بسم الله الرحمن الرحيم

دانشگاه اصفهان
دانشکده علوم
گروه آمار

پایان نامه‌ی کارشناسی ارشد رشته‌ی آمار گرایش آمار ریاضی

مدل‌های بیزی با اثرات آمیخته با توزیع‌های دم- سنگین

استاد راهنما:
دکتر ایرج کاظمی

پژوهشگر:
لیلا جباری

بهمن ماه ۱۳۸۶

شیوه نگارش پایان نامه رعایت شده است.
تحصیلات تکمیلی دانشگاه اصفهان

دانشگاه اصفهان

دانشکده علوم

گروه آمار

پایان‌نامه‌ی کارشناسی ارشد رشته‌ی آمار- گرایش ریاضی

خانم لیلا جباری کوپایی

تحت عنوان

مدلهای بیزی با اثرات آمیخته با توزیع های دم — سنگین

در تاریخ۸۶/۱۱/۱۰ توسط هیأت داوران زیر بررسی با نمره ۱۹/۳۳ و با درجه عالی به تصویب نهایی رسید.

۱- استاد راهنمای پایان‌نامه دکتر ایرج کاظمی با مرتبه ی علمی استادیار — امضاء

۲- استاد داور داخل گروه دکترهوشنگ طالبی با مرتبه ی علمی استاد یار — امضاء

۳- استاد داور خارج از گروه دکتر عادل محمدپور با مرتبه ی علمی استادیار — امضاء

امضای مدیر گروه

سپاسگزاری:

بر خود لازم می‌دانم از استاد بزرگوار جناب آقای دکتر کاظمی که در تمامی مراحل تدوین این پایان نامه مرا یاری نمودند تشکر و قدردانی نمایم. همچنین از جناب آقای دکتر طالبی و دکتر محمدپور که زحمت مشاوره این پایان نامه را بر عهده داشته‌اند و با نظرات ارزنده خود زمینه پربار شدن آن را فراهم نمودند قدردانی می‌نمایم.

در پایان از مادر و برادر عزیزم که در تمامی مراحل زندگی مشوق و راهنمای من بودند و همچنین از همسر مهربانم که با عشق و علاقه فراوان در تهیه این رساله به من یاری رساندند و از کلیه اساتید و دوستانی که در طول دوران تحصیل از نظرات ارزنده و مفیدشان بهره برده‌ام، از صمیم قلب تشکر می‌نمایم.

تقدیم به مادرم

او که با آمدنم ذره ذره وجودش را نثارم کرد

و

تقدیم به همسرم

او که با آمدنش عشق، شادی و آرامش زندگیم را دو صد چندان کرد.

چکیده

مدل‌های خطی با اثرات آمیخته یک روش مهم برای تحلیل داده‌های طولی است که در علوم مختلف از جمله کشاورزی، زیست شناسی، اقتصاد و ژئوفیزیک استفاده می‌شود. فرض معمول در برازش این مدل‌ها آن است که داده‌ها دارای توزیع نرمال هستند. این فرض ممکن است استنباط را در صورت وجود داده‌های دورافتاده آسیب‌پذیر سازد. لذا در سال‌های اخیر تحقیقات زیادی بر جایگزین کردن توزیع‌های متقارن دم- سنگین و حالت‌های چوله آنها به جای توزیع نرمال صورت گرفته است. نشان داده شده است که این توزیع‌ها برای استنباط‌های آماری مناسب‌تر هستند. توزیع استیودنت t از جمله این توزیع‌ها است که بخصوص در تحلیل بیزی داده‌ها مورد توجه زیادی قرار گرفته است. علاوه بر این، پیشرفت در محاسبات بیزی و روش‌های شبیه‌سازی مونت کارلوی زنجیر مارکوفی باعث گسترش برازش مدل‌های مناسب‌تر به داده‌های واقعی شده است. به طور شگفت‌آوری این روش‌ها در کلاس جدیدی از توزیع‌های بیضوی چوله قابل استفاده می‌باشد. این کلاس که برای تحقیقات کاربردی مناسب بوده، شامل توزیع نرمال چوله، t چوله، نرمال و t می‌باشد. هدف از انجام این رساله کاربرد و مقایسه توزیع‌های بیضوی و بیضوی چوله در مدل‌های رگرسیون خطی، بخصوص با اثرات آمیخته و انتخاب بهترین مدل برازش شده است. بدین منظور از الگوریتم متروپلیس- هستینگز و نمونه‌گیر گیبز جهت برآورد پارامترهای مدل استفاده می‌کنیم.

