بسم الله الرحمن الرحيم

دانشگاه علوم بهزیستی وتوانبخشی

گروه آمار و کامپیوتر

پایان نامه کارشناسی ارشد

رشته آمار زیستی

عنوان

کاربرد روش بیزی در برآورد پارامترهای مدل رگرسیون لوجستیک با مقادیر گمشده تصادفی (MAR) در متغیر کمکی

نگارنده

الهه کاظمی

استاد راهنما

دکتر مسعود کریملو

استاد مشاور

دکتر مهدی رهگذر

تابستان ۱۳۹۰

شماره ثبت ۱۰۳-۶۰۰۰

University of Social Welfare and Rehabilitation Sciences

Department of Biostatistics and Computer

M. Sc. Thesis

Application of Bayesian Method in Parameters Estimation of Logistic Regression Model with Missing at Random (MAR) Covariate

By:

Elaheh Kazemi

Supervisor:

Dr. Masoud Karimlou

Dr. Mehdi Rahgozar

Summer 2011

تقدیم به :

پدر و مادر مهربانم

که در سختی ها و مشکلات زندگی همواره در کنارم بوده اند و مرا در تمامی مراحل زندگی حمایت کرده اند.

تقدیر و تشکر :

بدینوسیله بر خود واجب می دانم که از اساتید محترم گروه نهایت قدردانی و سپاس را داشته باشم.

استاد ارجمندم جناب آقای دکتر مسعود کریملو استاد راهنمای اینجانب که در کلیه مراحل انجام این پایان نامه از راهنمایی های فکری و علمی ایشان بهره مند بوده ام.

استاد ارجمندم جناب آقای دکتر مهدی رهگذر استاد مشاور اینجانب که افتخار شاگردی ایشان موهبت بزرگی برای من بوده است.

استاد ارجمندم جناب آقای دکتر عنایت الله بخشی که از نظرات ارزشمند ایشان در انجام این پایان نامه بهره برده ام.

سرکار خانم دکتر ایمانه عسگری و سرکار خانم مینو دیانت خواه که داده های مورد استفاده در این پژوهش را در اختیار اینجانب قرار داده اند.

دوستان هم دوره ای سرکار خانم زهره رزاقی و سرکار خانم لیلا چراغی که دوران سخت تحصیل در کنار ایشان بسیار خاطره انگیز بود.

چکیده :

رگرسیون لوجستیک مدلی عمومی برای تحلیل داده های پزشکی و اپیدمیولوژیکی می باشد و اخیراً محققین معدودی تحقیقات خود را به تحلیل مدل های رگرسیون لوجستیک با وجود مقادیر گمشده در متغیرهای کمکی معطوف داشته اند. در بسیاری از پژوهش ها محققین با مجموعه داده هایی مواجه هستند که دارای مقادیر گمشده است. گمشدگی تهدید عمده ای برای درستی نتایج حاصل از مجموعه داده ها محسوب می شوند و اجتناب از آن بسیار مشکل است.

تمامی روش های برآورد پارامترها بر پایه فرض کامل بودن مجموعه داده ها استوار است و تحت برقراری این شرایط منجر به برآوردهایی نا اریب می شوند. در مطالعات انجام شده ثابت شده است که در صورت وجود مقادیر گمشده تصادفی در مجموعه داده ها برآوردهای حاصل، دیگر نااریب نخواهند بود و با افزایش نسبت گمشدگی ، مقدار اریبی نیز افزایش خواهد یافت.

ساتن و کارول تابع درستنمایی ویژه ای را برای برآورد پارامترهای مدل رگرسیون لوجستیک وقتی که مقادیر کمکی Z به طور کامل مشاهده شده و مقادیر متغیر کمکی X دارای مقادیر گمشده از نوع مکانیسم گمشدگی تصادفی (MAR) باشند، معرفی کرده اند. در این پژوهش از این تابع درستنمایی در تحلیل بیزی برای برآورد پارامترهای مدل رگرسیون لوجستیک استفاده شده است و نتایج به دست آمده با روش های جانهی چندگانه و واحد کامل مقایسه شده است.

