازه خطر است که نشان دهنده آن است که چقدر شانس کسی که با یک عامل خطر مواجهه داشته از کسی که مواجهه نداشته بیشتر است. خطرنسی برای متغیر پاسخ گستته 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- داده‌ها: در این مقاله بجای استفاده از داده‌های واقعی از داده‌های شبیه سازی شده استفاده شده است. شبیه سازی فرآیندی است برای پاسخ دادن به سوالات 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 می‌شود.

۲- تولید داده‌ها به روش شبیه سازی: ابتدا برای تولید داده‌های دودویی همبسته به هم به روش زیر اجرا شد. نکته قابل توجه این است که متغیر مستقل یا همان Xها شبیه سازی می‌شوند به دلیل اینکه مطالعه موردي 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 حجم ۵۰ از توزیع نرمال دو متغیره با ضریب همبستگی ها در ۵ حالت (۰.۸، ۰.۶، ۰.۴، ۰.۲، ۰.۰) تولید می‌شود. یعنی ضریب همبستگی ۰.۲ افزایش پیدا می‌کند. از آنجایی که توزیع گسته‌صفر و یک ازتابع توزیع تجمعی استفاده می‌شود. یعنی نقاط برش مختلفی برای تبدیل این داده‌ها به صفر و یک استفاده می‌شود. دلیل دیگر این است که در مطالعات موردي شاهدی به عنوان مثال اهمیت یک ریسک فاکتور را باگذاشتن نقاط برش مختلف کاهش یا افزایش می‌دهیم حتی ممکن است متغیر مستقل به عنوان عامل حفاظتی در نظر گرفته شود. نقاط برش برای دو گروه مورد شاهد به ترتیب به صورت (۰.۰۲۵۰، ۰.۰۲۵۰)، (۰.۰۱۵۰، ۰.۰۱۵۰)، (۰.۰۰۵۰، ۰.۰۰۵۰) گرفته می‌شود یعنی اگر تابع توزیع تجمعی در گروه مورد از ۰.۰۲۵۰ کمتر باشد عدد صفر و اگر از ۰.۰۰۵۰ بیشتر باشد عدد یک در نظر گرفته می‌شود 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: در این مرحله سه مدل رگرسیون لجستیک در حالت استقلال، حاشیه‌ای (GEE) و شرطی برآش داده می‌شود. نکته بسیار مهم در این مرحله این است که در حالت استقلال هیچ شرطی در مورد ارتباط گروه مورد و شاهد نداریم ولی در دو حالت دیگر حتماً هر کنترل که با هرمورد همسان سازی شده به صورت یک خوشی یا یک طبقه در نظر گرفته شود برای این منظور با یک تغییر در نحوه چیدمان داده‌های شبیه سازی شده داده‌ها به صورت زیر خواهند بود:

جدول ۱: نحوه خوش بندی مشاهدات برای برآورد مدل شرطی

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

مثال در مدل لجیت لگاریتم تابع درستنمایی به صورت رابطه ۱ خواهد 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 مدل‌های هستند که تاثیر متغیر مستقل روی پاسخ را به طور مجزا از همبستگی بین پاسخ‌ها برای یک آزمودنی معین مدل بندی می‌کند. در این مدل‌ها امید حاشیه ای $E(Y_{ij})$ به صورت تابعی از متغیرهای توضیحی مدل بندی می‌شود.

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 $h(\mu_{ij}) = x_{ij}\beta$ وابسته است که در آن h تابع پیوند شناخته شده ای است. به عنوان

مثال برای پاسخ‌های دودویی تابع لگاریتم است.

۲- واریانس حاشیه ای به میانگین حاشیه ای به صورت Φ $\varphi(y_{ij}) = v(\mu_{ij})$ 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.

زمانی که متغیر پاسخ در یک مطالعه دارای توزیعی غیر از نرمال باشد همانند توزیع برنولی، دو جمله ای و یا پواسن از مدل‌های خطی تعمیم یافته که تعمیمی از مدل‌های خطی همانند رگرسیون می‌باشد استفاده می‌شود.

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

در واقع Y یک متغیر تصادفی برنولی با $P(y_{ij}=1) = \pi_i$ باشد. که مدل رگرسیون لجستیک یک رابطه خطی بین X و لگاریتم شانس رخداد پیشامد بیماری $(\frac{\pi_i}{1-\pi_i})$ برای فرد آم برقرار می‌کند و به صورت زیر نوشته می‌شود.

