
بسم الله الرحمن الرحيم

دانشگاه اصفهان

دانشکده علوم

گروه آمار

پایان‌نامه‌ی کارشناسی‌ارشد رشته‌ی آمار گرایش آمار ریاضی

مباحثی در استفاده از توزیع لاپلاس - چوله چندمتغیره در مدل‌های رگرسیونی

استاد راهنما:

دکتر ایرج کاظمی

پژوهشگر:

راضیه محمدی

مهرماه ۱۳۹۰

کلیه حقوق مادی مترتب بر نتایج مطالعات، ابتکارات و نوآوری های ناشی از تحقیق موضوع این پایان نامه متعلق به دانشگاه اصفهان است.

دانشگاه اصفهان
دانشکده علوم
گروه آمار

پایان‌نامه‌ی کارشناسی ارشد رشته‌ی آمار گرایش آمار ریاضی

خانم راضیه محمدی

تحت عنوان

مباحثی در استفاده از توزیع لاپلاس – چوله چند متغیره در مدل های رگرسیونی

در تاریخ ۹۰/۷/۲۷ توسط هیأت داوران زیر 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 محمد بهرامی با مرتبه‌ی علمی استادیار امضاء

۳- استاد داور خارج از گروه دکتر مجید جعفری خالدی با مرتبه‌ی علمی استادیار امضاء

امضای مدیر گروه

تشکر و قدردانی

شکر و سپاس یزدان پاک را که در لحظه لحظه‌ی زندگی یگانه یار و راهنمای ماست. چه زیباست سخن بزرگ‌مرد تاریخ، سعدی شیرازی که

از دست و زبان که برآید کز عهده‌ی شکرش به در آید

و سپاس فروان از استادم، جناب آقای دکتر ایرج کاظمی که مرا در انجام این پایان‌نامه بسیار یاری رساندند. سپاس برای تمام اوقاتی که به من اختصاص دادند و مرا از راهنمایی‌های ارزشمند خود بهره‌مند کردند. امید آن دارم که کیفیت این مجموعه‌ی گردآوری‌شده، رضایت ایشان را به‌دنبال داشته باشد و همچنین آرزو می‌کنم در تمامی مراحل زندگی، موفق و سربلند باشند.

چکیده

در تحلیل مدل‌های رگرسیونی، فرض معمول این است که مؤلفه‌های خطا دارای توزیع نرمال هستند. اما از آن‌جا که امکان نقض این فرض در تحلیل داده‌های واقعی امکان‌پذیر است، لذا مطالعه بر روی جایگزینی توزیع‌های منعطف‌تر از نرمال موضوع اصلی بسیاری از تحقیقات کاربردی می‌باشد. یکی از این توزیع‌ها لاپلاس- چوله‌ی چندمتغیره است که در دهه‌ی اخیر از سوی محققان زیادی مورد توجه قرار گرفته است. در تحقیقات مختلف، شکل‌های متفاوتی برای نشان دادن تابع چگالی آن ارائه شده است. در این پایان‌نامه، یکی از این شکل‌ها که نسبت به بقیه دارای ساختارساده‌تر و کاربرد بهتر در مدل‌سازی است بررسی می‌شود. این توزیع به دلیل وجود انعطاف‌پذیری بیشترکه شامل سنگینی دم‌ها، کشیدگی زیاد و چولگی است می‌تواند جایگزین مناسبی در مباحث مدل‌سازی رگرسیونی باشد. در این پایان‌نامه ابتدا مدل‌های رگرسیون خطی معمولی با توزیع لاپلاس- چوله معرفی خواهند شد و سپس آن‌ها را به مدل‌های آمیخته با اثرهای تصادفی، که در تحلیل داده‌های همبسته به وفور به‌کار می‌روند، تعمیم می‌دهیم. با توجه به آنکه استنباط پارامترهای مدل بر اساس روش حداکثر درستنمایی حاشیه‌ای منجر به محاسبات جبری پیچیده‌ای می‌شود ما از روش‌های عددی پیشرفته مانند الگوریتم EM از دیدگاه فراوانی‌گرا و نیز رهیافت مونت کارلوی زنجیر مارکوفی از دیدگاه بیزی بهره می‌گیریم. دراین راستا از نمایش تصادفی سلسله‌مراتبی توزیع لاپلاس- چوله‌ی چندمتغیره که آمیخته‌ای از نرمال و گاماست استفاده خواهیم کرد. اهمیت نظری نتایج حاصل را با ارائه‌ی تحلیل داده‌های واقعی نشان می‌دهیم.

