مجله دانشگاه علوم پزشکی خراسان شمالی
۱۳۹۰؛ ۳ (ویژه نامه آمار زیستی و اپیدمیولوژی):۱۵۸-۱۵۱

Journal of North Khorasan University of Medical Sciences
2011;3(Biostatistics and Epidemiology Supplement):151-158

مقایسه مدلهای آماری حاشیه ای و آمیخته در تحلیل داده های پزشکی

میر سعید یکانی نژاد[۱]، مهدی یاسری[*۱]، کرامت نوری جلیانی[۲]، آرش اکابری[۳]*

[۱]دانشجوی دوره دکتری تخصصی آمار زیستی، گروه اپیدمیولوژی و آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی تهران ،تهران، ایران
[۲]دانشیار آمار زیستی، گروه اپیدمیولوژی و آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی تهران ،تهران، ایران
[۳] کارشناس ارشد آمارزیستی، دانشگاه علوم پزشکی خراسان شمالی، بجنورد، ایران
[*]**نویسنده مسئول:** تهران، دانشگاه علوم پزشکی تهران، دانشکده بهداشت، گروه اپیدمیولوژی و آمار زیستی
پست الکترونیک: m.yaseri@yahoo.com

چکیده

زمینه و هدف: *امروزه داده های زیادی وجود دارند که در آنها فرض استقلال داده ها که پیش فرض اصلی بسیاری از مدلهای آماری است برقرار نیست. داده های حاصل از نمونه گیری خوشه ای، مطالعات طولی با اندازه گیری های مکرر و یا داده های زوجی مانند داده های دو چشم و همچنین مطالعاتِ همسان سازی شده نمونه هایی از این داده ها هستند*

مواد و روش کار: *دو مدل آماری با در نظر گرفتن همبستگی بین مشاهدات، مدلهای آمیخته و مدلهای حاشیه ای هستند که در این مقاله مورد مقایسه قرار گرفته اند. این مدلها از نظر روشی که همبستگی بین داده ها را در نظر می گیرند و تفسیر ضرایب رگرسیونی با یکدیگر تفاوت دارند.*

یافته ها: *در مدلهای غیر خطی ضرایب رگرسیونی مدل حاشیه ای تغییرات در سطح جامعه را نشان می دهد در حالیکه در مدلهای آمیخته این ضرایب تغییرات را در یک فرد یا خوشه خاص نشان می دهند. در مدلهای خطی تفسیر ضرایب رگرسیونی در هر دو مدل یکسان است.*

نتیجه گیری: *به عنوان یک نتیجه گیری کلی در مدلهای غیر خطی که تفسیر نتایج متفاوت است بر اساس کاربرد مورد نیاز بهتر است از یکی از این دو مدل استفاده شود. در واقع برای سیاستگذاران بهداشتی که دید جامعه نگر دارند استفاده از مدل حاشیه ای توصیه می شود در حالیکه برای پزشکان معالج که بیشتر به تغییرات فردی در یک بیمار علا قمند هستند استفاده از مدل آمیخته مناسبتر است.*

واژه های کلیدی: *مدل حاشیه ای، مدل آمیخته، مدل خطی، مدل غیر خطی*

مقدمه

در بسیاری از داده های پزشکی فرض استقلال داده ها برقرار نمی‌باشد برای مثال داده هایی که به روش نمونه گیری خوشه ای جمع آوری می شوند و یا مطالعاتی که بصورت طولی انجام می شوند و در آن اندازه گیری های مکرر وجود دارد، و یا داده هایی که به طور همزمان از یک یا چند عضو بدن جمع آوری می شود و همچنین هنگامیکه در یک مطالعه همسان سازی انجام می شود فرض استقلال داده ها را نمی‌توان پذیرفت زیرا داده های یک خوشه، یک فرد در طول زمان یا نمونه های همسان شده به هم وابسته هستند. پیش فرض مورد نیاز در برازش مدلهای خطی تعمیم یافته فرض استقلال داده هاست برقرار نمی باشد لذا استفاده از این مدلها گرچه ممکن است به اریبی برآوردگرها نیانجامد و یا مقدار این اریبی قابل اغماض باشد با این وجود خطای معیار برآورد ضرایب به شدت تحت تاثیر پیش فرض استقلال داده ها است[۱-۳]. به عبارت دیگر در صورتیکه داده ها از هم مستقل نباشند خطای معیار و در نتیجه فاصله اطمینان و نتیجه آزمون فرضها برای صفر بودن ضرایب رگرسیونی غیر قابل اعتماد خواهندبود. لذا در آنالیز این داده ها استفاده از روشهایی که بتواند این وابستگی را در نظر بگیرد ضروری است. روشهای مختلفی برای حل این مساله ابداع شده

