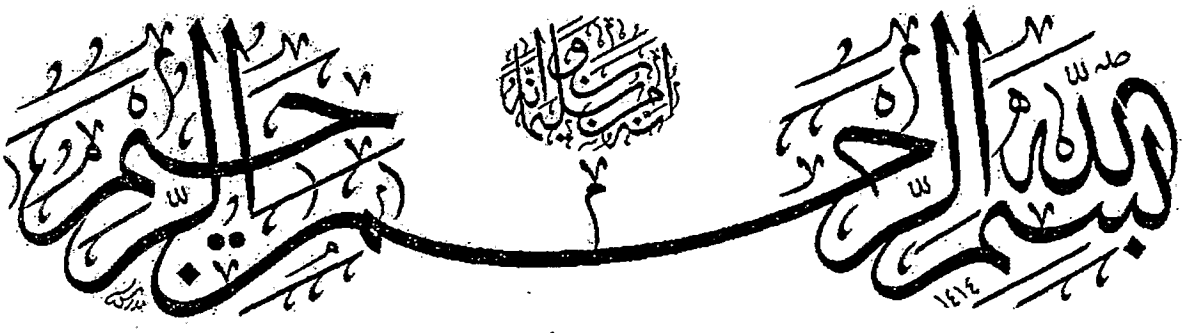

دانشگاه اصفهان

دانشکده علوم

گروه آمار

پایان نامه‌ی کارشناسی ارشد رشته‌ی آمار گرایش ریاضی

رهیافت بیزی نیمه پارامتری و کاربرد آن در مدل‌های خطی با اثرهای آمیخته

استاد راهنما :

دکتر ایرج کاظمی

پژوهشگر:

کیوان عسلی‌صاف

شهریور ماه ۱۳۸۸

دانشگاه اصفهان
دانشکده علوم
گروه آمار

پایان‌نامه‌ی کارشناسی ارشد رشته‌ی آمار گرایش ریاضی

آقای کیوان عسلی صاف

تحت عنوان

رهیافت بیزی نیمه پارامتری و کاربرد آن در مدلهای خطی با اثرهای آمیخته

در تاریخ ۸۸/۶/۲۵ توسط هیأت داوران زیر بررسی با نمره ۱۸/۱۹ و با درجه **عالی** به تصویب نهایی رسید.

۱- استاد راهنمای پایان‌نامه دکتر ایرج کاظمی با مرتبه‌ی علمی استادیار امضاء

۲ – استاد داور داخل گروه دکتر هوشنگ طالبی با مرتبه‌ی علمی استادیار امضاء

۳- استاد داور خارج از گروه دکتر حمید پزشک با مرتبه‌ی علمی استاد امضاء

گروه آمار امضای مدیر گروه

سپاسگزاری

سپاس خدای را مرا شوق دانستن و توان ادراک آموخت.

مراتب سپاس و تشکر خود را از استاد گرانقدرم جناب آقای دکتر ایرج کاظمی ابراز می‌دارم که در تمام مراحل تدوین این رساله با تلاش‌های بی‌وقفه و راهنمایی‌های ارزنده خود مرا یاری نمودند. همچنین از زحمات بی‌دریغ اساتید داور جناب آقای دکتر هوشنگ طالبی و جناب آقای دکتر حمید پزشک کمال تشکر را دارم.

از پدرم که سربلند زیستن و از مادرم که صبر و ایثار را به من آموختند، نهایت تشکر و قدردانی را دارم.

تقدیم به

پدر و مادرم

به پاس تعبیر عظیم و انسانی‌شان از واژه ایثار،

به پاس عاطفه سرشار و گرمای امیدبخش وجودشان

و به پاس محبت‌های بی‌دریغشان که هرگز فروکش نمی‌کند.