واژه‌های کلیدی: استنباط بیزی، الگوریتم EM، نمونه‌گیری‌گیبز، حداکثر درستنمایی، مدل‌رگرسیونی با اثرهای آمیخته، نمایش تصادفی سلسله‌مراتبی.

فهرست مطالب

عنوان **صفحه**

عنوان **صفحه**

عنوان **صفحه**

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

فصل اول

مقدمه

۱-۱ موضوع و پیشینه‌ی تحقیق

در تحلیل مدل‌های رگرسیونی فرض معمول این است که توزیع متغیر پاسخ به‌شرط متغیرهای توضیحی نرمال است. اما از آن‌جا که در بسیاری از مثال‌های تجربی، توزیع داده‌ها از توزیع نرمال انحراف زیادی دارد، لذا استفاده از توزیع‌های جایگزین دیگر، از زمان‌های پیش مورد توجه آماردانان قرار گرفته است. بعضی از این توزیع‌های جایگزین به جهت داشتن ویژگی‌های خاصی چون کشیدگی بیشتر و دم‌های سنگین‌تر و این‌که قابلیت تحت پوشش قرار دادن مشاهدات دورافتاده را دارند، به توزیع‌های مقاوم[1] معروف هستند. یکی از این توزیع‌ها، لاپلاس است که آمیخته‌ی نامتناهی از توزیع‌های نرمال و نمایی است (کاتز و همکاران [2] ۲۰۰۱). این توزیع که علاوه بر داشتن دم‌های سنگین‌تر، نسبت به نرمال از کشیدگی بیشتری نیز برخوردار است، در استنباط‌های مقاوم اهمیت فراوانی دارد. کاربردهایی از این توزیع در مباحث امور مالی، علوم اقتصادی و زیستی یافت می‌شود.

نظر به این‌که در تحلیل داده‌های واقعی، مثال‌هایی نیز وجود دارند که توزیع مشاهدات متقارن نیست، لذا مطالعه بر روی تعمیم‌هایی از توزیع لاپلاس به حالت چوله‌ی آن نیز مورد توجه محققان قرار گرفته است. از بین آن‌ها، کاتز و همکاران (۲۰۰۱) توزیع لاپلاس- چوله‌ی تعمیم‌یافته را به‌صورت زیر تعریف می‌کنند

---

[1] Robust
[2] Kotz

تعریف: فرض کنید متغیر تصادفی $\boldsymbol{Z}$ دارای توزیع نرمال p متغیره با میانگین صفر و واریانس $\boldsymbol{I}_p$ و متغیر تصادفی نامنفی V که مستقل از $\boldsymbol{Z}$ است، دارای توزیع گاما با پارامترهای θ و ۱ باشد. در این صورت متغیر تصادفی $\boldsymbol{Y}$ با نمایش تصادفی

$$\boldsymbol{Y} = \boldsymbol{\mu} + V\boldsymbol{\gamma} + \sqrt{V}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$$

