کومش- جلد ١٤، شماره ١ (پیاپی ٤٥)، پائیز ١٣٩١

# تحلیل داده‌های پنلی در مطالعات پزشکی

حبیب‌الله اسماعیلی [١](Ph.D)، ملوک هادی علیجانوند [*٢](M.Sc)، حسن دوستی [٣](Ph.D)، محمدتقی شاکری [١](Ph.D)

١- *دانشگاه علوم پزشکی مشهد، دانشکده بهداشت، گروه آمار زیستی و اپیدمیولوژی*

٢- *دانشگاه علوم پزشکی شهرکرد، دانشکده بهداشت، گروه آمار زیستی و اپیدمیولوژی*

٣- *دانشگاه تربیت معلم تهران، گروه آمار*

## چکیده

**سابقه و هدف: مطالعه‌ای که موقعیت تکرار عضوهای نمونه، نقاط زمانی باشند را مطالعه طولی نامیده می شودم (Longitudinal study). مطالعه پنلی(Panel study) دسته‌ای از مطالعه طولی است که با ترکیب داده‌های مقطعی و سری زمانی، چند اتفاق را در چند دوره زمانی مورد بررسی قرار می‌دهد و جایگاه خاصی در مطالعات پزشکی از جمله کارآزمایی بالینی دارد. در این مقاله ابتدا ماهیت داده‌های پنلی، انواع مدل‌بندی آماری داده‌های پنلی آن مورد بررسی قرار گرفته است، با توجه به هم‌بستگی درون مجموعه مشاهدات هر فرد در داده‌های پنلی، نمی‌توان تحلیل آماری را با استفاده از رگرسیون معمولی انجام داد. در ادامه معادلات برآوردی تعمیم یافته ( Generalized estimation equation GEE) برای برآورد ضرایب رگرسیون پنلی ضمن منظور نمودن هم‌بستگی بین مشاهدات پیشنهاد شد.**

**مواد و روش‌ها: اهمیت مدل‌بندی داده‌های پنلی با روش GEE در نمونه‌ای از داده‌های واقعی مربوط به تأثیر پچ (چسب)های استروژنی در درمان افسردگی بعد از زایمان نشان داده شد.**

**یافته‌ها: در برآورد ضرایب مدل رگرسیون پنلی داده‌های واقعی، اثر متغیر افسردگی قبل مطالعه با ضریب ۰/۴۲۸ و p<۰/۰۰۱، گروه‌درمانی با ضریب ۴/۰۲۵- و p<۰/۰۰۱ و متغیر دوره درمانی با ۱/۲۱۸- و p<۰/۰۰۱ با استفاده از روش نیرومند نسبت به بدتشخیصی ساختار هم‌بستگی GEE معنادار شد.**

**نتیجه‌گیری: با توجه به این‌که ساختار هم‌بستگی بین پاسخ‌های داده‌های پنلی می‌تواند نقش مهمی در تحلیل داده‌ها در مطالعات پزشکی داشته باشد، برآورد ضرایب رگرسیون داده‌های پنلی با استفاده از روش GEE پیشنهاد می‌شود.**

**واژه‌های کلیدی: مطالعه طولی، داده‌های پنلی، پژوهش در زیست پزشکی**

## مقدمه

کار آزمایی بالینی دسته مهمی از مطالعات پزشکی جهت بررسی سیر بیماری و تأثیر روش‌های مختلف درمانی است. در این گونه مطالعات متغیر پاسخ برای هر فرد در چندین نوبت متوالی مشاهده می‌شوند. به مطالعه‌ای که اندازه‌گیری مربوط به یک صفت در طول زمان مورد بررسی قرار می‌گیرد مطالعه طولی می‌گویند. یکی از اشکالات این‌گونه مطالعات این است که به دلایل مختلف، مشاهدات مربوط به برخی واحدها تا انتهای مطالعه ثبت نمی‌گردد و با داده‌های گم‌شده روبه‌رو خواهیم شد [١]. تحلیل این‌گونه از داده‌ها را برای پاسخ‌های دو حالتی [٢] و هم‌چنین برای پاسخ‌های رتبه‌ای به کار برده‌اند [٣]. اما یک سری از داده‌های دیگر که به آن داده‌های پنلی می‌گویند نیز در علوم پزشکی مورد توجه قرار می‌گیرد. این داده‌ها در مطالعات طولی، ترکیب دوبعدی از

* نویسنده مسئول، تلفن: ٠٩٣٨١٥٥١٠٨٩ نمابر: ٣٣٣٤٦٧٨ - ٠٣٨١
Email: moluk.hadi@gmail.com

تاریخ دریافت: ١٣٨٩/١١/١١ تاریخ پذیرش: ١٣٩١/١/٢٧

داده‌های مقطعی (مشاهداتی در یک نقطه زمانی) و داده‌های سری زمانی (مشاهده یک واحد خاص در یک دوره زمانی) می‌باشند. روش‌های مدل‌بندی آماری مطالعات طولی براساس اندازه‌گیری‌های مکرر، سری‌های زمانی و آنالیز بقا انجام می‌شود [۴]. و مدل‌بندی آماری داده‌های پنلی، با ترکیبی از اندازه‌گیری‌های مکرر و سری‌های زمانی انجام می‌شود. کاربرد تئوری آماری تحلیل داده‌های پنلی در مباحث اقتصادی، شناسایی و تحلیل داده‌های پنلی در مقایسه با مطالعات پزشکی بیش‌تر مورد توجه قرار گرفته‌اند و مطالعات مشابه زیادی در راستای شناخت و تحلیل آماری داده‌های پنلی انجام شده است که در ادامه به اختصار، به برخی از آن‌ها اشاره می‌شود.