کلید واژه: نمونه‌گیر گیبز، چولگی، مدل با اثرات آمیخته، استنباط بیزی

فهرست مطالب

عنوان صفحه

فصل اول : پیشگفتار

فصل دوم : مفاهیم مقدماتی

عنوان صفحه

فصل سوم : تحلیل بیزی مدل های خطی با اثرات آمیخته با استفاده از توزیع نرمال و t

فصل چهارم: تحلیل بیزی مدل‌های رگرسیون با استفاده از توزیع‌های نرمال چوله و t چوله

عنوان صفحه

فهرست جدول ها

عنوان | صفحه

فهرست شکل‌ها

عنوان **صفحه**

فصل اول

پیشگفتار

۱-۱ پیشینه تحقیق

یک فرض متداول در مدل‌های خطی آن است که داده‌ها دارای توزیع نرمال هستند. این فرض ممکن است استنباط را در صورت وجود داده‌های دور افتاده آسیب پذیر سازد. لذا در سال‌های اخیر بر روی جایگزین کردن توزیع‌های متقارن دم‌سنگین و توزیع‌های نامتقارن دم‌سنگین به جای توزیع نرمال، تحقیقات زیادی صورت گرفته است و در این رابطه کلاس‌های جدیدی از توزیع ها که برای استنباط‌های آماری مناسب‌تر هستند معرفی شده‌اند. توزیع t از جمله توزیع‌هایی است که (خصوصاً در تحلیل بیزی داده‌ها) از اوخر دهه ۱۹۷۰ میلادی مورد توجه زیادی قرار گرفته است. از کسانی که در این زمینه تحقیق کرده‌اند می‌توان به راجرز و توکی[1] (۱۹۷۲)، دمپستر[2] و همکاران (۱۹۸۰) و لانگ[3] و همکاران (۱۹۸۹) در برازش مدل‌های خطی به روش کلاسیک و به استرندن و گیانولا[4] (۱۹۹۹)، پینهرو[5]

---

1 -Rogers and Tukey
2 -Dempster
3 -Lange
4 -Stranden and Gianola
5 - Pinheiro

(۲۰۰۱) و رزا و همکاران[1] (۲۰۰۴ و۲۰۰۳) در مدل‌های خطی با اثرات آمیخته اشاره کرد. در دهه گذشته آماردانان روی کلاس کامل‌تری از توزیع‌ها، که توزیع نرمال و استیودنت-t چند متغیره را نیز شامل می شود و به کلاس توزیع‌های بیضوی چوله معروف است، توجه خاص نشان داده‌اند. این توزیع‌ها برای برازش به داده‌های واقعی که به طور کامل از توزیع نرمال پیروی نمی‌کنند، بسیار مناسب هستند. اولین بار آزالینی[2] توزیع نرمال چوله یک متغیره را در سال ۱۹۸۵ معرفی کرد. سپس آزالینی و دالا واله[3] (۱۹۹۶) و آزالینی و کاپیتانیو[4] (۱۹۹۹) توزیع نرمال چوله چند متغیره و خواص آن را بررسی کردند. در ادامه گوپتا[5] (۲۰۰۰) خواص دیگر توزیع t چوله چند متغیره را بررسی و تعمیمی از آن را ارائه کرد. برانکو و دی[6] (۲۰۰۱)، آرنولد و بیور[7] (۲۰۰۲و۲۰۰۰) خواص بیشتری از این توزیع‌ها را بررسی کردند. همچنین فرناندز و استیل[8] (۱۹۹۹)، ساهو[9] و همکاران (۲۰۰۳) و آرلانو-واله[10] و همکاران (۲۰۰۵) کاربرد این توزیع ها را در مدل‌های خطی ارائه داده و در تحقیقات خود برتری استفاده از توزیع‌های بیضوی چوله را نسبت به توزیع نرمال نشان دادند، خصوصاً زمانی که مشاهدات دور افتاده در مجموعه داده‌ها وجود دارد و یا توزیع‌های مورد بررسی متقارن نیستند.