است که مهمترین آنها مدلهای آمیخته[1] و مدلهای حاشیه-ای[2] هستند. گرچه از ابداع این مدلها مدت زمان زیادی می گذرد اما هنوز اتفاق نظر در مورد برتری یکی از این مدلها بر دیگری وجود ندارد. [۱, ۴-۶]برخی از آمار شناسان برای هر یک از این دو مدل هویت مستقل قائل هستند و برتری هریک از این مدل‌ها را بر اساس کاربرد آنها تعریف می کنند .[۴،۵،۷] در این مقاله به بررسی تفاوتهای مدلهای آمیخته و مدلهای حاشیه ای می پردازیم. این مقاله شامل دو بخش اصلی است که در این دوبخش تفاوتهای میان مدلهای آمیخته و مدلهای حاشیه ای در دو کلاس بزرگ مدلهای خطی[3] و مدلهای خطی تعمیم[4] یافته مورد بررسی قرار می گیرد. در هر یک از این دو بخش ابتدا تفاوت در نحوه در نظر گرفتن همبستگی در این دو مدل بررسی می شود و سپس تفسیر ضرایب رگرسیونی در آنها توضیح داده می شود و در پایان با یک مثال عددی و یک مثال واقعی حاصل از طرح سلامت و بیماری در سال ۱۳۹۰ اختلاف در تفسیر ضرایب در این دو مدل نشان داده شده و تفاوت درکاربرد آنها بیان می شود.

بخش اول: مدلهای خطی

در مدلهای کلاسیک آماری میانگین متغیر پاسخ به صورت تابعی خطی از متغیر های مستقل بیان می شود و به منظور در نظر گرفتن همبستگی مشاهدات از دو روش عمده استفاده می شود در مدلهای حاشیه ای یک الگوی همبستگی تعریف می شود که به صورت یک ماتریس همبستگی است

(۱)

$$E(y_i) = X_i\beta$$

(۲)

$$\sum i = cor\ (y_i)$$

روش دوم معرفی اثرات تصادفی در مدل برای میانگین متغیر پاسخ است.

(۳)

$$E(y_i \mid b_i) = X_i\beta + z_i b_i$$

که b یک بردار از اثرات تصادفی است که از خوشه ای به خوشه ی دیگر در تغییر است. معرفی اثرات تصادفی منجر به ایجاد ساختار کوواریانس اثرات تصادفی می شود .

(۴)

$$\sum i = Z_i G Z_i' + \delta^2 I_{ni}$$

در دو مدل مذکور تفسیر ضرایب β یکسان است و در هر دوی آنها ضریب β نشان دهنده‌ی آن است که متغیرهای کمکی به چه میزان ، میانگین متغیر پاسخ را در جامعه مورد مطالعه تغییر می دهند.[۲,۳] این تفسیر از β در معادله (۱) کاملا آشکار است و مدلهای حاشیه‌ای مستقیما به مدل (۱) اشاره دارند از این رو می‌توان تفسیر β را به تغییرات در میانگین جامعه تعمیم داد. جالب است که در مورد مدلهای آمیخته ی خطی نیز می توان تفسیر حاشیه ای از میانگین متغیر پاسخ داشت زیرا:

$$E(y_i) = E\{E(y_i \mid b_i)\}$$

(۵)

$$= E(X_i\beta + z_i b_i)$$
$$= X_i\beta + z_i E(b_i)$$
$$= X_i\beta$$

پارامتر β در معادلات (۱) و (۳) تفسیر یکسانی دارد. زیرا معادله‌ی (۳) یک مدل آمیخته خطی است. (به عبارتی سمت راست (۳) یک تابع خطی از b است.)

به بیان ساده‌تر در مدل اثرات آمیخته خطی، ضرایب β تفسیر حاشیه‌ای دارد زیرا میانگین تغییرات خطی متغیر پاسخ بر اساس متغیرهای کمکی در هر سطح برابر تغییرات میانگین متغیر پاسخ در جامعه بر اساس متغیر های کمکی است.

بخش دوم: مدلهای تعمیم یافته‌ی خطی

در این بخش به مقایسه‌ی مدلهای حاشیه‌ای و مدلهای تعمیم یافته آمیخته می‌پردازیم همانطور که می‌دانیم یکی از مولفه‌های مدلهای تعمیم یافته، تابع ربط[5] است.