داده‌های پنلی به دو دسته متوازن و نامتوازن تقسیم‌بندی می‌شوند، در داده‌های پنلی متوازن، مشاهدات در هر مقطع زمانی وجود دارد ولی در داده‌های پنلی نامتوازن نمی‌توان مشاهداتی در هر مقطع زمانی در طول دوره مطالعه داشت [۵]. اغلب در مطالعات پزشکی از جمله کارآزمایی بالینی و مطالعات هم‌گروهی تحلیل داده‌های پنلی بایستی در نظر گرفته شود. در کارآزمائی بالینی ممکن است به دلایل مختلف از جمله ملاحظات اخلاقی، شرکت‌کننده می‌تواند هر زمان که بخواهد از مطالعه‌ای که به صورت طولی طراحی گردیده است، خارج شود. بنابراین با داده‌های پنلی نامتوازن نیز رو به رو خواهیم گردید. لازم به ذکر است که در داده‌های پنلی متوازن مشاهدات در همه مقطع زمانی ثبت شده‌اند. ما در این مقاله ضمن بررسی مدل‌های مختلف تحلیل در داده‌های پنلی، کاربرد آن را در مطالعات پزشکی با یک مثال واقعی نشان خواهیم داد تا علاوه بر شناخت این نوع داده‌ها در مطالعات پزشکی به مدل‌بندی، روش برآورد ضرایب رگرسیون پنلی ضمن در نظر گرفتن ساختار هم‌بستگی و به منظور دقت لازم در روش برآورد ضرایب رگرسیون پنلی، استفاده از معادلات تعمیم یافته در روش برآورد ضرایب رگرسیون پنلی در قالب یک مثال کاربردی در مطالعات پزشکی واضح شود.

## مدل‌بندی داده‌های پنلی

**مدل با ضرایب ثابت (رگرسیون ادغام شده) Constant coefficient model (Pooled regression)**. عرض از مبدأ و شیب آن در مدل رگرسیونی پنلی، ثابت می‌باشند و مواقعی که اثر مقاطع (گروه‌های مختلف) و دوره زمانی معنادار نباشد، از این مدل استفاده می‌شود، در این مدل برای همه داده‌ها به طور ادغام شده، یک مدل رگرسیونی معمولی برازش داده می‌شود [۶].

**مدل با اثر ثابت(Fixed effects model)**. شیب‌های رگرسیون پنلی، ثابت هستند ولی عرض از مبدأ به دلیل تفاوت بین گروه‌های مختلف (برای مثال گروه‌های درمانی مختلف)، ثابت ولی متفاوت است. تأثیر دوره زمانی معنادار نیست، ضرایب ثابت مدل رگرسیونی در هر گروه با گروه دیگر متفاوت است و ممکن است خطای مدل، هم‌بستگی با اثر افراد داشته باشد [۶].

**مدل با اثر تصادفی(Random effect model)**. مدل رگرسیونی با ضرایب تصادفی است که در آن پارامترها در همه مقاطع مختلف هستند و خطای مدل، هم‌بستگی با اثر مقاطع دارد [۶].

**مدل پنلی دینامیکی(Dynamic panel models)**. مدل رگرسیونی پنلی است که مشاهدات در زمان‌های مختلف، خود هم‌بسته‌اند [۷]. که آزمونی برای بررسی این خودهم‌بستگی در داده‌های پنلی متوازن با تعمیم آزمون دوربین واتسون (بررسی خودهم‌بستگی مشاهدات در رگرسیون معمولی) پیشنهاد شده است [۸] و در ادامه با اصلاح آزمون آن‌ها، آزمون خودهم‌بستگی را برای داده‌های پنلی متوازن و نامتوازن تعمیم داده‌اند [۹].

داده‌های پنلی استفاده از روش‌های درست‌نمائی ماکزیموم استاندارد (Maximum likelihood estimation MLE) را مشکل‌ساز می‌کند. در نتیجه از روش‌های شبه درست‌نمائی (Quasi-likelihood Q-LE)از جمله روش تعمیم‌یافته گشتاوری (Generalized methods of moments GMM) و معادلات برآوردی تعمیم‌یافته (GEE) بایستی