چکیده

متداول‌ترین مدل‌ها برای تحلیل داده‌های وابسته در علوم مختلف، از جمله مطالعات زیستی، اقتصادسنجی و پزشکی و غیره، مدل‌های خطی تعمیم‌یافته با اثرهای آمیخته است. از آنجایی که استفاده از روش‌های مناسب برآوردیابی پارامترهای این مدل‌ها حائز اهمیت فراوان در حصول نتایج می‌باشد و با توجه به استفاده چشم‌گیر از آمار بیز در برازش مدل‌های پیچیده، این پایان‌نامه با به‌کارگیری روش‌های بیز نیمه‌پارامتری به مطالعه این موضوع می‌پردازد. یک مسأله اساسی در استفاده از این مدل‌ها بکارگیری فرآیند مناسبی برای تولید توزیع اثرهای آمیخته است. این پایان‌نامه، با فرض اینکه توزیع این اثرها از فرآیند دیریکله با توزیع پایه و پارامتر دقت مشخص پیروی می‌کنند، مدل‌های بیزی سلسله مراتب نیمه‌پارامتری را برای تحلیل داده‌های وابسته در رگرسیون خطی تعمیم‌یافته مورد بررسی قرار می‌دهد. بدین منظور، ابتدا فرآیند دیریکله معرفی و چگونگی تولید آن توسط روش‌های مختلف شرح داده می‌شود. سپس روش بکارگیری آن به عنوان فرآیند تولید توزیع‌های پیشین در مدل‌های مختلف توضیح داده خواهد شد. در ادامه، توزیع‌های پسین شرطی کامل پارامترهای مدل‌های رگرسیونی با اثرهای آمیخته را یافته و توسط رهیافت بیز، استنباط آماری را با الگوریتم نمونه‌گیرگیبز انجام خواهیم داد. همچنین مدل‌های متفاوتی به داده‌های واقعی از علوم مختلف برازش داده خواهد شد تا اهمیت کاربرد روش‌های بیز نیمه‌پارامتری در مقایسه با روش‌های متداول مشخص شود.

کلید واژه‌ها : مدل‌های رگرسیون خطی تعمیم‌یافته، مدل‌های بیز سلسله مراتب، اثرهای آمیخته، فرآیند دیریکله، نمونه‌گیری گیبز، توزیع‌های پسین شرطی کامل.

فهرست مطالب

عنوان صفحه

عنوان | صفحه

فصل چهارم : برازش مدل آمیخته فرآیند دیریکله

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

پیوست‌ها

عنوان صفحه

فصل اول

کلیات

۱-۱ مقدمه

در این فصل ابتدا موضوع تحقیق را بیان کرده و سپس هدف اصلی از انجام این تحقیق را ذکر می‌کنیم. در ادامه به بررسی اهمیت و کاربرد موضوع موردنظر خواهیم پرداخت و در این راستا به بیان زمینه و تاریخچه موضوع تحقیق نیز اشاره کامل خواهد شد. لازم به ذکر است که در تمام فصول این پایان‌نامه از حرف بزرگ لاتین برای نمایش متغیر تصادفی و از پررنگ کردن حروف بزرگ لاتین برای نمایش بردار و ماتریس استفاده کرده‌ایم.

۱-۲ موضوع و هدف تحقیق

موضوع این پایان‌نامه رهیافت بیزی نیمه پارامتری و کاربرد آن در مدل‌های با اثرهای آمیخته می‌باشد. در این پایان‌نامه سعی شده است که روش‌های جدید مرتبط با موضوع تحقیق مورد بحث قرار گیرند. از جمله به بررسی مسئله برآوردیابی پارامترها و مدل‌سازی در حالت‌های مختلف خطی خواهیم پرداخت. هدف از انجام این تحقیق، معرفی راهکار جدیدی برای برآورد پارامترهای مجهول مدل‌های مورد بررسی می‌باشدکه با توجه به مفهوم توزیع‌های آمیخته با رویکرد بیزی و اتخاذ فرآیند دیریکله به عنوان پیشین برای اثرات تصادفی، به بحث در زمینه تحلیل داده‌های وابسته و بخصوص موضوع تغییرپذیری بین واحدها خواهیم پرداخت.

۱-۳ اهمیت و کاربرد موضوع

از آنجایی که برای تحلیل داده‌های وابسته در علوم مختلف، به طور متداول از مدل‌های خطی تعمیم‌یافته با اثرهای آمیخته استفاده می‌شود، لذا استفاده از روش‌های مناسب برازش مدل و برآوردیابی پارامترهای این مدل‌ها حائز اهمیت فراوان در حصول نتایج می‌باشد. با توجه به گسترش زیاد آمار بیز در برازش مدل‌ها، در این پایان‌نامه با به‌کارگیری روش‌های بیز نیمه‌پارامتری به مطالعه برآورد پارامترهای مدل‌های مذکور می‌پردازیم.