که در آن $\boldsymbol{\mu} \in \mathbb{R}^p$ ، $\boldsymbol{\gamma} \in \mathbb{R}^p$، $\boldsymbol{\Sigma}$ ماتریس معین مثبت و $\boldsymbol{\Sigma}^{1/2}$ جذر $\boldsymbol{\Sigma}$ است، دارای توزیع لاپلاس- چوله‌ی تعمیم‌یافته است. در تعریف فوق، به متغیر تصادفی V که موجب تصادفی کردن میانگین و واریانس توزیع نرمال گردیده است، متغیر آمیختگی[1] گفته می‌شود. همچنین به پارامتر $\boldsymbol{\gamma}$ که نقش چوله کردن توزیع لاپلاس را بر عهده دارد، پارامتر چولگی[2] می‌گویند (اگر $\boldsymbol{\gamma} = \boldsymbol{0}$ آن‌گاه توزیع لاپلاس تعمیم‌یافته حاصل می‌شود). کاتز و همکاران (۲۰۰۱) نشان دادند که تابع چگالی متناظر با توزیع لاپلاس- چوله‌ی تعمیم‌یافته دارای یک نمایش پیچیده بر حسب تابع بسل تعدیل‌شده‌ی نوع سوم[3] می‌باشد. در حالت خاص، اگر $\theta = 1$ آن‌گاه توزیع لاپلاس- چوله‌ی چندمتغیره به دست می‌آید که ویژگی‌های آن از سوی محققان زیادی مطالعه شده است. از جمله کاتز و همکاران (۲۰۰۱) ضمن بررسی ویژگی‌های این توزیع، به محاسبه‌ی برآورد حداکثر درستنمایی پارامترهای آن در حالت یک‌متغیره پرداخته‌اند. ویسک[4] (۲۰۰۹) موضوع برآوردیابی پارامترهای توزیع فوق را با استفاده از روش گشتاوری مورد بررسی قرار داده است. کاربردهایی از توزیع لاپلاس- چوله‌ی چندمتغیره در امور مالی و علوم زیستی توسط کزوبسکی و پدگرسکی[5] (۲۰۰۱)، لیندسی و لیندسی[6] (۲۰۰۶) مطالعه شده است.

ارسلان[7] (۲۰۰۹) نشان داده است که اگر در تعریف فوق $\theta = \frac{p+1}{2}$ باشد (p بعد متغیر تصادفی است) آن‌گاه تابع چگالی لاپلاس- چوله‌ی تعمیم‌یافته شکل ساده‌تری پیدا می‌کند که می‌تواند در مباحث مدل‌سازی مفید واقع شود. وی همچنین با استفاده از الگوریتم EM، برآورد حداکثر درستنمایی پارامترها را به‌دست آورده و الگوریتم شبیه‌سازی ساده‌ای برای این توزیع ارائه داده است.

توزیع لاپلاس- چوله به لحاظ انعطاف‌پذیری زیادی که نسبت به نرمال دارد، می‌تواند در مدل‌های رگرسیونی به‌عنوان جایگزینی مقاوم مورد استفاده قرار گیرد. با توجه به پیچیدگی محاسبات در ارتباط با مسئله‌ی برآوردیابی

---

[1] Mixing variable
[2] Skewness parameter
[3] Modified Bessel function of the third kind
[4] Visk
[5] Kozubowski and Podgorski
[6] Lindsey and Lindsey
[7] Arslan

پارامترها در مدل‌های فوق، روش‌های جایگزین همچون نمونه‌گیری گیبز که در تحلیل‌های بیزی مورد استفاده قرار می‌گیرند، در این پایان‌نامه در نظر گرفته می‌شود. در این راستا با توجه به از نمایش سلسله‌مراتبی[1] مدل‌های فوق، از روش داده‌افزایی[2] که توسط تنر و ونگ[3] (۱۹۸۷) معرفی گردید و موجب سهولت در انجام محاسبات آماری می‌شود، استفاده می‌کنیم.