۱-۲ اهمیت و اهداف تحقیق

مدل‌های خطی با اثرات آمیخته ابزار مهمی برای تحلیل داده‌های طولی هستند که در علوم مختلف از جمله کشاورزی، زیست شناسی، اقتصاد و ژئوفیزیک استفاده می‌شوند. شهرت فزاینده این مدل‌ها به خاطر انعطاف پذیری آنها در مدل سازی تغییرات بین و درون گروه‌ها است که اغلب در داده‌های طولی اتفاق می‌افتد. فرض معمول در برازش این مدل ها، نرمال بودن اثرات تصادفی و باقیمانده‌ها است. با توجه به ویژگی‌های توزیع نرمال، این فرض ممکن است مدل را در برابر مشاهدات دور افتاده آسیب پذیر سازد. وجود این مشاهدات در مدل‌های خطی با اثرات آمیخته در مقایسه با

1 - Rosa
2 - Azzalini
3 - Dalla valle
4 - Capitanio
5 - Gupta
6 - Branco and Dey
7- Arnold and Beaver
8 - Fernandez and Steel
9 - Sahu
10 -Arellano-Valle

مدل‌های با اثرات ثابت مسئله سازتر است زیرا آنها ممکن است بر توزیع های فرضی اثرات تصادفی، باقیمانده‌های درون گروهی و یا بر هر دو آنها تأثیر گذار باشد که تشخیص آن در عمل مشکل است. از این رو به نظر می‌رسد در صورت وجود مشاهدات دور افتاده، انتخاب توزیع‌های دم- سنگین متقارن و یا توزیع‌های دم-سنگین نامتقارن برای اثرات تصادفی و باقیمانده‌ها مناسب باشد. در سال‌های اخیر با دسترسی به نرم افزارهای پیشرفته آماری، امکان برازش مدل‌های پیچیده توسط روش‌های مختلف آماری به خصوص روش‌های الگوریتمی مانند نمونه‌گیری گیبز[1] امکان پذیر شده است. لذا در این رساله قصد داریم مقایسه‌ای بین فرض نرمال، t ، نرمال چوله و t چوله در مدل‌های خطی انجام داده و با استفاده از الگوریتم متروپلیس- هستینگز[2] و نمونه گیری گیبز، برآورد پارامترها را به روش بیز و با کمک نرم افزار وین باگز[3] بدست آوریم.

۱-۳ ساختار پایان نامه

در فصل دوم ابتدا خلاصه‌ای از آمار بیز را ارائه داده، سپس روش شبیه سازی مونت کارلوی زنجیر مارکوفی[4] (MCMC) را بیان می‌کنیم و خلاصه‌ای از نظریه فرآیندهای تصادفی را عنوان خواهیم کرد. در ادامه الگوریتم متروپلیس-هستینگز و الگوریتم نمونه‌گیر گیبز را که به عنوان یک روش مؤثر در برازش مدل‌های بیزی شناخته شده است، ارائه می‌کنیم و همگرایی آن را مورد بررسی قرار می‌دهیم.

در فصل سوم مدل با اثرات آمیخته با فرض نرمال بودن اثرات تصادفی و باقیمانده‌ها از دیدگاه بیز بررسی می شود و توزیع‌های پسین شرطی کامل پارامترها که برای برآورد پارامترها به روش نمونه‌گیر گیبز لازم هستند، محاسبه می‌شوند. سپس خانواده‌ای از توزیع‌های دم- سنگین را که به توزیع‌های نرمال/ مستقل[5] معروف هستند معرفی می‌نماییم. این خانواده شامل توزیع t استیودنت، اسلش[6] و نرمال ناخالص[7] است که دم سنگین‌تر از نرمال هستند و اغلب برای استنباط‌های استوار استفاده می‌شوند. در ادامه مدل خطی با اثرات آمیخته را با استفاده از توزیع‌های نرمال/

[1] - Gibbs sampling
[2] - Metropolis- Hastings algorithm
[3] - Winbugs
[4] - Markov chain Monte Carlo
[5] – Normal/independent distribution
[6] - Slash
[7] - Contaminated normal

مستقل بررسی کرده و توزیع‌های لازم را برای استفاده از روش MCMC ارائه می‌کنیم. در انتهای این فصل با استفاده از داده های واقعی مربوط به دندان پزشکی، عملکرد مدل‌های نرمال و t را با هم مقایسه می‌کنیم.