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

هرچند مدل‌های لجستیک برای مطالعات موردی شاهدی قابل استفاده هستند ولی یک محدودیت بزرگ دارند و آن هم پیش بینی خطر برای یک فرد را نمی‌توانند انجام دهند. در واقع فقط خطرنسبی برای این مطالعات قابل محاسبه است. که در آن e^{β} همان خطرنسبی OR است. برای برآورد ضرایب رگرسیونی از روش معادلات درستنمایی استفاده می‌شود که با فرض مستقل بودن مشاهدات معادلات درستنایی نوشته می‌شوند به عنوان

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

برش (0.25,0.25) یعنی زمانی که مواجهه درگروه مورد و شاهد یکسان است عدد یک را شامل می شود که همانطور که انتظار می رو د 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. همینطور که احتمال مواجهه درگروه مورد بالا می رو د 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. در نقطه برش (0.25,0.25) هرچند هر سه مدل میانه یک را نشان می دهد. ولی صدک ۲.۵ و ۹۷.۵ در مدل استقلال و حاشیه ای شبیه هم بین (۰.۴۰،۰.۴) ولی در مدل شرطی بین OR (۳۰،۳۳) است. همینطور که نقاط برش را تغییر می دهیم 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 عده ۶۰. تفاوت در برآش سه مدل آشکارتر می شود. به عنوان مثال در نقطه برش (۰.۲۵، ۰.۰۱) 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 در مدل حاشیه ای و استقلال شبیه هم اما با مدل شرطی بسیار متفاوت خواهد بود.

بحث

هدف این مطالعه آشکار کردن میزان خطای بود که زمانی که مشاهدات به هم وابسته هستند در انتخاب مدل استقلال یا حاشیه ای خواهیم داشت بود. همانطور که در بالا ذکر شد استفاده از مدل های شرطی زمانی که همبستگی بین مشاهدات زیاد است قطعاً 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 میانه ها و فاصله اطمینان را نشان میدهد. هرچند مدل حاشیه ای درگروه مدل هایی قرار می گیرد که همبستگی بین مشاهدات را لاحاظ می کند ولی چون روی میانیگ حاشیه ها مدل بندی می کند، ضرایب را بهبود نمی بخشد.

۳- همبستگی y_{ij} و y_{ik} به صورت تابعی از میانگین حاشیه ای آنها و پارامترهای اضافی α است یعنی $\text{corr}(y_{ij}, y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \alpha)$ که در آن ρ تابع شناخته شده ای از μ_{ij} هاست. می توان گفت مدل های حاشیه ای برای پاسخ های وابسته تعمیمی از GLM ها برای پاسخ های مستقل هستند. یکی از مدل های معروف حاشیه ای معادلات برآورده تعمیم یافته (GEE) است که بوسیله Liang و zeger در سال ۱۹۸۶ معرفی شد.

رهیافت معادلات برآورده 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 میانیگ حاشیه ای است. به علاوه ماتریس همبستگی عملی برای مشاهدات هر آزمودنی معین می شود.