1 - Mixed model
2 - Marginal model
3 - Linear model
4 - Generalized linear models

5 - Link function

$g(\mu_i)$ که میانگین متغیر پاسخ را به تابعی خطی از متغیر های کمکی مرتبط می سازد. در بحثی که در مقایسه ی دو مدل، در بخش اول انجام گرفت به عنوان یک حالت خاص از مدل‌های تعمیم یافته، تابع ربط یک تابع همانی بود. $g(\mu_i) = \mu_i$ که در این حالت تفسیر پارامترهای رگرسیونی β در دو مدل آمیخته و مدل حاشیه ای یکسان است. در اینجا حالتی را مورد بررسی قرار خواهیم داد که تابع ربط غیر خطی است. مدل حاشیه‌ای تعمیم یافته ی خطی در این حالت بصورت زیر بیان می‌شود.

(۶)

$$g(\mu_i) = g\{E(y_i)\} = X_i\beta$$

که در آن g(.) یک تابع ربط غیر خطی متناسب با متغیر پاسخ است. در مدلهای حاشیه‌ای تفسیر پارامترهای رگرسیونی β به صورت تغییر در میانگین متغیر پاسخ در جامعه مورد مطالعه بر اساس تغییرات متغیرهای کمکی است. برای مثال، زمانی که پاسخ بصورت بردار y_i دو وضعیتی است و تابع ربط لجیت برای آنها استفاده می‌شود تفسیر پارامترهای رگرسیونی β به صورت میزان تغییر در لگاریتم نسبت شانس در جامعه است

(۷)

$$\log it(\mu_i) = X_i\beta$$

برای هر تابع ربط مفروض، g(.)، میانگین جامعه را می توان بر اساس معکوس تابع ربط بیان کرد.

(۸)

$$h\{g(\mu_i)\} = \mu_i = E(y_i) = h(X_i\beta)$$

به عنوان مثال، زمانی که y_iها دو وضعیتی هستند و تابع ربط لجیت است مدلی که برای μ_i در نظر گرفته می شود به صورت زیر است :

که در اینجا h(.) تابع معکوس لجیت است. پارامترهای رگرسیونی β، در هر دو فرم (۶) و(۸) نشان دهنده ی تغییر در میانگین متغیر پاسخ در جامعه به ازای تغییرات در متغیر های کمکی هستند .

اکنون مدلهای تعمیم یافته ی آمیخته ی خطی را برای میانگین شرطی متغیر پاسخ، به شرط برداری از اثرات تصادفی b_i در نظر می گیریم.

(۹)

$$g[E(y_i \mid b_i)] = X_i\beta^* + Z_i b_i$$

در اینجا پارامترهای ثابت را با β^* نمایش می دهیم تا از نظیر آنها در مدلهای حاشیه ای متمایز باشند. در اینجا پارامترهای رگرسیونی تفسیر شرطی برای یک خوشه یا یک فرد دارند. یعنی β^* بر حسب یک واحد تغییر در متغیر کمکی، زمانی که b_i ثابت در نظر گرفته می شودبیان می شود (۸). معمول ترین روش برای ثابت در نظر گرفتن اثر تصادفی آن است که میانگین شرطی متغیر پاسخ را در نظر بگیریم بنابراین بر خلاف β در مدلهای حاشیه ای، β^* بر اساس تغییر در میانگین متغیر پاسخ در هر خوشه و یا هر فرد تفسیر می شود و پارامترهای رگرسیونی β^* در مورد تغیرات متغیر پاسخ در جامعه اطلاعی را نمی‌دهند.

در صورتی که بخواهیم میانگین حاشیه ای متغیر پاسخ را بدست آوریم باید از توزیع اثرات تصادفی b_i میانگین بگیریم. که این مستلزم میانگین گرفتن از یک تابع غیر خطی از اثرات تصادفی ، b_i،است.

$$\mu_i = E(y_i)$$
$$= E\{E(y_i \mid b_i)\}$$
$$= E\{h(X_i\beta^* + z_i b_i)\}$$

برای میانگین گرفتن از یک تابع غیر خطی از اثرات تصادفی، میانگین وزنی بر اساس توزیع اثرات تصادفی می گیریم.

$$\mu_i = E(y_i) = E\{h(X_i\beta^* + z_i b_i)\} = \int_{-\infty}^{\infty} h(X_i\beta^* + z_i b_i) f(b_i) db_i$$

در معادله بالا ، انتگرال گیری معادل میانگین گرفتن می باشد و مقادیر تابع چگالی احتمال $f(b_i)$ وزنهای مورد استفاده هستند همانطور که از انتگرال فوق بر می آید

میانگین متغیر پاسخ برابر با $h(X_i\beta)$ نمی باشد.