واقف بودن محقق به پاره‌ای از اطلاعات قبلی در مورد مدل‌های با اثرات آمیخته و پارامترهای آن اهمیت ویژه‌ای در حصول نتایج برآوردیابی دارد بنابراین یک مسأله اساسی در استفاده از این مدل‌ها بکارگیری فرآیند مناسبی برای تولید توزیع پیشین اثرهای آمیخته است. در این پایان‌نامه، فرض می‌کنیم که توزیع این اثرها از فرآیند دیریکله با توزیع پایه و پارامتر دقت مشخص پیروی می‌کنند. در نهایت برای مقایسه روش ارائه شده با روش‌های متداول بیز به برآورد پارامترهای مجهول و اثرات آمیخته در مدل‌های خطی می‌پردازیم و سپس با توجه به معیارهای اعتبار مدل‌ها از قبیل معیار اطلاع پراکندگی (DIC[1]) ، AIC[2] و پراکندگی[3] و غیره که در فصل سوم به طور مفصل مورد بحث قرار خواهند گرفت، به مقایسه و سنجش روش ارائه شده با روش‌های معمول بیز خواهیم پرداخت.

۱-۴ زمینه و تاریخچه موضوع تحقیق

کارهای تحقیقاتی بر روی توزیع‌های آمیخته دارای قدمت زیادی در علم آمار می‌باشد. برای مثال باتاچاریا در سال ۱۹۶۷ توزیع طول ماهی‌ها را به صورت یک جامعه آمیخته در نظر گرفته و از آن در تعیین سن گروه‌ها استفاده کرد. دقت کنید که هر گروه سنی یک توزیع متفاوت از نظر طولی دارد به طوری که طول ماهی‌ها، برخی اطلاعات را در مورد سن فراهم می آورد.

جهت یک آشنایی کلی با موضوع مورد بحث، فرض کنید که برای Y_i ، $i = 1, 2, ..., k$ ها به شرط پارامترهای θ_i به صورت آمیخته‌ای از توزیع‌های $f(.)$ که در پارامترهایشان متفاوت هستند، پیروی کنند. همچنین فرض کنید که θ_iها دارای توزیع یکسان G_0 باشند. فرگوسن[4] (۱۹۷۳) این مدل را به صورت سلسله‌مراتبی بیزی به صورت زیر بیان نمود

---

[1] Deviance Informatin Criterion
[2] Akaike Information Criterion
[3] Deviance
[4] Ferguson (1973)

$$
\begin{aligned}
Y_i \,|c_i, \boldsymbol{\theta} &\sim f\left(Y \,|\theta_{c_i}\right) \\
\mathbf{c}|\pi_{1:k} &\sim \text{discrete}\left(\pi_1, \pi_2, ..., \pi_k\right) \\
\theta_i &\sim \mathrm{G}_0(.) \\
\left(\pi_1, \pi_2, ..., \pi_k\right) &\sim \mathrm{Dir}\left(\alpha, M\right)
\end{aligned}
$$

که در آن c_iها نشانگرهایی هستند که مقادیر Y_i را به یک مقدار θ_{c_i} نسبت می‌دهند و π_iها ضرایب آمیختگی هستند که نامنفی بوده و مجموعشان یک است. $\mathrm{Dir}(\alpha,\mathrm{M})$ نیز نشانگر توزیع دیریکله با پارامترهای α و M می‌باشد. برای بیان مفهوم توزیع دیریکله فرض کنید $\boldsymbol{\Theta}=\{\theta_1,...,\theta_n\}$ یک توزیع احتمال روی فضای گسسته $\mathbf{X}=\{X_1,X_2,...,X_n\}$ باشد، به طوری که $P(X=X_i)=\theta_i$ که در آن X یک متغیر تصادفی در فضای $\mathbf{X}$ می باشد. توزیع دیریکله روی $\boldsymbol{\Theta}$ به صورت زیر می باشد

$$
P\left(\boldsymbol{\Theta}|\alpha,\mathbf{M}\right)=\frac{\Gamma(\alpha)}{\prod_{i=1}^{n}\Gamma(\alpha m_i)}\prod_{i=1}^{n}\theta_i^{\alpha m_i-1}
$$