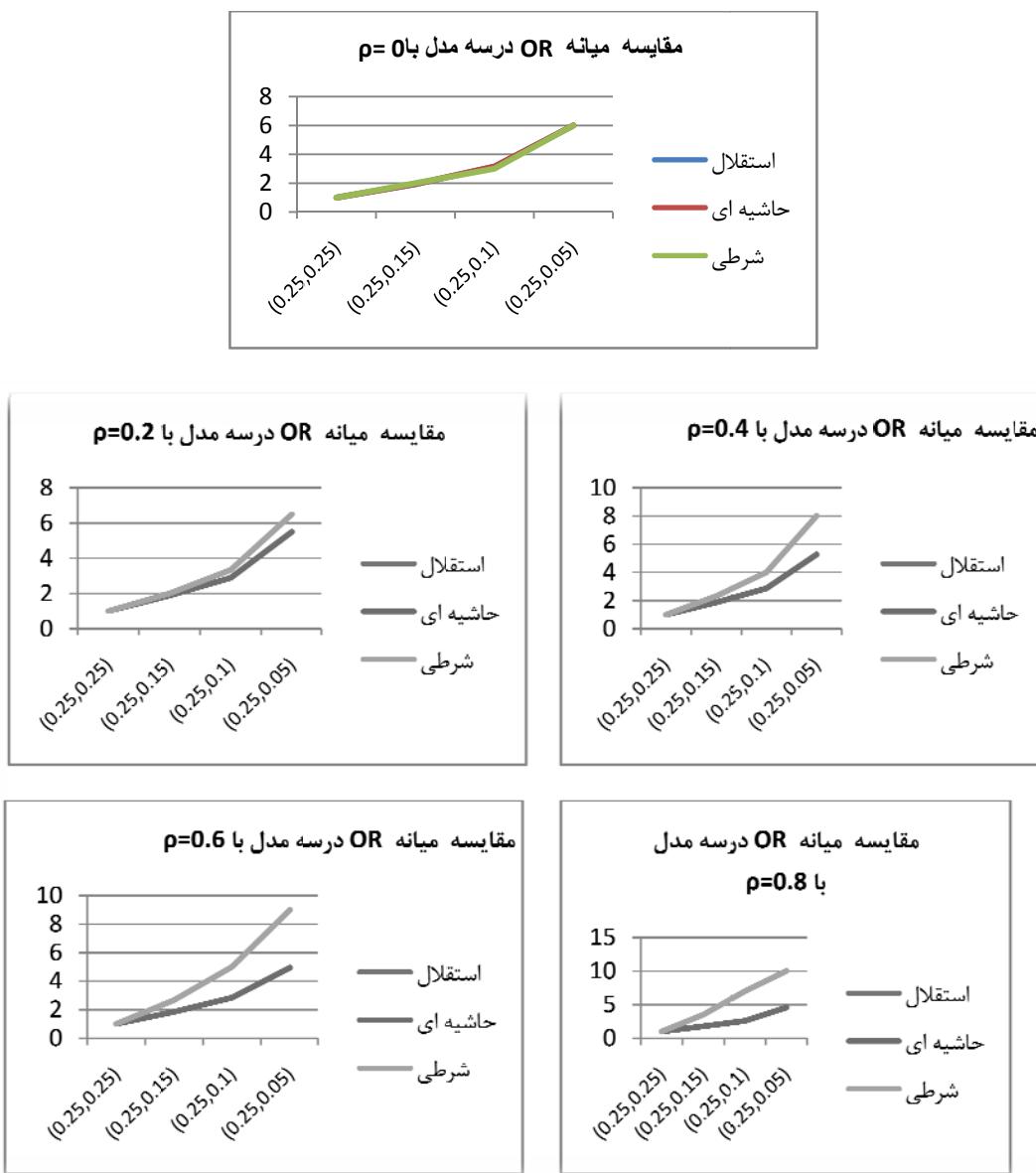
ب- مدل اثرات تصادفی (شرطی) در این مدل فرض می شود ، مدل منحصر به یک آزمودنی خاص است. در این صورت ضرایب نشان دهنده تغییرات پاسخ در کلاس مورد نظر به ازاء یک واحد تغییر در متغیر پیشگوست. یا چنین تصور کنیم که ضرایب نشان دهنده میزان تغییر در داخل آزمودنی است وقتی متغیر پیشگو یک واحد تغییر می کند (وقتی بقیه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

یافته ها

با استفاده از نرم افزار R و زن ۱۵ فرایند تولید داده تصادفی ۱۰۰۰۰ بار با تغییر ضریب همبستگی و نقاط برشی که در بالا ذکر شد تکرار شد و در هر تکرار، پس از برآش مدل ها مقادیر OR از رابطه $\text{EXP}(\beta)$ 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 می شود محاسبه گردید. زمانی که همبستگی بین مشاهدات صفر است همانطور که در جدول ۴ ملاحظه می شود هر سه مدل شرطی، حاشیه ای و استقلال برای OR مقادیر مشابهی را نشان می دهند. یعنی اگر مشاهدات مستقل باشند تفاوتی در ضرایب مدل رگرسیونی و به همین ترتیب در خطرنسی مشاهده در سه مدل نخواهیم کرد. بازه اطمینان OR در نقطه

جدول ۲: مقایسه OR حاصل از سه مدل استقلال، حاشیه‌ای و شرطی در نقاط برش و ضریب همبستگی‌های متفاوت