$$E(y_i) \neq h(X_i\beta) \qquad (۱۱)$$

برای مثال در مدل رگرسیون لجستیک با عرض از مبدا تصادفی داریم :

$$, b_i \approx N(0, \delta_b^2)$$

$$\log it\{E(y_i \mid b_i)\} = X_i\beta^* + b_i$$

میانگین حاشیه ای از متغیر پاسخ برابر است با:

$$\mu_i = E(y_i)$$

$$= E\{E(y_i \mid b_i)\}$$

$$= E\{\frac{e^{(X_i\beta^* + b_i)}}{1+e^{(X_i\beta^* + b_i)}}\}$$

$$= \int_{-\infty}^{\infty} \frac{e^{(X_i\beta^* + b_i)}}{1+e^{(X_i\beta^* + b_i)}} \frac{1}{\sqrt{2\pi}\delta_b^2} e^{\frac{-1}{2}\frac{b_i^2}{\delta_b^2}} db_i$$

این انتگرال به راحتی قابل محاسبه نیست و واضح است که از فرم رگرسیون لجستیک نمی باشد .

(۱۲)

$$\frac{e^{(X_i\beta^* + b_i)}}{1+e^{(X_i\beta^* + b_i)}}$$

رابطه حاصله را می توان بصورت تقریبی بفرم زیر بیان کرد.

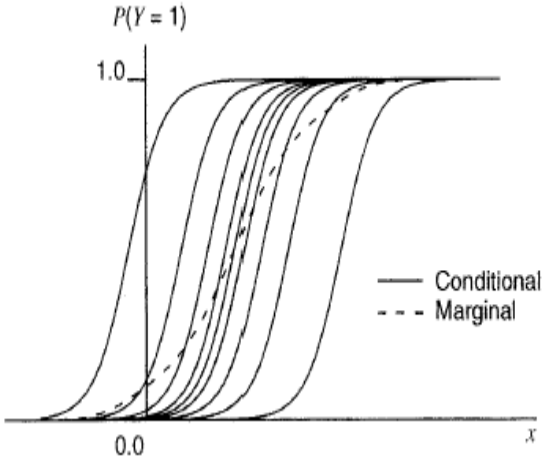
$$\log itE(y_i) \approx (1+k\sigma_b^2)^{-\frac{1}{2}X_i\beta^*}$$

که در آن $k = \frac{16\sqrt{3}}{15\pi}$ طرز رسیدن به این فرمول خیلی مهم نیست. آنچه که آهمیت دارد این است که چگونه ضریب در مدل با عرض از مبدا تصادفی بیشتر از مدل های حاشیه‌ای می گردد.

$$\beta = \frac{\beta^*}{\sqrt{1+0.346\sigma_b^2}} \qquad k^2 = 0.346$$

بنابراین هر گاه $Var(b_i) = \sigma_b^2 > 0$ پارامترهای مدل حاشیه ای رگرسیون لجستیک β مقادیر کوچکتر از پارامترهای مدل رگرسیونی لجستیک با تاثیر تصادفی β^* می‌گردند. و این تفاوت با افزایش σ_b^2افزایش می یابد. در شکل ۱ علت کوچکتر بودن ضرایب در مدل‌های حاشیه‌ای نسبت به مدل با عرض از مبدا تصادفی به نمایش درآمده است

برای یک متغیر کمکی با نام x نمودارهای مختص فردی[1] $P(Y_{ij}=1 \mid b_i)$ را برای افراد گوناگون وقتی تغییرپذیری قابل ملاحظه‌ای وجود دارد نمایش داده است. این تغیر پذیری به علت وجود یک تاثیر تصادفی با σبزرگ است.



در هر مقدار ثابت از x، تغییر پذیری زیادی، در میانگین‌های شرطی وجود دارد $E(Y_{ij} \mid b_i)$ میانگین این میانگین های شرطی (با تاثیر تصادفی) همان میانگین حاصل از مدل حاشیه ای است یعنی $E(Y_{ij})$ این میانگین گیری برای مقادیر مختلف x موجب ایجاد یک منحنی کلی برای مدل حاشیه ای می شود. این منحنی شیب کمتری نسبت به تمامی منحنی‌های مدل‌های با تاثیر تصادفی دارد. در حقیقت این منحنی دقیقا از توزیع لجستیک تبعیت نمی‌کند. چنین خاصیتی برای بسیاری از توابع ربط نیز وجود دارد.