که در آن $i=1,...,n$ و $\mathbf{M}=\{m_1,...,m_n\}$ اندازه پایه تعریف شده روی $\mathbf{X}$ و بردار میانگین $\boldsymbol{\Theta}$ و α پارامتر دقتی است که چگونگی تمرکز توزیع حول $\mathbf{M}$ را نشان می دهد. بزرگ بودن α اطمینان ما را درمورد انتخاب مناسب اندازه پایه $\mathbf{M}$ بیشتر می‌کند و از این رو توزیع بیشتر حول $\mathbf{M}$ متمرکز می‌شود. برای مقادیر مفروض ضرایب آمیختگی فرض معمول بر آن است که نشانگرهای c_i دارای توزیع چندجمله‌ای باشند. حال اگر تعداد مولفه‌های جامعه آمیخته زیاد باشد و به عبارت دیگر وقتی که $k\to\infty$ آنتونیاک[1] (۱۹۷۴) مدل آمیخته فوق را به صورت

$$
\begin{aligned}
Y_i \,|\boldsymbol{\theta} &\sim f\left(Y \,|\theta_i\right) \\
\theta_i &\sim \mathrm{G}(.) \\
G &\sim \mathrm{DP}\left(\alpha G_0(\theta)\right)
\end{aligned}
$$

[1] Antoniak (1974)

نشان داد که در آن G یک توزیع عمومی و آزاد می‌باشد و از فرآیند دیریکله $\mathrm{DP}(\alpha G_0(\theta))$ با اندازه پایه G_0 و پارامتر دقت α پیروی می‌کند.

کاربردهای اساسی ساختار فوق را که به مدل بیزی سلسله‌مراتبی نیمه پارامتری معروف است، اسکوبار و وست (۱۹۹۸) و همچنین کلینمن و ابراهیم[1] (۱۹۹۸) در مدل‌های خطی تعمیم‌یافته با اثرات آمیخته بیان کردند. واژه نیمه‌پارامتری از آن جهت به این مدل‌ها اطلاق می‌گردد که شکل توزیع $f(Y|\theta_i)$ برای محقق معلوم می‌باشد و به عبارت دیگر این توزیع بخش پارامتری ساختار را شرح می‌دهد و از طرفی $\theta_i \sim \mathrm{G}(.)$ ، که در آن G یک توزیع عمومی و آزاد می‌باشد، بخش ناپارامتری ساختار را بیان می‌کند. لذا با تلفیق دو بخش ذکر شده مدل فوق را مدل بیزی سلسله‌مراتبی نیمه پارامتری می‌گوییم.

ایشوارن و جیمز[2] (۲۰۰۱) توزیع‌های شرطی کامل پارامترهای مدل با اثرات آمیخته را برای انجام استنباط بیزی به دست آوردند. اسپیگل هالتر[3] (۲۰۰۲) معیارهای سنجش اعتبار مدل‌ها از قبیل DIC و پراکندگی را برای مقایسه چند مدل برازش داده‌شده به داده‌ها معرفی و به کار برد. به طوری که مدل با DIC کمتر به بقیه مدل‌ها ارجحیت دارد. اولسن و همکاران[4] (۲۰۰۶) یک مدل رگرسیون خطی تعمیم‌یافته با اثرات آمیخته را برای بررسی اثرات بیمارستان‌های تحت بررسی بر روی بهبود بیماران به داده‌ها برازش دادند و با تخصیص فرآیند دیریکله به عنوان پیشین به اثرات بیمارستان‌ها با استفاده از روش‌های بیز نیمه‌پارامتری و نیز با بکارگیری طرح نمونه‌گیرگیبز با نرم‌افزار WinBUGS به برآورد و تحقیق در مورد اثرات بیمارستان‌ها پرداختند.