یکی از مدل‌های معروف آماری که برای تحلیل داده‌های وابسته و آشیانه‌ای به کار می‌رود، مدل چندسطحی[4] است (فریز[5] ۲۰۰٤). از آن‌جا که این مدل‌ها به دلیل وجود اثرهای تصادفی دارای ساختار پیچیده‌تری هستند، لذا استنباط آن‌ها بر مبنای توزیع‌های مقاوم، توسط روش‌های معمول کلاسیک به آسانی امکان‌پذیر نمی‌باشد. از این‌رو استفاده از روش‌های بیزی در مطالعه‌ی مدل‌های چندسطحی، به موضوع اصلی بسیاری از تحقیقات کاربردی تبدیل شده است. از جمله آرلانو- واله[6] و همکاران (۲۰۰۷) به مطالعه‌ی مدل رگرسیون خطی با اثرهای آمیخته[7] با استفاده از توزیع نرمال- چوله پرداختند. در این پایان‌نامه نیز مدل رگرسیون خطی با اثرهای آمیخته با توزیع لاپلاس- چوله‌ی چندمتغیره معرفی می‌شود.

۱-۲ اهداف تحقیق

هدف از انجام این پایان‌نامه، ضمن معرفی توزیع لاپلاس- چوله و مروری بر شکل‌های متفاوتی از این توزیع که در دهه‌ی اخیر معرفی شده است، بررسی کاربرد این توزیع در مدل‌های رگرسیونی می‌باشد. همچنین با توجه به کاربرد گسترده‌ی مدل‌های چندسطحی در تحلیل داده‌ها با ساختار سلسله‌مراتبی، در این پایان‌نامه به‌جای توزیع نرمال که به‌طور متداول به‌کار می‌رود، توزیع لاپلاس- چوله‌ی چندمتغیره را منظور کرده و استنباط پارامترهای این مدل‌ها را با استفاده از دو دیدگاه کلاسیک و بیزی انجام می‌دهیم.

۱-۳ اهمیت و کاربرد نتیجه‌های تحقیق

با توجه به اهمیت مطالعه‌ی توزیع‌های مقاوم در مباحث مدل‌سازی، نتایج حاصل از این پایان‌نامه که شامل معرفی مدل‌های رگرسیونی با توزیع لاپلاس- چوله‌ی چندمتغیره می‌باشد، می‌تواند در زمینه‌ی تحلیل داده‌های تجربی

---

[1] Hierarchical representation
[2] Data augmentation
[3] Tanner and Wong
[4] Multilevel model
[5] Frees
[6] Arellano-Valle
[7] Linear mixed- effect model

مفید واقع شود. از سوی دیگر با توجه به فراگیر شدن کاربرد روش‌های بیزی مبتنی بر روش‌های مونت کارلوی زنجیر مارکوف[1] (MCMC) در دهه‌های اخیر، که موجب سهولت انجام محاسبات آماری گردیده است، در این پایان‌نامه نیز از روش‌های فوق بهره برده‌ایم. همچنین دو مثال کاربردی در این پایان‌نامه ارائه می‌شوند تا اهمیت استفاده از مدل‌های رگرسیونی با توزیع مذکور نشان داده شود.

۱-٤ جنبه‌های محاسباتی

در این پایان‌نامه برای رسم توابع چگالی مختلف، نرم‌افزار Maple به‌کار رفته‌است. همچنین جهت برازش مدل رگرسیونی معمولی از نرم‌افزار SPSS و جهت برازش مدل رگرسیون خطی با اثرهای آمیخته از نرم‌افزار Stata استفاده شده‌است. برآوردیابی حداکثر درستنمایی پارامترهای توزیع لاپلاس- چوله‌ی چندمتغیره توسط نرم‌افزار SAS (فصل چهارم را ببینید) و محاسبات مربوط به تحلیل بیزی مثال‌های ارائه شده در پایان‌نامه توسط نرم‌افزار وین‌باگز[2] انجام شده است.