استفاده شود. از جمله خواص برآوردگر با روش شبه درست‌نمائی این است که برآوردگری سازگار برای ضرایب رگرسیونی است حتی اگر ساختار هم‌بستگی به درستی تشخیص داده نشود ضمن این‌که در صورت درست بودن ماتریس هم‌بستگی این برآوردگر کاراتر از حالت برآوردگری است که نادرست از آن استفاده می‌شود. با وجود این معادلات برآوردی تعمیم‌یافته که از همین روش برای برآورد ضرایب استفاده می‌کند با استفاده از خطای معیار تجربی علاوه بر سازگاری برآورد، مشابه روش شبه درست‌نمائی نیرومند به بد تشخیصی ساختار هم‌بستگی است [۱۰]. لذا برآوردی مناسب برای پارامترهای مدل رگرسیونی داده‌های پنلی با ساختار هم‌بستگی بین مشاهدات ضمن در نظر گرفتن تأثیر مقاطع مکانی (گروه‌های مختلف) و زمانی را با استفاده از معادلات برآوردی تعمیم‌یافته (GEE) می‌توان به دست آورد. معادلات برآوردی تعمیم‌یافته در برآورد پارامترها نسبت به بدتشخیصی توزیع متغیر پاسخ و ساختار هم‌بستگی، نیرومند است. و با به کارگیری این روش در مدل رگرسیون پنلی برای متغیر پاسخ با توزیعی از خانواده نمایی (نرمال، دوجمله‌ای، پواسن، دو جمله‌ای منفی، گاما، بتا و...) می‌توان ضرایب رگرسیون را نااریب و کارا برآورد کرد [۱۱].

**موارد احتیاط استفاده از روش (GEE).** در مواقعی که حجم نمونه در پنل‌ها (مقاطع) کم باشد، برآورد پارامترها با این روش اریب می‌باشد.

**مدل‌بندی آماری داده‌های پنلی**

مدل رگرسیونی در داده‌های پنلی به طور کلی به صورت ذیل می‌باشد:

$$y_{it} = \alpha + \beta' X_{it} + u_{it} \quad ; i = 1,\ldots, N \;\&\; t = 1,\ldots, T$$

که در آن $y_{it}$ مقادیر متغیر پاسخ می‌باشد که کمی بود و Xit ماتریس طرح و uit جمله خطاست.

T ابعاد زمانی و N ابعاد فردی را نشان می‌دهد.

**مدل با اثر ثابت**

$$u_{it} = \mu_i + \nu_{it}$$

در مدل اثر ثابت μiها ویژگی‌هایی هستند که نسبت به زمان پایا می‌باشد.

**مدل با اثرات تصادفی**

در مدل با اثرات تصادفی فرض می‌کنیم که μi و vit دارای توزیع می‌باشند که معمولاً توزیع آن‌ها را نرمال در نظر می‌گیرند:

$$\nu_{it} \sim N\left(0, \sigma_\nu^2\right), \mu_i \sim N\left(0, \sigma_\mu^2\right)$$

و این دو مؤلفه خطای μi و vit از یک‌دیگر مستقل‌اند.

**مدل پنلی دینامیکی**

در این مدل uitها ساختارهای هم‌بستگی مختلفی دارند از جمله:

ماتریس‌های هم‌بستگی به فرم ام–وابسته (M-dependent correlation matrix M-D CM)، بی‌ساختار (Unstructured correlation matrix US CM)، تبادل‌پذیر (Exchangeable correlation catrix EC CM) و خودهم‌بسته (Auto regressive correlation matrix AR CM) مرتبه اول ساختار هم‌بستگی متغیر وابسته را با در نظر گرفتن (مدل اثر ثابت، تصادفی، دینامیکی) و ماهیت داده‌ها نشان می‌دهند که این ساختارها به شکل ذیل است:

و در مدل پنلی دینامیکی با خودهم‌بستگی مرتبه اول (AR (1)) داریم:

$$u_{it} = \rho u_{i(t-1)} + v_{it}$$

انتخاب مدل مناسب

حال اگر بخواهیم یک کارآزمایی بالینی (به عنوان مثال دو یا چند روش درمانی را در گروه‌های مختلف) از نوع داده‌های پنلی را مدل‌بندی کنیم، با پاسخ به سوال پژوهشی مدل مناسب را انتخاب می‌کنیم:

سوال پژوهشی: آیا مشاهدات را می‌توان نمونه‌ای تصادفی از یک جمعیت فرض کرد؟

اگر پاسخ منفی باشد، بایستی داده‌ها با مدل اثر ثابت تحلیل شوند و اگر پاسخ مثبت باشد با هر دو روش مدل‌بندی شود که در ادامه اگر آزمون دوربین ـ ووـ هازمن، (H) معنادار

باشد، مدل با اثر ثابت و در غیر این صورت، مدل با اثر تصادفی به‌کار می‌رود. اگر آماره این آزمون (H)، بزرگ‌تر از مقدار کای اسکوار با K (تعداد عناصرβ) درجه آزادی باشد، آزمون معنادار بوده و مدل با اثر ثابت به کار رفته و در غیر این صورت، مدل با اثر تصادفی به کار می‌رود.

$$H = \left(b^{random} - b^{fixed}\right)' \Sigma^{-1} \left(b^{random} - b^{fixed}\right)$$

bها برآورد β در دو روش با اثر ثابت و اثر تصادفی، Σ تفاضل ماتریس واریانس- کوواریانس برآورد شده در دو روش مذکور است.