البته لازم به ذکر است که علاوه بر تابع ربط همانی در تابع ربط لگاریتم نیز هر دو ضریب یک تفسیر دارند. مدل با تاثیرات تصادفی عرض از مبدا را در نظر بگیرید.

$$\ln[E(y_i \mid b_i)] = X_i\beta^* + b_i$$

$$\mu_i = E(y_i) = E[E(y_i \mid b_i)] = E\left[e^{X_i\beta^* + b_i}\right] = e^{X_i\beta^*} \times E(e^{b_i})$$

$$\Rightarrow \ln[E(y_i)] = X_i\beta^* + \ln E(e^{b_i})$$

1 -Subject specific

و از آنجایی که e^{b_i} یک عدد ثابت است (اگر $b_i \sim N(0,\sigma_b^2)$ آنگاه $\ln[E(e^{b_i})]=\sigma_b^2$). بنابراین برای دو تابع ربط همانی و لگاریتم ضرایب رگرسیونی یکسان هستند.

اما بطور کلی حتی با انتگرال گیری روی کلیه مقادیر b_i نیز بصورت مستقیم ضرایب مدل حاشیه ای قابل حصول نیست.

بحث

حال با بیان یک مثال عددی از داده های فرضی و یک مثال کاربردی از داده های طرح سلامت وبیماری به بررسی بیشتر مدلهای حاشیه ای و آمیخته و مقایسه بین آنها می پردازیم.

مثال عددی

برای روشن تر شدن مطلب به بیان یک مثال ساده می پردازیم تا تفاوتهای اساسی مدلهای آمیخته و حاشیه ای را مشخص کنیم. یک مجموعه از داده های فرضی را در جدول (۱) در تظر بگیرید .این مجموعه از داده ها نشان دهنده ی ریسک واقعی بیمار شدن سه فرد در شروع مطالعه و پس از درمان با یک داروی جدید به منظور کاهش دادن خطر بیماری می باشد. این سه فرد از نظر

ریسک بیمار شدن در ابتدای مطالعه کاملا متفاوت هستند.که این تفاوت را می توان به صورت اثرات تصادفی b_i بیان کرد. یعنی سه فرد A, B, C دارای ریسک های شدید، متوسط، ضعیف از بیمار شدن هستند اگر فرض کنیم که همه جامعه نیز به نسبت مساوی دارای ریسک های فوق می باشند. آنگاه سطر آخر جدول (۱) میانگین های ریسک جامعه را قبل از درمان ، پس از درمان و اختلاف ریسک را نشان می دهند.

اگر تفاضل ریسک را به عنوان یک مدل خطی از ریسک بیمار شدن در نظر بگیریم.(در اینجا تفاضل ریسک بیماری پس از درمان و قبل از درمان، به عنوان شاخصی از اثر بخشی درمان بکار می رود.) این اختلاف ریسک ها در ستون چهارم جدول (۱) نشان داده شده است و مقادیر آن از ۰.۰۹- تا ۰.۱۷- متغیر است.

این مقادیر اثرات فردی درمان می باشند. حال به دو طریق می توان یک شاخص از اثر بخشی دارو ارائه داد در روش اول از اثرات فردی دارو برای هر سه فرد میانگین می گیریم .

$$\frac{-۰.۱۳-۰.۱۷-۰.۰۹}{۳}=-۰.۱۳$$

در روش دیگر میانگین ریسک بیمار شدن را قبل از درمان (۰.۵)با پس از درمان (۰.۳۷) مقایسه می‌کنیم. این روش در واقع مقایسه ریسک بیمار شدن در جامعه است .

(۰/۳۷-۰/۵۰)= -۰/۱۳

بنابراین تفاضل میانگین های ریسک بیماری قبل از درمان و پس از درمان در جامعه برابر میانگین اختلاف ریسک ها در افراد است .بطور خلاصه « تفاضل میانگین ها » برابر « پمیانگین تفاضل » است. این مثال ساده‌ی عددی موید آن است که تفسیر ضرایب رگرسیونی در مدلهای *خطی* آمیخته و مدلهای حاشیه ای یکسان است.

حال، یک تابع غیر خطی از ریسک بیمار شدن را در نظر می گیریم . لگاریتم نسبت شانس را به عنوان یک تابع غیر خطی، که ریسک بیمار شدن را پس از درمان و قبل از درمان مقایسه می کند در نظر می گیریم. لگاریتم نسبت شانس برای افراد A,B,C در ستون پنجم جدول (۱) نمایش داده شده است.