۱-۵ ساختار پایان‌نامه

در فصل دوم به معرفی آمار بیز و مباحث مرتبط با آن می‌پردازیم. همچنین الگوریتم EM را که یک روش عددی در محاسبه‌ی برآورد حداکثر درستنمایی محسوب می‌شود، در این فصل شرح می‌دهیم. در فصل سوم توزیع لاپلاس- چوله را معرفی می‌کنیم و از این توزیع در برازش مدل‌های رگرسیونی استفاده می‌کنیم. در فصل چهارم به معرفی توزیع لاپلاس- چوله‌ی چندمتغیره پرداخته و موضوع برآوردیابی پارامترهای آن را توسط روش حداکثر درستنمایی بررسی می‌کنیم. در فصل پنجم مدل‌های چندسطحی را معرفی می‌کنیم و از توزیع لاپلاس- چوله‌ی چندمتغیره جهت تحلیل این مدل‌ها استفاده می‌کنیم. در این راستا مدل رگرسیونی عرض از مبدأ تصادفی[3] تصادفی[3] را با فرض آن‌که اثرها و مؤلفه‌های خطا دارای توزیع لاپلاس- چوله‌ی چندمتغیره باشند، معرفی کرده و استنباط بیزی پارامترهای این مدل را با استفاده از رهیافت مونت کارلوی زنجیر مارکوفی انجام می‌دهیم. علاوه بر آن، این موضوع را از دیدگاه فراوانی‌گرا، توسط روش EM، بررسی می‌کنیم. در پایان، مطالعه‌ای بر برازش مدل رگرسیونی خطی با اثرهای آمیخته، توسط توزیع لاپلاس- چوله‌ی چندمتغیره، انجام خواهیم داد.

قابل توجه آن‌که در کل پایان‌نامه، نماد * به‌منظور نشان دادن نتایج مستخرج از پایان‌نامه به‌کار رفته‌است.

---

[1] Markov Chain Monte Carlo
[2] WinBUGS
[3] Random intercept model

فصل دوم

کلیات

۲-۱ مقدمه

در این فصل ابتدا به معرفی آمار بیز و روش‌های مونت کارلوی زنجیر مارکوف که از معروف‌ترین روش‌های محاسباتی در آمار بیز محسوب می‌شوند، می‌پردازیم. سپس الگوریتم متروپلیس- هستینگس[1] و نمونه‌گیری گیبز[2] را به‌عنوان دو روش مهم از روش‌های مونت کارلوی زنجیر مارکوف معرفی می‌کنیم. همچنین مسئله‌ی همگرایی و روش‌های تشخیص آن را که از مباحث مهم در استفاده از این روش‌ها هستند، بررسی می‌کنیم. در ادامه به موضوع پیش‌بینی در مدل‌های بیزی و همچنین معرفی معیارهایی جهت مقایسه‌ی مدل‌ها می‌پردازیم. در پایان فصل نیز الگوریتم EM را به‌عنوان یک روش برآوردیابی عددی معرفی می‌کنیم که در فصل‌های بعد از آن استفاده خواهیم کرد.

[1] Metropolis-Hastings
[2] Gibbs Sampling

۲-۲ آمار بیز

تفاوت اصلی آمار بیز با آمار کلاسیک این است که در آمار بیز، پارامترهای مجهول نیز به‌عنوان کمیت‌های تصادفی در نظر گرفته می‌شوند و توزیع آماری مناسبی به آن‌ها تخصیص داده می‌شود که به آن توزیع پیشین[1] می‌گویند. این توزیع، بیانگر اعتقاد محقق درمورد پارامتر قبل از جمع‌آوری داده‌هاست. در این حالت استنباط آماری بر مبنای قضیه‌ی بیز انجام می‌گیرد. بر این اساس، اطلاعات موجود در توزیع پیشین و اطلاعات حاصل از جمع‌آوری داده‌ها با هم ترکیب شده و توزیع جدیدی حاصل می‌شود که به آن توزیع پسین[2] می‌گویند. به بیان دیگر، ادغام اطلاعات پیشین و کنونی موجب به‌روز شدن اطلاعات محقق از پارامتر می‌گردد و در نتیجه توزیع پسین حاصل می‌شود که استنباط بر مبنای آن انجام می‌گیرد.