متغیرهای پاسخ داده‌های پنلی می‌توانند گسسته یا پیوسته باشند و پارامترها با روشی مناسب مانند GEE در مدل‌های رگرسیون خطی تعمیم‌یافته با در نظر گرفتن توزیع متغیر پاسخ که الزامی برای برقراری فرض نرمال ندارد، برآورد شوند.

برآورد ضرایب مدل رگرسیون پنلی در یک کارآزمایی بالینی جهت بررسی تأثیر روش درمانی جدید در دو گروه دارونما و تحت درمان داروی جدید، در چند مقطع زمانی با استفاده از معادلات برآوردی تعمیم‌یافته به شرح زیر است:

فرض کنید تعداد افراد در گروه درمانی جدید na، تعداد افراد در گروه کنترل np و N = na + np باشد.

$y_{it}$ پاسخ مشاهده شده در زمان (j=1,…,T) و $t_j$ برای فرد i ام (i=1,…,N ) است، نقطه شروع درمان، $t_T$ پایان دوره درمانی $z_{ij}$ تغییرات پاسخ‌های مشاهده شده را از زمان شروع تا $t_j$ نشان می‌دهد.

$$z_{ij} = y_{ij} - y_{i0} = \beta_1 t_j + \beta_2 I\left(t_j > t_1\right)\left(t_j - t_1\right) + \varepsilon_{ij}$$

توزیع $\varepsilon_{i1}, \ldots, \varepsilon_{ij}$ نرمال با میانگین صفر می‌باشد ولی واریانس آن‌ها ثابت نیست و با مشاهدات فرد i ام هم‌بستگی دارند.

$$X_{ij} = \left(t_j, I\left(t_j > t_1\right)\left(t_j - t_1\right)\right) \ , \ b = \left(b_1, b_2\right)^T$$

با استفاده از معادلات برآوردی تعمیم‌یافته (GEE) در صورتی b برای β ناریب خواهد بود که S(β)=0 باشد و

$$S_{na}(\beta) = \frac{1}{\sqrt{na}} \sum_{i=1}^{na} \sum_{j=1}^{T} \left(z_{ij} - \beta^T X_{ij}\right) X_{ij}$$

بنابراین

$$b = \left( \sum_{i=1}^{na} \sum_{j=1}^{T} X_{ij}^T X_{ij} \right)^{-1} \sum_{i=1}^{na} \sum_{j=1}^{T} X_{ij} z_{i_j}$$

در این‌جا استقلال برای ساختار هم‌بستگی فرض شده است.

گام‌های اساسی در برآورد ضرایب رگرسیونی با معادلات برآوردی تعمیم‌یافته، انتخاب تابع ربط مناسب، توزیع متغیر وابسته و ساختار همبستگی آن می‌باشند.

## مواد و روش‌ها

در ادامه با تشخیص داده‌های پنلی نامتوازن در یک کارآزمائی انجام شده به تحلیل آماری آن با روش مذکور پرداختیم که به شرح ذیل است:

گریگوری، کامراوریت، هندرسن و استد در سال ۱۹۹۶، تأثیر پچ (چسب)های استروژنی را در درمان افسردگی بعد از زایمان بر روی ۶۱ زن مورد بررسی قرار دادند. به طور تصادفی، ۲۷ نفر از زنان، دارونما و ۳۴ نفر از آن‌ها پچ (چسب)های استروژنی را استفاده کردند و با اندازه‌گیری ماهیانه مقیاس افسردگی بعد از زایمان ادینبرگ، افسردگی در بیماران به مدت ۶ ماه مطالعه شده است [۱۲]. داده‌های مذکور در جدول ۱ دسته‌بندی شده‌اند.

برای مدل‌بندی این داده‌ها، ۷ متغیر افسردگی قبل از مطالعه، افسردگی ماه اول، افسردگی ماه دوم، افسردگی ماه سوم، افسردگی ماه چهارم، افسردگی ماه پنجم، افسردگی ماه ششم که به ترتیب میزان افسردگی (با مقیاس ادینبرگ) قبل از شروع دوره مطالعه، میزان افسردگی در اولین ماه، میزان افسردگی در دومین ماه و ... و میزان افسردگی در ماه ششم و برای متغیر گروه‌درمانی در گروه کنترل (دارونما) کد صفر و برای گروه‌درمانی استروژنی کد یک در نظر گرفته شده است. داده‌ها از نوع داده‌های پنلی نامتوازن‌اند و مدل‌بندی با اثر تصادفی و دینامیکی را به کار برده‌ایم. با استفاده از نرم‌افزار SPSS در ابتدا ماتریس هم‌بستگی بین متغیرهای افسردگی قبل از مطالعه، افسردگی ماه اول، افسردگی ماه دوم، افسردگی ماه سوم، افسردگی ماه چهارم، افسردگی ماه پنجم، افسردگی

