

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

حبیب اله اسماعیلی<sup>۱</sup>، مریم سالاری<sup>۲</sup>، آزاده ساکی<sup>۳\*</sup>، بهزاد قلی زاده<sup>۴</sup>، مصطفی بسکابادی<sup>۴</sup>، حسین لشکردوست<sup>۵</sup>

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 آمار زیستی و اپیدمیولوژی

پست الکترونیک: saki-a@ajums.ac.ir

### چکیده

**زمینه و هدف:** یکی از متداول ترین مطالعات در حیطه علوم پزشکی جهت یافتن ریسک فاکتورها و عوامل مرتبط با بیماری ها مطالعات موردی - شاهدی هستند که شاخص مهم قابل محاسبه در آن  $OR$  یا خطر نسبی است. اما در این بین بعضی عوامل مخدوش گر که بر پاسخ موثرند اعتبار  $OR$  به دست آمده را زیر سوال می برند و  $OR$  را کمتر یا بیشتر نشان می دهد یکی از روش های حذف اثر مخدوش گر، طراحی مطالعات  $matching$  است. رگرسیون لجستیک یکی از روش های متداول مدل سازی 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 همبستگی های  $(0.02, 0.04, 0.06, 0.08)$  تولید می شوند. سپس با انتخاب نقاط برش  $(0.25, 0.25)$ ،  $(0.25, 0.15)$ ،  $(0.25, 0.1)$ ، برای تابع احتمال تجمعی آنها این داده ها که از توزیع پیوسته هستند به توزیع گسسته صفر و یک که به هم وابسته هستند تبدیل می شوند. سپس سه مدل رگرسیون لجستیک در حالت استقلال، حاشیه ای و شرطی به داده ها برازش داده می شود و  $OR$  محاسبه می شود. با ۱۰۰۰ بار تکرار مقدار صدک ۲.۵ و ۹۷.۵ و همچنین میانه  $OR$  سه مدل در نقاط برش ذکر شده با هم مقایسه می شوند.

**یافته ها:** 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 حاشیه ای و استقلال متفاوت نیست ولی مقدار آن با مدل شرطی متفاوت خواهد بود. به عنوان مثال در نقطه برش  $(0.25, 0.1)$  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 دو برابر مقدار برازش شده است.

**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 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، مخدوش گر

## 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 تصوری رود سرطان‌زا قرار می‌دهند و سپس فراوانی سرطان را در هر دو گروه ثبت می‌کنند. روش معمول‌تر این است که پژوهشگر تنها می‌تواند رویداد بیماری را در گروه‌های مردم که بر اساس تجربه مواجهه (مانند متاهل‌ها در مقابل مجردها یا سیگاری‌ها در مقابل غیرسیگاریها) از یکدیگر جدا شده‌اند را مشاهده می‌کند. مشکل اساسی در مطالعات مشاهده‌ای این است که معمولاً گروه‌های مشاهده‌شده علاوه بر عامل خاص تحت مطالعه از جانب پاره‌ای خصوصیات دیگر با هم متفاوتند. به دلیل این عوامل مخدوشگر که اغلب غیرقابل اندازه‌گیری هستند نمایش نقش عامل تحت بررسی مشکل‌تر می‌شود. یکی از انواع مطالعات علت شناسی در پزشکی مطالعات موردی شاهده‌ای است [۱].

مطالعات موردی شاهده‌ای یکی از انواع مطالعات گذشته‌نگر هستند. در این مطالعات افرادی که تشخیص داده شده‌اند بیمار هستند (موارد) با افرادی که بیمار نیستند (شاهدها) مقایسه می‌شوند. در طراحی این مطالعات روشی که می‌توان اثر مخدوشگر را از بین برد بسیار مهم است. جورکردن یا همسان‌سازی موارد با شاهدها از نظر عوامل مخدوشگر معمول‌ترین روش است. جورکردن عبارت است از انتخاب شاهدها به نحوی که از لحاظ پاره‌ای مشخصات اختصاصی با موارد مشابه باشند. موارد و شاهدها را ممکن است تک‌تک با هم جورکرد. سن، جنس و نژاد معمول‌ترین متغیرهایی هستند که در جورکردن به کارگرفته می‌شود. یکی از شاخص‌های مهم برای تعیین میزان اهمیت عوامل خطر در بیماری‌ها محاسبه می‌شود خطرنسبی (OR) است. [۲،۳]. خطرنسبی (OR) یک اندازه خطر است که نشان‌دهنده آن است که چقدر شانس کسی که با یک عامل خطر مواجهه داشته از کسی که مواجهه نداشته بیشتر است. خطرنسبی برای متغیر پاسخ گسسته هم از طریق جداول توافقی و هم از روش مدل‌سازی محاسبه کرد. در این مقاله روش مدل‌سازی موردنظر است [۴،۵].

## روش کار

۱- داده‌ها: در این مقاله بجای استفاده از داده‌های واقعی از داده‌های شبیه‌سازی 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 شبیه‌سازی استفاده می‌شود.