---

1 Kleinma. and Ibrahim (1998)
2 Ishwaran and James (2001)
3 Spiegelhalter et al (2002)
4 Ohlssen et al (2006)

فصل دوم

تعاریف و مفاهیم مقدماتی

۲-۱ مقدمه

در این فصل ابتدا مطالعه‌ای مقدماتی برمفهوم آمار بیز و استفاده از آن در برآوردیابی پارامترها خواهیم داشت. از این رو برای آشنایی بیشتر با آمار بیز، نخست به بررسی مفاهیم توزیع پیشین و پسین خواهیم پرداخت که به نوعی اساس و بنیاد مبحث آمار بیز می‌باشند، سپس با بررسی توابع زیان، مخاطره و مخاطره بیز به بیان مفهوم برآوردگرهای پسین بیزی و همچنین برآوردگرهای بیزی خواهیم پرداخت. در ادامه به دلیل اینکه روش‌های عددی و تحلیلی برای استنباط و استخراج نمونه به جهت برآورد پارامترهای مجهول، در بسیاری از توزیع‌های پسین پیچیده میسر نمی‌باشد، از این رو یکی از روش‌های متداول مبتنی بر دنباله‌های مونت کالو را که به نمونه‌گیر گیبز معروف است، بیان می‌کنیم. در انتهای این فصل نیز به بیان تعاریف، مفاهیم و قضایایی خواهیم پرداخت که در فصل‌های بعدی ازآنها استفاده می‌کنیم. لازم به ذکر است که تمام توابع چگالی احتمال را با نماد $P(.)$ یا $f(.)$، متغیرهای تصادفی را با حروف بزرگ لاتین و بردارها و ماتریس‌ها را با پررنگ کردن حروف نمایش داده‌ایم.

۲-۲ برآوردگرهای پسین بیز[1]

در بررسی مسائل برآورد، اغلب فرض می‌کنیم که نمونه تصادفی از یک تابع چگالی احتمال $P(.;\theta)$ که در آن $P(.;\theta)$ معلوم فرض می‌شود، بدست آمده است. علاوه بر این، فرض می‌کنیم که θ هر چند برای ما نامعلوم است، ثابت می‌باشد. در بسیاری از کاربردهای واقعی که از چگالی $P(.;\theta)$ استفاده می‌شود، اغلب اطلاعات بیشتری درباره θ وجود دارد. برای مثال، پژوهشگر ممکن است شواهدی داشته باشد که خود θ در نقش یک متغیر تصادفی عمل کند و در نتیجه بتواند برای آن یک تابع چگالی احتمال منطقی را در نظر بگیرد. برای مثال، فرض کنید ماشینی که قطعات اتومبیل را تولید می‌کند قرار است از نقطه نظر نسبت معیوب‌های تولید بررسی شود. در یک روز معین، ۱۰ قطعه از خروجی ماشین مورد آزمایش قرار گرفته و مشاهدات را با $X_{10},...X_2,X_1$ نشان می‌دهیم که در آن اگر قطعه iام معیوب باشد، $X_i=1$ واگر سالم باشد $X_i=0$ که آنها را می توان به عنوان یک نمونه تصادفی به حجم ۱۰ از چگالی احتمال برنولی

$$P(X;\theta)=\theta^x(1-\theta)^{1-x}$$

در نظر گرفت که در آن $x=0,1$ و $0\le\theta\le1$. این چگالی احتمال نشان می‌دهد که احتمال اینکه قطعه مفروضی معیوب باشد برابر با عدد نامعلوم θ است. چگالی احتمال توأم ۱۰ متغیر تصادفی $X_{10},...,X_2,X_1$ عبارت است از

$$\theta^{\sum_{i=1}^{10}x_i}(1-\theta)^{10-\sum_{i=1}^{n}x_i}$$

می‌دانیم که برآوردگر ماکزیمم درستنمایی[2] θ، برابر $\hat{\theta}=\bar{X}$ است. روش برآورد گشتاوری[3] هم همین برآوردگر را نتیجه می‌دهد. با وجود این، فرض کنید که پژوهشگر دارای اطلاعات بیشتری درباره θ است و او مشاهده کرده است که در روزهای مختلف، مقدار θ تغییر می‌کند و چنین می‌نماید که این تغییر را می‌توان به عنوان یک متغیر تصادفی با چگالی احتمال $g(\theta)$ نشان داد. یک سؤال مهم این است که، چگونه می‌توان از این اطلاعات اضافی درباره θ برای برآورد θ_0 استفاده کرد، که θ_0 مقداری است که θ در روز استخراج نمونه با آن برابر بوده است.