در فصل چهارم ابتدا به معرفی توزیع های بیضوی و توزیع‌های بیضوی چوله که نرمال چوله و t چوله حالت‌های خاص آن هستند پرداخته، سپس کلاس جدیدی از توزیع‌های چوله که در موارد تجربی کاربرد دارند و توسط ساهو و همکارانش در سال ۲۰۰۳ معرفی شده است را ارائه نموده و ویژگی‌های آن را بررسی می‌کنیم. در انتهای این فصل با استفاده از داده‌های واقعی مربوط به پذیرش دانشجو در دانشگاه انگلستان مدل های نرمال، t ، نرمال چوله ساهو و t چوله ساهو را با هم مقایسه می‌کنیم و نشان می‌دهیم که مدل t چوله بهترین مدل برای داده‌ها است.

فصل دوم

مفاهیم مقدماتی

۲-۱ مقدمه

در این فصل ابتدا خلاصه‌ای از آمار بیز را ارائه داده و چند اصطلاح مربوط به آن را معرفی خواهیم کرد. سپس در بخش ۲-۳ روش شبیه سازی مونت کارلوی زنجیره مارکوف را بیان و خلاصه‌ای از نظریه فرآیندهای تصادفی را که در فصول بعدی نیاز داریم عنوان خواهیم کرد. در بخش ۲-۴ الگوریتم متروپلیس-هستینگز را معرفی و راجع به بررسی مانایی آن بحث خواهیم کرد. در بخش ۲-۵ الگوریتم نمونه‌گیر گیبز را که به عنوان یک روش مؤثر در برازش مدل‌های بیزی شناخته شده است، ارائه داده و بحث همگرایی آن را بیان می‌کنیم. در ادامه توزیع‌هایی را که در این پایان نامه مورد استفاده قرار می‌گیرند معرفی خواهیم کرد.

۲-۲ آمار بیز

امروزه در علوم مختلف، به طور گسترده از روش بیز برای استنباط‌های آماری استفاده می‌شود. تفاوت اساسی آمار کلاسیک با آمار بیز آن است که در روش کلاسیک پارامتر θ که می‌خواهیم آن را برآورد کنیم، کمیتی ثابت و نامعلوم در جامعه است که بر اساس یک نمونه تصادفی از آن جامعه در مورد آن آگاهی بدست می‌آوریم. در حالی که در روش بیز θ کمیت ثابتی نیست بلکه متغیری تصادفی است که در حوزه مقادیرش تغییر کرده و تغییرات آن توسط یک توزیع احتمال پیشین[۱] بیان می‌شود. این توزیع که بر اساس باور آزمایشگر است بایستی قبل از این که داده ها مشاهده شوند تعیین گردد. در استنباط بیزی توزیع پیشین با اطلاعات نمونه‌ای ترکیب و توزیع جدیدی حاصل می‌شود که آن را توزیع پسین[۲] می‌نامیم. این تعدیل توسط فرمول بیز انجام می‌شود که در زیر آن را بیان می کنیم.

فرض کنید $\mathbf{y}$ نشان دهنده مشاهدات نمونه بوده که دارای تابع چگالی $f(\mathbf{y}|\theta)$ است. اگر چگالی پیشین پارامتر θ را $p(\theta)$ در نظر بگیریم آنگاه چگالی پسین θ عبارت است از

$$p(\theta|\mathbf{y})=\frac{p(\theta)f(\mathbf{y}|\theta)}{f(\mathbf{y})}, \qquad (۱-۲)$$

که در آن $f(\mathbf{y})$ چگالی حاشیه‌ای $\mathbf{y}$ است. در روش بیز، کلیه استنباط‌ها در مورد پارامتر نامعلوم θ براساس توزیع پسین انجام می گیرد.

ویژگی اساسی استنباط بیز، تبدیل آگاهی و اطلاعات غیر نمونه‌ای به صورت توزیع‌های احتمال و ترکیب آن با اطلاعات نمونه‌ای است. در صورت عدم وجود اطلاعات غیر نمونه‌ای، بکارگیری پیشین‌های بی اطلاع[۳] یا ناسره[۴] و ترکیب آن با اطلاعات نمونه‌ای می‌تواند نتایجی معادل با استنباط کلاسیک را به دست دهد. درادامه‌ی این بخش اصطلاحات مورد نیاز در این زمینه را بیان می کنیم.

[1] -Prior distribution
[2] -Posterior distribution
[3] -Noninformative prior
[4] -Improper prior

۲-۲-۱ توزیع پیشین

توزیع پیشین توزیع احتمالی است که اطلاعات پارامتر را قبل از حصول اطلاعات نمونه‌ای بیان می‌کند. این توزیع می‌تواند به پارامترهایی بستگی داشته باشد که آنها را ابر پارامترها[1] می‌نامیم. توزیع پیشین عموماً با توجه به وجود یا عدم وجود اطلاعات قبلی در مورد پارامترهای مدل انتخاب می‌شود. توزیع‌های پیشین متداول به صورت زیر می‌باشند.