روش های مذکور را بر روی داده های شبیه سازی شده و داده های دندانپزشکی اجرا کرده و پس از مقایسه نتایج حاصل از سه روش مذکور نتیجه گرفته شد که اگر مکانیسم گمشدگی تصادفی باشد، به کارگیری تحلیل بیزی با تکنیک MCMC منجر به برآوردهای دقیق تری نسبت به روش جانهی چند گانه و روش واحد کامل می شود.

واژه های کلیدی : رگرسیون لوجستیک، مکانیسم گمشدگی تصادفی (MAR)، روش بیزی، زنجیرهای مارکوف مونت کارلویی (MCMC)، جانهی چندگانه، روش واحد کامل، DMFT.

فهرست مطالب

فهرست جداول

فهرست نمودارها و تصاویر

فصل اول :

کلیات تحقیق

۱-۱ : بیان مسئله

۱-۱-۱ : مقدمه

مدل رگرسیون لوجستیک روشی تحلیلی است که به طور وسیعی در تحقیقات پزشکی و اپیدمیولوژیکی کاربرد دارد. با گسترش و تنوع این مدل ها تجزیه و تحلیل داده های حاصل از تحقیقات در علوم مختلف با به کار گیری این مدل ها روز به روز افزایش یافته و به لحاظ نظری لزوم تحقیق در زمینه های گوناگون این مدل ها را بیش از پیش فراهم نموده است.

هدف از تحلیل رگرسیون لوجستیک همانند مدل های رگرسیون معمولی دستیابی به مدلی مناسب و در عین حال ساده جهت بررسی ارتباط بین متغیر پاسخ (وابسته)[۱] با یک یا مجموعه ای از متغیر های مستقل (کمکی)[۲] است. با این ویژگی که در این گونه مدل ها متغیر پاسخ بر خلاف رگرسیون معمولی عموماً از نوع رسته ای دو یا چند حالتی می باشد.

از آن جا که مدل رگرسیون لوجستیک مدلی غیر خطی است، در آمار کلاسیک، تحلیل این گونه مدل ها مبتنی بر برآورد پارامترها از طریق ماکسیمم درستنمایی[۳] است، این روش ایده آل ترین روش است، چرا که در آن هیچ گونه محدودیتی برای متغیرهای مستقل در نظر گرفته نمی شود. به طور معمول دو روش متفاوت ماکسیمم درستنمایی برای برآورد پارامترهای مدل رگرسیون لوجستیک وجود دارد که ماکسیمم درستنمایی غیر شرطی[۴] و شرطی[۵] نامیده می شود.

1. Response variable ، dependent variable
2 .Independent (Covariate) variable
3 .Maximum likelihood
4 .Unconditional
5 .Conditional

یک روش استنباط آماری، روش بیزی[1] است که در آن اطلاعات ناشی از مطالعات قبلی و یا تجارب شخصی را می توان به کمک توزیع پیشین پارامترها در مدل لحاظ نمود. در روش بیزی لازم است با تلفیق داده های مشاهده شده و توزیع های پیشین[2]، توزیع های پسین[3] پارامترهای مدل را به دست آورد و بر اساس آن ها استنباط درباره پارامترهای مدل را انجام داد. هدف این پژوهش بررسی و تحلیل بیزی با استفاده از تابع درستنمایی ساتن و کارول و مقایسه این روش با روش های واحد کامل[4] و جانهی چندگانه[5] می باشد. انتظار بر این است که با استفاده از روش روش بیزی برآوردهای دقیق تر و نتایج بهتری نسبت به روش های معمول به دست بیاوریم.