۲- تولید داده‌ها به روش شبیه‌سازی: ابتدا برای تولید داده‌های دودویی همبسته به هم به روش زیر اجرا شد. نکته قابل توجه این است که متغیر مستقل یا همان  $X$ ها شبیه‌سازی می‌شوند به دلیل اینکه مطالعه موردی شاهده‌ای است و متغیر  $Y$  از قبل مشخص است که برای گروه مورد برابر یک و برای گروه شاهد برابر صفر است. نمونه‌هایی به حجم ۵۰ از توزیع نرمال دو متغیره با ضریب همبستگی‌ها در ۵ حالت (۰.۰۲، ۰.۴، ۰.۶، ۰.۸، ۰.۲۰) تولید می‌شود. یعنی ضریب همبستگی ۰.۲ افزایش پیدا می‌کند. از آنجایی که توزیع نرمال یک توزیع پیوسته است برای تبدیل آن به توزیع گسسته صفر و یک از تابع 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 موردی شاهده‌ای به عنوان مثال اهمیت یک ریسک فاکتور را با گذاشتن 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 (۰.۲۵، ۰.۲۵)، (۰.۲۵، ۰.۱۵)، (۰.۲۵، ۰.۱)، (۰.۲۵، ۰.۰۵)، (۰.۲۵، ۰.۰۵) گرفته می‌شود یعنی اگر تابع توزیع تجمعی در گروه مورد از ۰.۲۵ کمتر باشد عدد صفر و اگر از ۰.۲۵ 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 شبیه‌سازی ثابت خواهد ماند.

مدل‌سازی آماری: در این مرحله سه مدل رگرسیون لجستیک در حالت استقلال، حاشیه‌ای (GEE) و شرطی برازش 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 هر مورد همسان سازی شده به صورت یک خوشه یا یک طبقه در نظر گرفته شود. برای این منظور با یک تغییر در نحوه چیدمان داده‌های شبیه‌سازی شده داده‌ها به صورت زیر خواهند بود:

جدول ۱: نحوه خوشه بندی مشاهدات برای برازش مدل شرطی

ردیف	شماره خوشه	مورد=۱ شاهد=۰	مواجهه=۱ عدم 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

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 شاهدهی همسان سازی شده یعنی زمانی که مشاهدات به هم وابسته هستند استفاده میشود.

۲- مدل رگرسیون لجستیک حاشیه ای: مدلهای حاشیه ای مدل هایی هستند که تاثیر متغیر مستقل 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  $E(Y_{ij})$  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  $\mu_{ij} = \beta_0 + \beta_1 X_{ij}$  وابسته است که در آن  $h$  تابع 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  $\Phi$   $var(y_{ij}) = v(\mu_{ij})$  وابسته است که  $v$  تابع واریانس شناخته شده ای است و  $\Phi$  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، دوجمله ای و یا پواسن از مدلهای خطی تعمیم یافته که تعمیمی از مدلهای خطی همانند رگرسیون می باشند استفاده می شود.

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

درواقع  $Y$  یک متغیر تصادفی 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  $i$ ام 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 شاهدهی قابل استفاده هستند ولی یک محدودیت بزرگ دارند و آن هم پیش بینی خطر برای یک فرد را نمی توانند انجام دهند. درواقع فقط خطر نسبی برای این مطالعات قابل محاسبه است. که در آن  $e^{\beta}$  همان خطر نسبی OR است. برای برآورد ضرایب رگرسیونی از روش معادلات درستنمایی استفاده می شود که با فرض مستقل بودن مشاهدات معادلات درستنمایی نوشته می شوند به عنوان

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

۳- همبستگی  $y_{ij}$  و  $y_{ik}$  به صورت تابعی از میانگین حاشیه ای آنها و پارامترهای اضافی  $\alpha$  است یعنی

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

که در آن  $\rho$  تابع شناخته شده ای از  $\mu_i$  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 برآورد تعمیم یافته (GEE) است که بوسیله Liang و Zeger در سال 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  $u_i$  مقادیر  $y_{ij}$  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 برازش می دهد در حالی که مدل شرطی قابلیت برازش مدل شرطی را هم دارد.