---

[1] Bayesian Posterior Estimators
[2] Maximum Likelihood Estimator
[3] Moment Estimation

برای بررسی این مسأله، علاوه بر این فرض که نمونه تصادفی از چگالی احتمال $P(.;\theta)$ استخراج شده است فرض می‌کنیم که پارامتر نامعلوم θ مقداری از یک متغیر تصادفی، مانند Θ است. باز هم علاقه‌مند به برآورد کردن تابعی از θ، مانند $\tau(\theta)$ خواهیم بود. اگر Θ متغیری تصادفی باشد پس دارای یک توزیع است. فرض می‌کنیم $G(.)=G_\theta(.)$ تابع توزیع تجمعی Θ و $g(.)=g_\theta(.)$ تابع چگالی احتمال Θ را نشان می‌دهد و این توابع دارای پارامترهای مجهول نیستند.

اگر فرض کنیم توزیع Θ معلوم است، دارای اطلاعات اضافی می‌باشیم. بنابراین یک سؤال مهم این است که چگونه این اطلاعات اضافی را در برآورد به کار برد؟ این مسئله‌ای است که در ادامه به آن می‌پردازیم. در بسیاری از مسائل، این فرض که θ مقداری از یک متغیر تصادفی است، ممکن است غیر واقعی باشد. در سایر مسائل، گرچه فرض اینکه θ مقداری از یک متغیر تصادفی است مناسب به نظر می رسد، ممکن است توزیع Θ معلوم نباشد، یا حتی اگر معلوم باشد، ممکن است شامل پارامترهای نامعلوم دیگری باشد. با این وجود، در برخی مسائل فرض معلوم بودن توزیع Θ واقع بینانه است.

۲-۲-۱ توزیع‌های پیشین و پسین[1]

اکثراً برای هر θ در فضایی که به آن متعلق است، نماد $P(X;\theta)$ را برای نشان دادن تابع چگالی احتمال متغیر X به کار می بریم. هرگاه بخواهیم نشان دهیم که پارامتر θ مقداری از متغیر تصادفی Θ است، چگالی X را به جای $P(X;\theta)$ به صورت $P(X|\theta)$ خواهیم نوشت و ممکن است به جای P از f برای نشان دادن تابع چگالی احتمال استفاده کنیم. باید توجه کنیم که $P(X|\theta)$ یک چگالی شرطی است، یعنی چگالی X به شرط $\Theta=\theta$ است.

فرض کنید $X_1,X_2,...X_n$ یک نمونه تصادفی به حجم n از چگالی $P(.|\theta)$ باشد، که در آن θ مقدار متغیر تصادفی Θ می‌باشد. فرض کنید چگالی Θ، $g(.)=g_\theta(.)$، معلوم بوده و دارای هیچ پارامتر مجهولی نیست و فرض کنید می‌خواهیم $\tau(\theta)$ را برآورد کنیم. چگونه باید اطلاعات اضافی معلوم بودن $g_\theta(.)$ را در روش‌های برآورد وارد کنیم؟ تابع درستنمایی را به عنوان عبارتی که همه اطلاعات را در بر دارد، می‌شناسیم. عبارت تابع درستنمایی، نمونه $x_1,x_2,...,x_n$ و همچنین شکل چگالی $P(X;\theta)$ که از آن نمونه گرفتیم را در برداشت. اکنون عبارتی لازم است که علاوه بر دارا بودن تمام اطلاعات موجود در تابع درستنمایی، اطلاعات اضافی معلوم بودن چگالی $g_\theta(.)$ را نیز دارا باشد. $g_\theta(.)$ توزیع پیشین Θ نامیده می‌شود. این چگالی آنچه را که ما پیش از

---

[1] Prior and Posterior Distribution