ماه ششم را برای بررسی ساختار هم‌بستگی به‌دست آوردیم. بعد از آن، آزمون تی در دو گروه دارونما و استروژن قبل از شروع درمان انجام شد تا از نبود اریبی به دلیل انتخاب گروه‌ها اطمینان حاصل کنیم. در ادامه تحلیل‌های آماری، پنلی و نامتوازن بودن داده‌ها در نظر گرفته شود و یک متغیر جدید با ترکیب متغیرهای افسردگی ماه اول، افسردگی ماه دوم، افسردگی ماه سوم، افسردگی ماه چهارم، افسردگی ماه پنجم، افسردگی ماه ششم را با استفاده از دستور Varstocases به نام افسردگی دوره درمان می‌سازیم و با استفاده از دستور genlin برآوردهای نااریب و کارا توسط معادلات برآوردی تعمیم یافته (GEE) با ساختارهای مختلف هم‌بستگی را به دست می‌آوریم.

جدول ۱. ساختار داده‌های جمع آوری شده در تأثیر پچ‌های استروژنی بر افسردگی بعد از زایمان

| تعداد افراد | گروه درمانی | افسردگی | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | قبل از مطالعه | ماه اول | ماه دوم | ماه سوم | ماه چهارم | ماه پنجم | ماه ششم |
| ۱ | دارونما | ۱۸ | ۱۷ | ۱۸ | ۱۵ | ۱۷ | ۱۴ | ۱۵ |
| ۲ | دارونما | ۲۷ | ۲۶ | ۲۳ | ۱۸ | ۱۷ | ۱۲ | ۱۰ |
| ۳ | دارونما | ۱۶ | ۱۷ | ۱۴ | . | . | . | . |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ۵۹ | استروژن | ۱۷ | ۱۵ | ۱۲ | ۱۵ | . | . | . |
| ۶۰ | استروژن | ۲۲ | ۷ | . | . | . | . | . |
| ۶۱ | استروژن | ۲۶ | ۲۴ | . | . | . | . | . |

## نتایج

در ابتدای تحلیل آماری، آزمون تی در دو گروه دارونما و پچ‌های استروژنی قبل از شروع درمان انجام شد و با p=۰/۶۲۸ معنادار نبود و از نبود اریبی به دلیل انتخاب گروه‌ها اطمینان حاصل شد.

قبل از تحلیل‌های پنلی، تحلیل آنالیز واریانس اندازه‌گیری مکرر را برای مقایسه نتایج انجام دادیم. نتایج خروجی آنالیز واریانس اندازه‌های مکرر (uniANOVA) اثر گروه‌درمانی با p<۰/۰۰۱ و اثر دوره درمانی (افسردگی دوره درمان) با p<۰/۰۰۱ به‌دست آمد.

با در نظر گرفتن ساختار هم‌بستگی بین مشاهدات، از روش معادلات برآوردی تعمیم‌یافته برای برآورد پارامترهای رگرسیون پنلی استفاده شد. و نتایج خروجی دستور genlin برای به دست آوردن برآوردهای نااریب و کارا برای مدل رگرسیون پنلی توسط معادلات برآوردی تعمیم‌یافته GEE در جداول ذیل نشان داده شده است:

برآورد پارامترها در جدول ۲، با استفاده از GEE، تابع ربط identity و توزیع نرمال برای متغیر پاسخ افسردگی و ساختار ماتریس هم‌بستگی مستقل (همانی) است.

جدول ۲. برازش مدل رگرسیونی با روش GEE با ساختار همبستگی مستقل

| ضرایب | برآورد | خطای معیار | P- مقدار |
|---|---|---|---|
| عرض از مبدأ | ۸/۲۳۴ | ۱/۸۰٤ | ۰/۰۰۰ |
| افسردگی قبل از درمان | ۰/٤۷۷ | ۰/۰۸۰ | ۰/۰۰۰ |
| گروه درمانی | -٤/۲۹۱ | ۰/٦۰۷ | ۰/۰۰۰ |
| افسردگی دوره درمان | -۱/۳۰۸ | ۰/۱۷۰ | ۰/۰۰۰ |

اما با توجه به ماهیت داده‌ها و ضریب هم‌بستگی بین افسردگی قبل از شروع مطالعه و ماه‌های اول تا ششم بعید به نظر می‌رسد که یک میزان افسردگی در دوره درمانی از دیگری مستقل باشد. لذا مدل‌های ذیل را با ساختار هم‌بستگی تبادل‌پذیر و خودهم‌بسته مرتبه اول در نظر گرفتیم.

برآورد پارامترها در جدول ۳، با استفاده از GEE، تابع ربط همانی و توزیع نرمال برای متغیر پاسخ افسردگی و ساختار ماتریس هم‌بستگی تبادل‌پذیر (EC CM) در نظر می‌گیریم.

برای برآورد پارامترها در جدول ۴، با استفاده از GEE، توزیع نرمال برای متغیر پاسخ افسردگی و ماتریس هم‌بستگی خودهم‌بسته مرتبه اول، برای لحاظ کردن ساختار هم‌بستگی مشاهدات در زمان‌های مختلف، در نظر می‌گیریم.