یافته ها

با استفاده از نرم افزار R ورژن 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 برازش 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 مقادیرت مشابهی را نشان می دهند. یعنی اگر مشاهدات مستقل باشند تفاوتی در ضرایب مدل رگرسیونی و به همین ترتیب در خطر نسبی مشاهده در سه مدل نخواهیم کرد. بازه اطمینان OR در نقطه

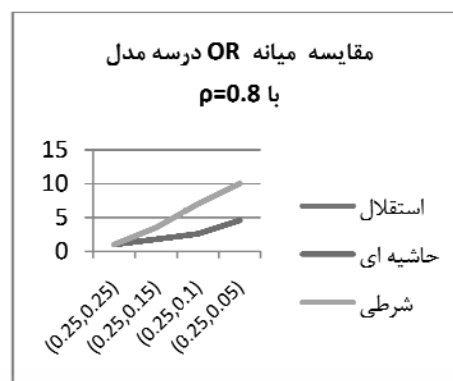
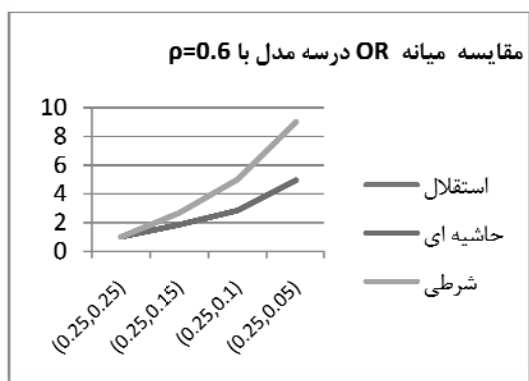
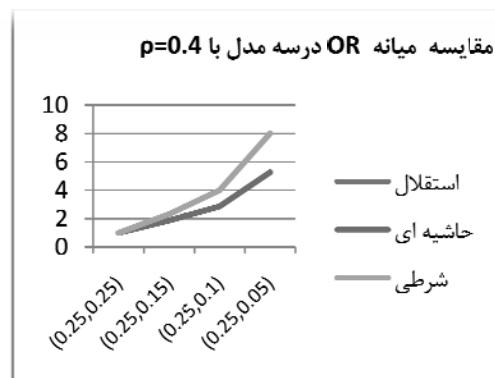
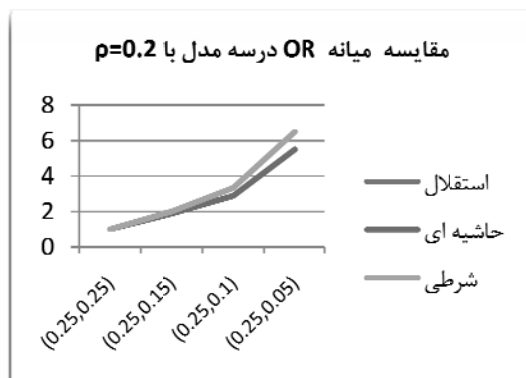
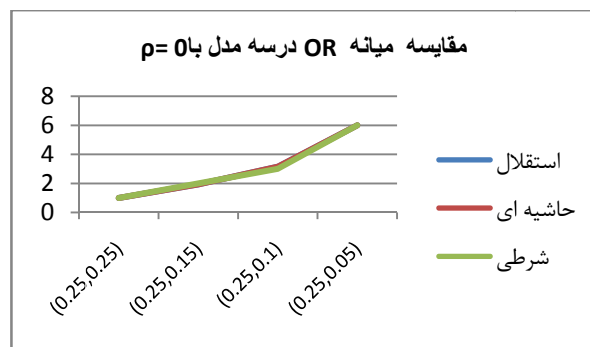
برش (0.25,0.25) یعنی زمانی که مواجهه در گروه مورد و شاهد یکسان است عدد یک را شامل می شود که همانطور که انتظار می رود نشان دهنده این است که متغیر مورد بررسی یک عامل خطر نیست. اما در نقطه برش (0.25,0.15) هر چند بازه شامل یک می شود ولی میانه فاصله عدد ۱.۹ در مدل استقلال و حاشیه ای و ۲ در مدل شرطی است . همینطور که احتمال مواجهه در گروه 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 . در نقطه برش (0.25,0.25) هر چند 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 (۳.۳، ۰.۳) است. همینطور که نقاط برش را تغییر می 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 عدد ۰.۶ 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 در مدل حاشیه ای و استقلال شبیه هم اما با مدل شرطی بسیار متفاوت خواهد بود.

#### بحث

هدف این مطالعه آشکار کردن میزان خطایی 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 هر چه 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. هر چند مدل حاشیه ای در گروه مدل هایی قرار می گیرد که همبستگی بین مشاهدات را لحاظ می کند ولی چون روی میانگین حاشیه ها مدل بندی می کند ، 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
$\rho=0.2$									
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
$\rho=0.4$									
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
$\rho=0.6$									
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
$\rho=0.8$									
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



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

### 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 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<sup>1</sup>, Salari M<sup>2</sup>, saki A<sup>\*3</sup>, Gholizadeh B<sup>4</sup>, Boskabadi M<sup>4</sup>, Lashkardost<sup>5</sup>

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

<sup>2</sup>M.Sc of Biostatistics. School of Health, Mashhad University of Medical Sciences, Mashad, Iran

<sup>3</sup> Assistant professor, Department Of Environmental Statistic and Epidemiology, Jondi Shapur University of Medical Sciences, Ahvaz, Iran

<sup>4</sup>M.Sc of Statistics, Ferdosi University, Mashad, Iran

<sup>5</sup> 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.

---