در بسیاری از مطالعات با مجموعه داده هایی مواجه می شویم که بخشی از آن ها گزارش نشده اند از قبیل خود داری از پاسخ، عدم تکمیل کامل پرسشنامه ها یا پرونده ها، ناقص بودن چارچوب مطالعه و غیره. در این صورت با داده های گمشده سروکار داریم که می تواند در متغیر پاسخ یا در متغیرهای کمکی بوجود آید. در این پژوهش گمشدگی در متغیرهای کمکی مورد نظر می باشد. مواجهه با داده گمشده مشکل بومی تحقیقات اجتماعی، پزشکی و اپیدمیولوژیکی است و در هنگام تحلیل، وجود این گونه موارد مشکلات عدیده ای را فراهم می سازد و عملاً تجزیه و تحلیل آماری را به سوی نتایج اریب سوق داده و نهایتاً دستیابی به یک نتیجه گیری مفید از داده های جمع آوری شده را با مشکل مواجه می سازد. ساده ترین روش برای تجزیه و تحلیل چنین داده هایی صرفه نظر کردن از موردهای دارای مقادیر گمشده و انجام آنالیز با داده های کامل می باشد (روش واحد کامل)، که این روش در عمل کارا نیست [1].

به طور کلی سه روش در نحوه بررسی داده های گمشده مورد استفاده قرار می گیرد [2] :

۱: روش های مبتنی بر واحد های کامل

۲: روش های مبتنی بر جانهی[6]

1 .Bayesian method
2 .Prior distributions
3 .Posterior distributions
4 .Complete Case
5 .Multiple Imputation
6 .Imputation

۳: روش های مبتنی بر مدل

تعریف :

داده ی گمشده : در تحلیل های آماری معمولاً داده ها به صورت مجموعه ای ماتریسی اند. سطرهای ماتریس، مشاهدات، آزمودنی ها[1] یا واحدهای مورد مطالعه و ستون های آن متغیرها هستند. متغیرها به دو گروه متغیرهای مستقل یا کمکی و متغیرهای پاسخ تقسیم بندی می شوند. هر گاه برخی از مشاهدات این ماتریس اندازه گیری نشده و یا مقدار آن در دسترس نباشد با موضوع مقادیر گمشده سروکار داریم. گمشدگی مقادیر ممکن است در متغیر پاسخ یا متغیرهای کمکی رخ دهد. گم شدن در متغیرهای کمکی به دلایل مختلفی می تواند رخ دهد. به عنوان مثال در یک مطالعه گذشته نگر به علت نقص مدارک و سوابق ممکن است برخی از اطلاعات در دسترس نباشد، مانند مطالعات اپیدمیولوژیکی که از موارد ثبت شده در پرونده پزشکی بیماران برای تعیین مقادیر متغیرها استفاده می شود ممکن است پرونده ها برای همه بیماران شرکت کننده در پژوهش به طور کامل پر نشده باشند. در برخی مطالعات گمشدگی در مقادیر برخی از متغیرها ممکن است به علت نوع طرح پژوهشی رخ داده باشد، به طور مثال در مطالعه ای برای جمع آوری داده های مورد نظر، یک نمونه گیری دو مرحله ای انجام می شود، در مرحله اول مقادیر متغیرهایی که اندازه گیری آن ها آسان و ارزان است برای تمامی افراد شرکت کننده در مطالعه جمع آوری می شود و سپس در مرحله دوم مقادیر متغیرهایی که اندازه گیری آن ها پر هزینه و پیچیده می باشد برای زیر مجموعه ای از افراد شرکت کننده در مطالعه جمع آوری می شود [3]. همچنین این احتمال هست که به علت نقص یا ضعف دستگاه و تجهیزات، امکان مشاهده و اندازه گیری وجود نداشته باشد و یا در بعضی از مطالعات نظر سنجی، افرادی قادر به اظهار نظر دقیق نباشند. در این حالت ها با داده های گمشـده روبرو هستیم [1,4].

1.Observation ، Case، Unit