نقاط برش	مدل حاشیه‌ای				مدل استقلال				مدل شرطی	
	OR _{+/+25}	OR _{+/-5}	OR _{-/+25}	OR _{-/-25}	OR _{/+5}	OR _{-/+25}	OR _{-/-25}	OR _{/+5}	OR _{-/+25}	OR _{-/-25}
(+/25,+/25)	0/39	1/00	2/58	0/39	1/00	2/58	0/36	1/00	2/75	
(+/25,+/15)	0/70	1/94	6/00	0/70	1/94	6/00	0/70	2/00	7/00	
(+/25,+/1)	1/00	3/14	11/29	1/00	3/14	11/29	1/00	3/00	13/00	
(+/25,+/-5)	1/83	6/00	23/06	1/83	6/00	23/06	1/75	6/00	17/00	
$\rho = 0.2$										
نقاط برش	مدل حاشیه‌ای				مدل استقلال				مدل شرطی	
	OR _{+/+25}	OR _{+/-5}	OR _{-/+25}	OR _{-/-25}	OR _{/+5}	OR _{-/+25}	OR _{-/-25}	OR _{/+5}	OR _{-/+25}	OR _{-/-25}
(+/25,+/25)	0/40	1/00	2/45	0/40	1/00	2/45	0/33	1/00	3/00	
(+/25,+/15)	0/72	1/88	5/50	0/72	1/88	5/50	0/70	2/00	9/00	
(+/25,+/1)	1/00	2/89	10/76	1/00	2/89	10/76	1/00	3/33	14/00	
(+/25,+/-5)	1/87	5/50	21/00	1/87	5/50	21/00	2/00	6/50	16/00	
$\rho = 0.4$										
نقاط برش	مدل حاشیه‌ای				مدل استقلال				مدل شرطی	
	OR _{+/+25}	OR _{+/-5}	OR _{-/+25}	OR _{-/-25}	OR _{/+5}	OR _{-/+25}	OR _{-/-25}	OR _{/+5}	OR _{-/+25}	OR _{-/-25}
(+/25,+/25)	0/44	1/00	2/25	0/44	1/00	2/25	0/30	1/00	3/00	
(+/25,+/15)	0/79	1/88	5/06	0/79	1/88	5/06	0/75	2/33	10/00	
(+/25,+/1)	1/17	2/88	9/33	1/17	2/88	9/33	1/20	4/00	14/00	
(+/25,+/-5)	1/94	5/27	19/06	1/94	5/27	19/06	2/25	8/00	16/00	
$\rho = 0.6$										
نقاط برش	مدل حاشیه‌ای				مدل استقلال				مدل شرطی	
	OR _{+/+25}	OR _{+/-5}	OR _{-/+25}	OR _{-/-25}	OR _{/+5}	OR _{-/+25}	OR _{-/-25}	OR _{/+5}	OR _{-/+25}	OR _{-/-25}
(+/25,+/25)	0/47	1/00	2/06	0/47	1/00	2/06	0/25	1/00	3/67	
(+/25,+/15)	0/85	1/84	4/57	0/85	1/84	4/57	0/75	2/67	11/00	
(+/25,+/1)	1/23	2/84	7/98	1/23	2/84	7/98	1/40	5/00	14/00	
(+/25,+/-5)	1/98	4/95	17/22	1/98	4/95	17/22	2/67	9/00	16/00	
$\rho = 0.8$										
نقاط برش	مدل حاشیه‌ای				مدل استقلال				مدل شرطی	
	OR _{+/+25}	OR _{+/-5}	OR _{-/+25}	OR _{-/-25}	OR _{/+5}	OR _{-/+25}	OR _{-/-25}	OR _{/+5}	OR _{-/+25}	OR _{-/-25}
(+/25,+/25)	0/54	1/00	1/83	0/54	1/00	1/83	0/17	1/00	5/00	
(+/25,+/15)	1/00	1/77	3/86	1/00	1/77	3/86	1/00	3/50	11/00	
(+/25,+/1)	1/31	2/58	6/77	1/31	2/58	6/77	2/00	7/00	13/00	
(+/25,+/-5)	1/98	4/57	15/47	1/98	4/57	15/47	3/50	10/00	16/00	



نمودار ۱: مقایسه OR حاصل از سه مدل استقلال، حاشیه ای و شرطی در نقاط برش و ضریب

تشکر و قدردانی

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

نتیجه گیری

استفاده از مدل های شرطی زمانی که همبستگی بین مشاهدات زیاد است قطعاً روش صحیح تری است و هرچه میزان این همبستگی بالا بروز میزان خطای ما در استفاده از مدل استقلال یا حاشیه ای بالا می رود. اما زمانی که همبستگی بین مشاهدات ناقص است استفاده از سه مدل برآوردهای یکسانی می دهد.

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 H¹, Salari M², saki A^{*3}, Gholizadeh B⁴, Boskabadi M⁴, Lashkardost⁵

¹Associated professor, Department Of Environmental Statistic and Epidemiology, School Of Health, Mashhad University of Medical Sciences, Mashad, Iran

²M.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

⁴M.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.