جدول ۳. برازش مدل رگرسیونی با روش GEE با ساختار همبستگی تبادل پذیر (با در نظر گرفتن r= ۰/۵۵۱)

| ضرایب | برآورد | خطای معیار | P- مقدار |
|---|---|---|---|
| عرض از مبدأ | ۸/۲۳۴ | ۳/۷۰۳۳ | ۰/۰۲۳ |
| افسردگی قبل از مطالعه | ۰/٤٦۰ | ۰/۱۷۲۱ | ۰/۰۰۸ |
| گروه درمانی | -٤/۰۲٦ | ۱/۰٤۷۸ | ۰/۰۰۰ |
| افسردگی دوره درمان | -۱/۲۲۷ | ۰/۱٦٤۹ | ۰/۰۰۰ |

جدول ۴. برازش مدل رگرسیونی با روش GEE با ساختار همبستگی خود همبسته مرتبه اول (با در نظر گرفتن r= ۰/۶۹۴)

| ضرایب | برآورد | خطای معیار | P- مقدار |
|---|---|---|---|
| عرض از مبدأ | ۹/۱۴۸ | ۳/٤٦٤۲ | ۰/۰۰۸ |
| افسردگی قبل از مطالعه | ۰/٤۲۸ | ۰/۱٦۰٦ | ۰/۰۰۸ |
| گروه درمانی | -٤/۰۲٥ | ۰/۹۷۰۸ | ۰/۰۰۰ |
| افسردگی دوره درمان | -۱/۲۱۸ | ۰/۱۷۸٤ | ۰/۰۰۰ |

جدول ۵. برازش مدل رگرسیونی با روش GEE با ساختار همبستگی خود همبسته مرتبه اول و اثر متقابل گروه درمانی و افسردگی دوره درمان (با در نظر گرفتن r= ۰/۶۹۴)

| ضرایب | برآورد | خطای معیار | P- مقدار |
|---|---|---|---|
| عرض از مبدأ | ۹/۱۴۸ | ۳/٤٦٤۲ | ۰/۰۰۸ |
| افسردگی قبل از مطالعه | ۰/٤۲۸ | ۰/۱٦۰٦ | ۰/۰۰۸ |
| گروه درمانی | -٤/۰۲٥ | ۰/۹۷۰۸ | ۰/۰۰۰ |
| افسردگی دوره درمان | -۱/۲۱۸ | ۰/۱۷۸٤ | ۰/۰۰۰ |
| افسردگی دوره درمان × گروه درمانی | -۰/۲۴۸ | ۰/۳۷۳٤ | ۰/۵۰۷ |

با توجه به ماهیت داده‌ها و ضریب‌های همبستگی ساختار همبستگی خودهمبسته اول منطقی‌تر به نظر می‌رسد ولی روش برآورد معادلات تعمیم‌یافته نسبت به بدتشخیصی ساختار همبستگی نیرومند است و نتایج برآورد تغییر چشم‌گیری ندارد.

همان‌طور که ملاحظه می‌شود با توجه به جداول ۲ تا ۵ برآوردهای رگرسیون پنلی با روش GEE در ساختارهای همبستگی مختلف، تفاوت قابل ملاحظه‌ای ندارند.

در مدل رگرسیونی که متغیر وابسته در دسته‌بندی‌های مختلف و متغیرهای مستقل وابسته به زمان در مطالعات طولی هستند (داده‌های پنلی)، کارایی برآوردهای پارامترهای رگرسیونی با زیاد شدن هم‌بستگی مشاهدات کاهش می‌یابد و این کاهش در هم‌بستگی‌های بیش‌تر از ۰/۴ چشم‌گیر است [۱۳]. بنابراین با توجه به این که مقدار اکثر هم‌بستگی‌های مشاهده شده بین متغیرهای افسردگی قبل از مطالعه، افسردگی ماه اول، افسردگی ماه دوم، افسردگی ماه سوم، افسردگی ماه چهارم، افسردگی ماه پنجم و افسردگی ماه ششم در یافته‌های مثال کاربردی بیش‌تر از ۰/۴ مشاهده شد، بیانگر این است که استفاده از رگرسیون بدون توجه به هم‌بستگی بین متغیرهای از دقت کافی برخوردار نیست، روش‌های آنالیز واریانس اندازه‌گیری‌های مکرر برای رفع این مشکل مناسب نیست. زیرا با این روش، هم‌بستگی درون مشاهدات تکراری در نظر گرفته نمی‌شوند. روش مذکور به طور معمول، برای مجموعه کامل و متوازنی از داده‌ها با متغیر پاسخ نرمال استفاده می‌شود و نمی‌توان مدل رگرسیونی که متغیرهای مستقل آن در زمان‌های مختلف تغییر می‌کنند را با این روش تحلیل کرد [۱۴]. در مثال کاربردی با استفاده از روش آنالیز واریانس اندازه‌گیری مکرر اثر گروه‌درمانی و طول دوره درمانی معنادار شد اما در این روش ساختار هم‌بستگی لحاظ نشد. لذا راه حل استفاده از معادلات برآوردی تعمیم‌یافته، برای برآوردهای رگرسیونی مطالعات طولی و طرح‌های اندازه‌گیری مکرر با متغیرهای پاسخ غیرنرمال با ساختارهای مختلف هم‌بستگی پیشنهاد شد [۱۱]. این مدل با نیرومند بودن به ساختارهای مختلف هم‌بستگی و معنادار نبودن اثر متقابل با p=۰/۵۰۷ نشان می‌دهد که با کنترل متغیرهای افسردگی دوره درمان و افسردگی قبل از مطالعه و استفاده از پچ (چسب)های استروژنی میزان افسردگی ۴ واحد کاهش و با کنترل گروه درمانی و افسردگی قبل از مطالعه و یک واحد افزایش متغیر افسردگی دوره درمان میزان افسردگی ۱/۲ واحد در ساختارهای هم‌بستگی مختلف، کاهش خواهد یافت. در آخر می‌توان گفت برآورد پارامترهای رگرسیونی ضمن در نظر گرفتن هم‌بستگی درون مشاهدات از دقت بیش‌تری برخوردار است. با استفاده از روش تعمیم‌یافته گشتاوری (GMM) نیز که