توزیع پیشینی که اطلاعات یکسانی را درباره مقادیر مختلف θ در فضای پارامتر آن در نظر بگیرد پیشین آگاهی نابخش نامیده می شود. پیشین‌های یکنواخت و ناسره از این نوع هستند. در زیر دو مثال از آنها آمده است.

۱) فرض کنید θ دارای توزیع یکنواخت بر بازه‌ی $(0,1)$ باشد آنگاه θ دارای توزیع پیشین آگاهی نابخش است.

۲) فرض کنید که θ دارای توزیع نرمال با میانگین θ_0 و واریانس σ^2 است. اگر $\sigma^2 \to \infty$ آنگاه انتگرال روی توزیع پیشین بی نهایت می‌شود. یعنی

$$\int_{-\infty}^{+\infty} p(\theta)\, d\theta = \infty,$$

به این توزیع پیشین، پیشین ناسره گوییم.

هنگامی که از ترکیب توزیع پیشین با تابع درستنمایی توزیع پسینی حاصل شود که متعلق به همان خانواده توزیع پیشین باشد به آن توزیع پیشین مزدوج[2] گوییم.

۲-۲-۲ توزیع پسین

از ترکیب توزیع پیشین $p(\theta)$ و تابع درستنمایی $L(\theta|\mathbf{y})$ توزیعی حاصل می شود که به آن توزیع پسین گوییم. براساس قانون بیز توزیع پسین $p(\theta|\mathbf{y})$، متناسب با حاصل ضرب $p(\theta)$ و $L(\theta|\mathbf{y})$ است. یعنی

$$p(\theta|\mathbf{y}) \propto L(\theta|\mathbf{y})\, p(\theta).$$

---

[1] -Hyperparameters

[2] -Conjugate prior

۲-۳ شبیه سازی مونت کارلوی زنجیره مارکوف

پارامتر $\boldsymbol{\theta}$ عموماً یک بردار kبعدی است و جهت انجام استنباط بیزی پارامتر θ_i برای $i=1,...,k$ نیاز به محاسبه توزیع‌های پسین کناری است که اغلب نیازمند انتگرال گیری از تابع‌هایی با ابعاد زیاد بوده و مستلزم محاسبات بسیار پیچیده می‌باشد. یک روش مناسب برای محاسبه این انتگرال ها استفاده از زنجیره مارکوف مونت کارلو است که از توزیع‌های پیچیده مورد علاقه شبیه سازی می کند. یک شکل عمومی از روش MCMC توسط متروپلیس[1] (۱۹۵۳) و هستینگز[2] (۱۹۷۰) ارائه شد که به الگوریتم متروپلیس- هستینگز معروف است و نمونه‌گیری گیبز که گمن و گمن[3] (۱۹۸۴)آن را معرفی کردند یک حالت خاص آن می‌باشد. استفاده فراگیر از نمونه‌گیر گیبز را می توان در اوایل سال ۱۹۹۰ توسط گلفند و اسمیت[4] دانست. قبل از معرفی این روش مروری بر روش انتگرال گیری مونت کارلو و زنجیره‌های مارکوف و اینکه چگونه می تواند در روش‌های بیز استفاده شود خواهیم داشت.

۲-۳-۱ انتگرال گیری مونت کارلو[5]

روش مونت کارلو در اصل یک روش بسط داده شده توسط فیزیکدانان می‌باشد که از تولید اعداد تصادفی برای محاسبه انتگرال ها استفاده می کند. فرض کنید هدف محاسبه انتگرال

$$\int_a^b h(x)\,dx,$$

است. اگر بتوانیم $h(x)$ را به صورت حاصل ضرب یک تابع $p(x)$ و یک تابع چگالی احتمال $f(x)$ روی فاصله (a,b) تعریف کنیم آنگاه

$$\int_a^b h(x)\,dx = \int_a^b p(x)f(x)\,dx = E_{f(x)}\left[p(X)\right]. \qquad (۲-۲)$$

1 -Metropolis
2 -Hastings
3 -Geman and Geman
4 -Gelfand and Smith
5 -Monte Carlo integration