جدول ۱: داده های فرضی از ریسک بیمار شدن، قبل و پس از درمان، برای سه فرد با ریسک متفاوت برای بیمار شدن

شخص	قبل از درمان	پس از درمان	تفاضل	لگاریتم نسبت شانس
A	۰/۸۰	۰/۶۷	-۰/۱۳	-۰/۶۸
B	۰/۵۰	۰/۳۳	-۰/۱۷	-۰/۷۱
C	۰/۲۰	۰/۱۱	-۰/۹۰	-۰/۷۰
میانگین جامعه	۰/۵۰	۰/۳۷	۰/۱۳	

برای مثال لگاریتم نسبت شانس برای فرد A برابر است با:

$$\log\{\frac{۰.۶۷/(۱-۰.۶۷)}{۰.۸/(۱-۰.۸)}\} = -۰.۶۸$$

مقادیر نسبت شانس برای افراد مورد مطالعه خیلی شبیه هستند. و از ۰.۶۸- تا ۰.۷۱- در حال تغییرند. که این مقادیر اثرات فردی دارو بر ریسک بیماری می باشند. بار دیگر به دو روش می توانیم یک شاخص کلی از اثر بخشی دارو ارائه بدهیم. در روش اول از اثرات فردی، که همان مقادیر لگاریتم نسبت شانس برای تک تک افراد هستند میانگین می‌گیریم:

$$\frac{-۰.۶۸-۰.۷۱-۰.۷۰}{۳} = -۰.۶۹۷$$

این مقدار نشان دهنده ی آن است که درمان، خطر بیمار شدن را به نصف کاهش می دهد. (زیرا $e^{-۰.۶۹۷} \approx ۰.۵$)

در روشی دیگر به عنوان شاخصی از اثر بخشی درمان لگاریتم نسبت شانس قبل از درمان، $\log(\frac{۰.۵}{۰.۵}) = ۰$ را با لگاریتم نسبت شانس پس از درمان، $\log(\frac{۰.۳۷}{۰.۶۳}) = -۰.۵۳۲$ مقایسه می‌کنیم که در نتیجه این مقایسه مقدار ۰/۵۳۲- برای تفاضل لگاریتم نسبت شانس برای هر فرد حاصل می شود که از ۰/۶۹۷- کوچکتر است. بنابراین در یک تابع غیر خطی از ریسک بیمار شدن « مقایسه غیر خطی از میانگین ها » برابر با « میانگین مقایسه های غیر خطی » نیست.

حال این سوال پیش می آید که کدام یک از دو مقدار بدست آمده برای اثر بخشی دارو واقعیت را نشان می دهند؟ جواب این سوال این است که هر دوی این مقادیر ،مقادیری واقعی هستند اما به سوالهای متفاوتی پاسخ می دهند.

برآورد ۰/۶۹۷- یک معیار از تغییرات مورد انتظار در ریسک بیماری برای فردی است که با دارو درمان شده است. یعنی خطر بیماری برای فردی که دارو را استفاده می کنند به میزان ٪۵۰ درصد کاهش می یابد. ($e^{-۰.۵۳۲} \approx ۰.۵۰ - ۱$)

این برآورد برای پزشک معالج بیمار ارزش ویژه ای دارد زیرا او علاقه مند به دانستن اثر متوسط درمان بر روی هر فرد است. از طرف دیگر، مقدار برآورد ۰/۵۳۲- نشان دهنده ی تغییر در شیوع بیماری در جامعه مورد مطالعه است زمانیکه تمام افراد جامعه تحت درمان باشند. یعنی با مصرف دارو در جامعه شانس بیمار شدن در جامعه ٪۴۰ کاهش می یابد. ($e^{-۰.۵۳۲} - ۱ \approx ۰.۴$) این مقدارِ برآورد برای محققان و سیاست گذاران بهداشت حائز اهمیت است. زیرا نشان دهنده ی اثر کلی درمان در سطح جامعه است.

مثال کاربردی

در این بخش به مقایسه مدلهای آماری رگرسیون لجستیک معمولی، حاشیه ای و آمیخته در تعیین عوامل موثر بر تغذیه با شیر مادر خواهیم پرداخت. داده‌های مورد استفاده در این مثال مربوط به مطالعه‌ی بررسی سلامت و بیماری است. طرح سلامت و بیماری در سال ۱۳۷۸ در سطح کشور جمهوری اسلامی ایران اجرا شده که در آن ۱۳۴۷۵ خانوار بر اساس یک نمونه گیری چند مرحله ایی خوشه ائی و به نسبت یک هزارم جمعیت کل خانوارهای کشور و بطور تصادفی انتخاب گردیده اند

از مجموع ۱۳۴۷۵ خانوار طرح سلامت ۱۷۵۶خانوار کودک زیر دو سال دارند، که در این ۱۷۵۶خانوار مجموعا۲۰۱۰ کودک کمتر از۲سال وجود دارند.که نمونه مورد مطالعه را تشکیل می دهند. جهت تعیین عوامل موثر بر تغذیه با شیر مادر در کودکان کمتر از دو سال، بر اساس اطلاعات پرسشنامه کودکان به دو گروه استفاده کننده از شیر مادر (Y=1) و کودکان که با شیر مادر تغذیه نمی شوند (Y=0) تقسیم شدند.