فرض کنید متغیر تصادفی Y دارای تابع چگالی $f(y|\theta)$ باشد که در آن متغیر تصادفی θ دارای توزیع پیشین $\pi(\theta)$ است. اگر $Y_1, \dots, Y_n$ را یک نمونه‌ی تصادفی از $f(y|\theta)$ در نظر بگیریم و تابع درستنمایی را با $L(\theta|y)$ نشان دهیم داریم

$$L(\theta|y) = f(y_1, \dots, y_n) = \prod_{i=1}^{n} f(y_i|\theta). \qquad (۱-۲)$$

در این صورت با توجه به قضیه‌ی بیز، توزیع پسین به‌صورت

$$f(\theta|y) = \frac{L(\theta|y)\pi(\theta)}{f(y)} \qquad (۲-۲)$$

به‌دست می‌آید که می‌توان آن را متناسب با حاصل‌ضرب تابع درستنمایی در توزیع پیشین به‌صورت

$$f(\theta|y) \propto L(\theta|y)\pi(\theta) \qquad (۳-۲)$$

در نظر گرفت. با استفاده از توزیع پسین می‌توان برآورد نقطه‌ای پارامتر θ را محاسبه کرد که متداول‌ترین آن، میانگین این توزیع است.

۲-۲-۱ توزیع پیشین

انتخاب توزیع پیشین به دلیل آن‌که توزیع پسین را تحت تأثیر قرار می‌دهد، اهمیت زیادی دارد. در این قسمت به معرفی انواع توزیع‌های پیشین می‌پردازیم.

---

[1] Prior distribution
[2] Posterior distribution

۱- توزیع پیشین ناآگاهی‌بخش[1]

در اغلب مثال‌های کاربردی که هیچ‌گونه اطلاعاتی از پارامتر در دست نیست، توزیع پیشین را طوری در نظر می‌گیرند که توزیع پسین را تحت تأثیر قرار ندهد. چنین توزیع‌هایی را توزیع پیشین مبهم[2] یا ناآگاهی‌بخش می‌گویند (نتزوفراس[3] ۲۰۰۹).

یک توزیع پیشین ناآگاهی‌بخش می‌تواند به صورت

$$\pi(\theta) \propto 1$$

روی فضای پارامتر در نظر گرفته شود که در این صورت به آن توزیع یکنواخت یا مسطح[4] می‌گویند. در این حالت توزیع پسین با تابع درستنمایی معادل می‌شود و در نتیجه برآورد مد توزیع پسین با برآورد حداکثر درستنمایی در حالت کلاسیک یکسان می‌شود. همچنین اگرتوزیع پیشین طوری در نظر گرفته شود که

$$\int \pi(\theta)\, d\theta = \infty$$

آن‌گاه به آن، توزیع ناسره[5] می‌گویند. توزیع‌های ناسره در صورتی قابل استفاده هستند که منجر به یک توزیع پسین سره شوند.

جزئیات بیشتر در مورد توزیع‌های پیشین ناآگاهی‌بخش در یانگ و برگر[6] (۱۹۹۶) موجود است.

۲- توزیع پیشین آگاهی‌بخش

به توزیع‌های پیشینی که حاوی اطلاعات پیشین هستند و توزیع پسین را تحت تأثیر قرار می‌دهند پیشین آگاهی‌بخش می‌گویند. به‌عنوان یک مثال از این توزیع‌ها می‌توان توزیع پیشین مزدوج[7] را در نظر گرفت. هرگاه توزیع‌های پیشین و پسین هر دو از یک خانواده باشند، توزیع پیشین را مزدوج می‌گویند (نتزوفرس ۲۰۰۹).

۳-۲ روش‌های مونت کارلوی زنجیر مارکوفی

نظر به این‌که محاسبه‌ی توزیع پسین و برآورد بیز همیشه به سادگی امکان‌پذیر نمی‌باشد و در اغلب موارد مستلزم حل انتگرال‌های پیچیده است، لذا در این مواقع باید از روش‌های عددی استفاده شود. یکی از روش‌های

[1] Noninformative
[2] Vague
[3] Ntzoufras
[4] Flat
[5] Improper
[6] Yang and Berger
[7] Conjugate