در مقدمه اشاره شد در مواقعی که متغیر وابسته گسسته یا پیوسته باشد، می‌توان برآورد نااریب، کارا و نیرومند نسبت به بدتشخیصی توزیع متغیر پاسخ برای ضرائب مدل رگرسیون پنلی به‌دست آورد اما نقصان این روش نسبت به روش GEE در این است که روش تعمیم‌یافته گشتاوری (GMM) نسبت به بد تشخیصی ساختار هم‌بستگی نیرومند نیست و در صورتی ساختار هم‌بستگی در این روش به درستی تشخیص داده نشود از کارائی برآورد ضرائب کاسته می‌شود. و یا به عبارتی دیگر زمانی تحت فرض این‌که ساختار ماتریس کوواریانس (ساختار هم‌بستگی) در روش GEE به درستی تشخیص داده شود برآوردگر GEE می‌تواند مانند روش‌های تعمیم‌یافته برآوردگر گشتاوری تفسیر شود. موارد احتیاط هر دو روش مذکور در این است که حجم نمونه در پنل‌ها (مقاطع) کم باشد، ضمن این‌که از روش GEE در برآورد ضرایب رگرسیون پنلی در داده‌های پنلی نامتوازن می‌توان استفاده کرد. لذا در مثال کاربردی برای داده‌های پنلی در مطالعات پزشکی ترجیح داده شد که از روش GEE که در واقع گسترش یافته روش GMM است که هم نسبت به توزیع متغیر پاسخ و هم نسبت به ساختار هم‌بستگی نیرومند است، در این پژوهش استفاده شود و سعی بر آن شده است به این روش برآورد پارامترها در داده‌های پنلی که به وفور در مطالعات پزشکی (کارآزمائی بالینی و مطالعه هم‌گروهی) مورد استفاده می‌گیرد، توجه کافی شود.

## بحث و نتیجه‌گیری

در پژوهش حاضر سعی بر آن شد تا با شناخت داده‌های پنلی در مطالعات پزشکی، اشاره به مدل‌بندی آماری و روش‌های برآورد ضرایب رگرسیون پنلی، تحلیل رگرسیونی این داده‌ها در قالب مثال کاربردی، اهمیت تحلیل داده‌های پنلی را در بررسی سیر بیماری و تأثیر روش‌های درمانی نشان دهیم. از آن‌جا که گام اساسی برای بررسی سیر بیماری و تأثیر روش‌های درمانی در مطالعات پزشکی استفاده مناسب از مدل‌بندی‌های آماری و به دنبال آن تصمیم‌گیری‌های آماری در جهت تصمیم‌گیری‌های پزشکی است. و با توجه به این‌که نمی‌توان رگرسیون معمولی و در همین راستا روش‌های آنالیز واریانس اندازه‌گیری‌های مکرر را به دلیل وجود هم‌بستگی مشاهدات و امکان نرمال نبودن متغیر پاسخ در این نوع از داده‌ها به کار برد، تحلیل آماری داده‌های پنلی را در مطالعات پزشکی ضرورت دانستیم. به دنبال روش برآورد ضرایب رگرسیون پنلی به دو روش تعمیم‌یافته گشتاوری (GMM) و معادلات برآوردی تعمیم‌یافته (GEE) اشاره شد که هر دو روش نسبت به بد تشخیصی توزیع متغیر پاسخ نیرومند هستند اما علاوه بر ویژگی مذکور این دو روش، بایستی به در نظر گرفتن ساختار هم‌بستگی در برآورد ضرایب مذکور توجه لازم را داشت که این دو روش ضمن در نظر گرفتن ساختار هم‌بستگی در رگرسیون پنلی برآوردی نااریب و کارا را برای ضرایب رگرسیون پنلی به‌دست می‌آورند. با این حال برتری روش GEE نسبت به روش GMM در این است که روش GMM نسبت به بد تشخیصی ساختار هم‌بستگی رگرسیون پنلی نیرومند نیست و به همین دلیل کارائی این روش در مواقعی که ساختار هم‌بستگی به درستی تشخیص داده نمی‌شود، کاهش می‌یابد. در ادامه با ساختار هم‌بستگی مختلف در داده‌های پنلی، از معادلات برآوردی تعمیم‌یافته (GEE) برای به دست آوردن برآوردی نااریب و کارا در مدل‌های رگرسیونی پنلی در یک مثال کاربردی کارآزمائی بالینی استفاده شد که نتایج برآورد ضرایب رگرسیون پنلی با استفاده از معادلات برآوردی تعمیم‌یافته، نمایانگر نیرومند بودن این روش نسبت به بد تشخیصی ساختار هم‌بستگی داده‌های پنلی است.