نتایج حاصل از سه مدل مورد بررسی در جدول ۲ ارائه شده است با نگاهی به این جدول در می یابیم که خطای معیار ضرائب متغیرها در مدل آمیخته و حاشیه‌ای نسبت به مدل معمولی که در آن همبستگی درون خوشه‌ای در نظر گرفته شده بیشتر برآورد شده است مقایسه P-مقدارهای این مدل‌ها نشان داد که در مدل معمولی P مقدار کمتری نسبت به دو مدل آمیخته و حاشیه‌ای دارد.

با مقایسه مدل حاشیه‌ای و آمیخته در می‌یابیم که با وجود اینکه در هر دو مدل خطای معیار ضرایب افزایش یافته، اما میزان افزایش خطای معیار در مدل حاشیه ای به اندازه مدل آمیخته نیست. از آنجایی که مدلهای آمیخته و حاشیه‌ای روشهای کاملا متفاوتی برای در نظر گرفتن

همبستگی مشاهدات دارند خطای معیار حاصل از این مدلها متفاوت است.

همانطور که در جدول ۲ مشاهده می‌شود متغیرهای

بنابراین تفسیر ضرایب رگرسیونی در مدل آمیخته وابسته به اثرات تصادفی است.

جدول ۲: مقایسه نتایج سه مدل رگرسیونی در بررسی عوامل موثر بر تغذیه با شیر مادر

متغیر	رده	رگرسیون لجستیک					
		معمولی		حاشیه‌ای		آمیخته	
		(خطای معیار) ضریب	P	(خطای معیار) ضریب	P	(خطای معیار) ضریب	P
سن کودک(ماه)		-۰/۱۰۶ (۰/۰۱۲)	<۰/۰۰۱	-۰/۰۹۸ (۰/۰۱۴)	<۰/۰۰۱	-۰/۱۰۴ (۰/۰۲۱)	<۰/۰۰۱
جنس کودک	مرد	پایه		پایه		پایه	
	زن	-۰/۴۲۸ (۰/۱۴۹)	۰/۰۰۲	-۰/۳۸۵ (۰/۱۷۴)	۰/۰۰۵	-۰/۴۴۷ (۰/۱۸۱)	۰/۰۲۷
سن اولین	۲۵ و کمتر	پایه		پایه		پایه	
	بالاتر از ۲۵	-۰/۶۰۶ (۰/۲۳۵)	۰/۰۴۰	-۰/۵۲۳ (۰/۲۷۶)	۰/۰۲۱	-۰/۵۷۵ (۰/۳۴۸)	۰/۰۵۹
فاصله دو تولد آخر	۲۴ و کمتر	پایه		پایه		پایه	
	بیش از ۲۴	۰/۵۷۵ (۰/۱۴۸)	<۰/۰۰۱	۰/۵۰۵ (۰/۱۷۳)	<۰/۰۰۱	۰/۵۸۷ (۰/۲۴۷)	۰/۰۰۴

جنس کودک، سن اولین حاملگی، و فاصله تولد در مدل حاشیه‌ای در مقایسه با مدل آمیخته اثر معنی دارتری بر روی تغذیه با شیر مادر دارند که این به دلیل است که خطای معیار در مدل آمیخته بیشتر است.

همانطور که دیدیم تفسیر ضرایب رگرسیونی در مدلهای حاشیه‌ای و مدلهای آمیخته کاملا متفاوت می‌باشد و این دو مدل به دو جامعه هدف متفاوت اشاره دارند در مدلهای حاشیه ای ضرایب رگرسیونی مقدار تغییرات در میانگین جامعه یا مقدار تبدیل یافته میانگین جامعه را به ازای تغییر در متغیر مستقل نشان می‌دهند. و یک مشخصه منحصربفرد مدلهای حاشیه‌ای در آن است که مدلهای رگرسیونی برای میانگین متغیر پاسخ و الگوی همبستگی مشاهدات به صورت مجزا تعیین می شوند. تفکیک مدل رگرسیونی برای میانگین متغیر پاسخ و مدل همبستگی داده ها از یکدیگر موجب می شود که تفسیر ضرایب رگرسیونی در مدل حاشیه ای مستقل از الگوی در نظر گرفته شده برای وابستگی مشاهدات باشد و همانطور که از لفظ حاشیه ای برمی آید ضرایب رگرسیونی وابسته به هیچ اثری جز متغیر های کمکی نیستند.

در مقایسه با مدلهای حاشیه ای اساس مدلهای آمیخته، تغییر پذیری ذاتی در خوشه‌هاست و همبستگی بین مشاهدات از این واقعیت ناشی می‌شود که ضرایب رگرسیونی برای مجموعه‌ای از مشاهدات یکسان است.

به عنوان مثال . در مدل حاشیه‌ای با هر ماه افزایش سن کودک لگاریتم شانس تغذیه با شیر مادر در جامعه به میزان ۰/۰۹۸ کاهش می‌یابد در حالیکه براساس مدل آمیخته با هر ماه افزایش سن کودک لگاریتم شانس تغذیه با شیر مادر در یک خوشه بطور متوسط ۰/۱۰۴ کاهش می‌یابد.

نتیجه گیری

زمانیکه تابع ربط به فرم همانی و یک تابع خطی باشد تفسیر پارامتر های رگرسیونی در مدل های آمیخته همانند مدلهای حاشیه ای است. در حالیکه وقتی تابع ربط مورد استفاده به شکل غیر خطی است تفسیر پارامترهای رگرسیونی در دو مدل کاملا متفاوت است. تفسیر نتایج برآوردهای حاصل از مدل آمیخته برای پزشکان معالج ارزش ویژه‌ای دارد زیرا آنها علاقه مند به دانستن اثر متوسط درمان بر روی هر فرد هستند. از طرف دیگر، تفسیر نتایج برآوردهای حاصل از مدل حاشیه‌ای تغییر در پارامتر مورد بررسی در جامعه مورد مطالعه است این مقدارِ برآورد برای محققان و سیاست گذاران بهداشت حائز اهمیت است. گرچه باید دقت کرد که در برخی از توابع ربط مانند همانی این دوضریب مفهوم یکسانی می‌یابند و هر دو کاربرد را دارند.

References

1. Lee Y, Nelder J, Conditional and Marginal Models: Another View, Statist Sci, 2004;27(2):20.
2. Fitzmaurice GM, Larid N, Ware J, Applied longitunal analysis, New Jersy: wiley & sons; 2004.
3. Agresti A, Categorical Data Analysis Hoboken, New Jersey :wiley & sons; 2003.
4. Carriere I, Bouyer J, Choosing marginal or random-effects models for longitudinal binary responses: application to self-reported disability among older persons, BMC Med Res Methodol, 2002; Dec 5;2:15.
5. Hadgu A, Koch G, Westrom L,Analysis of ectopic pregnancy data using marginal and conditional models, Stat Med , 1997 ;Nov 15;16(21):2403-17.
6. Ten Have TR, Ratcliffe SJ, Reboussin BA, Miller ME, Deviations from the population-averaged versus cluster-specific relationship for clustered binary data, Stat Methods Med Res, 2004 Feb;13(1):3-16.
7. Lindsey JK, Lambert P, On the appropriateness of marginal models for repeated measurements in clinical trials, Stat Med, 1998 Feb 28;17(4):447-69.
8. Larsen K, Petersen JH, Budtz-Jorgensen E ,Endahl L, Interpreting parameters in the logistic regression model with random effects, Biometrics , 2000 Sep;56(3):909-14.

Comparison of marginal and mixed models in medical data analysis

yekaninejad MS[1], Yaseri M[1], Nourijelyani K[2], Akaberi A[3]*

[1] PH.D student of Biostatistics, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
[2] Associated professor of Biostatistics, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
[3] M.Sc of Biostatistic,North Khorasan University of Medical Sciences, Bojnurd, Iran

***Corresponding Author:** Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran*
Email: m.yaseri@yahoo.com

Abstract

Background & Objectives: *In medical researchers, there are lots of correlated data which cannot be analyzed using the usual classical statistical methods because the assumption of independency between observations is not met. Data from cluster sampling, longitudinal studies, observations on paired organs and matched studies are examples of such data.*

Materials & Methods: *Two statistical methods that can be used to correctly handle these kinds of data are marginal and mixed models. These models are different in considering correlation between subjects and Interpretation of the regression coefficients. These models were compared in this paper.*

Results: *The regression coefficients in marginal models with non-identity link functions show the change in population whereas in mixed models they represent changes within a subject or a cluster.*

Conclusion: *In result, in nonlinear models, application of these two kinds of models depends on the areas of their usage. While the marginal models are more attractive to the Health policy makers who are considering the potential effects of a variable on the population as a whole, the mixed model will be of most interest to a physician in a physician/patients context.*

Key words: *Marginal model, Mixed model, Linear model, Non linear model*