## تشکر و قدردانی

نویسنده بر خود لازم می‌داند که مراتب سپاس و قدردانی خود را به حضور استاد راهنما و اساتید مشاور که از راهنمائی آنان در این پژوهش بهره‌مند بوده، ابراز دارد.

## منابع

[1] Esmaeili H, Meshkani MR, Arghami NR, Kazemnejhad A. Application of Mont Carlo sampling in analyse incomplete longitudinal binary responses with bayesian methods. J Rec Med Sci 1382; 2: 17-22. (Persian).

[2] Esmaily H, Meshkani MR, Arghami NR, Kazemnejad A. Application of analysis of incomplete longitudinal binary responses with Bayesian methods in the effects of Lidocaine and Lidocaine/Morphine on pain after root canal therapy. Iran J Basic Med Sci 1381; 1: 1-5. (Persian).

[3] Ghorbani R, Faghih Zadeh S, Meshkani MR, Alavi Majd H, Esmaeili H. Analysis of longitudinal ordered categorical response with missing data: A bayesian approach. Daneshvar Med 1383; 54: 71-60. (Persian).

[4] Petrie A, Sabin C. Medical statistics at a glance. Blackwell Sci 2000; 39-41,101-106.

[5] Doughtery C. Introduction to econometrics. 2nd Ed. Oxford Univ Press 2007; 408-421.

[6] Wooldridge JM. Analysis of cross section and panel data. MIT press 2002.

[7] Greene WH. Econometric analysis. 5th ed. Upper Saddle River. Pren Hall 2003; 285: 291,293,304.

[8] Bhargava A, Franzini L, Narendranathan W. Serial correlation and the fixed effects models. Econom Stud 1982; 49: 533-549.

[9] Baltagi BH, Wu PX. Unequally spaced panel data regression with AR (1) disturbances. Econometr Theor 1999; 15: 814-823.

[10] Raymond H, Myers Douglas C, Montgomery G. Geoffrey Vening. Translated by: Nirumand HA. Generalized linear models with applications in engineering and the sciences. Mashad Univ Med Sci 1384; 233-274. (Persian).

[11] Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986; 42: 121-130.

[12] Gregoire AJ, Kumar R, Everitt B, Henderson AF, Studd JW. Transdermal estrogen for the treatment of severe post-natal depression. Lancet 1996; 347: 930-933.

[13] Fitzmaurice GM. A caveat concerning independence estimating equations with multivariate binary data. Biometrics 1995; 51: 309-317.

[14] Diggle PJ, Heagerty P, Liang KY, Zeger SL. Analysis of longitudinal data. Oxford Univ Press 1994; 1-346.

# Panel data analysis in medical research

Habib ollah Esmaili (Ph.D)[1], Molok Hadialijanvand (M.Sc)[*2], Hasan Doosti (Ph.D)[3], Mohammad taghi Shakeri (Ph.D)[1]

*1 - Dept. of Biostatistics and Epidemiology, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran*
*2 – Dept. of Biostatistics & Epidemiology, School of Health, Shahrekord University of Medical Sciences, Shahrekord, Iran*
*3 - Tarbiat Moallem University, Tehran, Iran*

(Received: 31 Jan 2011 Accepted: 15 Apr 2011)

***Introduction:*** A longitudinal study involves repeated observations of the same items over long periods of time. Panel studies are longitudinal studies of batch which combine cross-sectional and time- series data in observations on a number of over time that play special role in medical research or clinical trials. In this paper, we primarily discuss about the panel data and various modeling of panel data. It is not possible to use ordinary regression methods due to inter- correlation of observation related to one unit. Generalized estimation equation (GEE) for estimation of panel regression coefficient by considering inter correlation was used among observations.

***Materials and Methods:*** We considerd the importance of panel data modeling with GEE in real panel data samples as applications in the effect of estrogen patches on the postnatal depression.

***Results:*** In the estimation of panel regression coefficients on real data, starting treatment effect (b(pre)=0.428, p<0.001), treatment group effect (b(group)=4.025, p<0.001) and treatment period effect (b(visit)=1.218,p<0.001 with use of robust method to misspecification of the correlation structure were significant.

***Conclusion:*** According to the important role of panel data inter- correlation structure in analyzing in medical research, we propose to estimate the coefficients of panel via GEE method.

***Keywords:*** Longitudinal Studies, Panel Data, Biomedical research

* Corresponding author: Tel: +98 9381551089; Fax: +98 381 3334678
moluk.hadi@gmail.com