
بسم الله الرحمن الرحیم

دانشگاه اصفهان
دانشکده علوم
گروه آمار

پایان نامه‌ی کارشناسی ارشد رشته‌ی آمار گرایش آمار ریاضی

توزیع‌های آمیخته– مقیاس نرمال– چوله و کاربرد آن‌ها در مدل‌های رگرسیونی غیرخطی

استاد راهنما:

دکتر ایرج کاظمی

پژوهشگر:

مهسا عابدینی

آبان ماه ۱۳۹۱

دانشگاه اصفهان
دانشکده علوم
گروه آمار

پایان‌نامه‌ی کارشناسی ارشد رشته‌ی آمارگرایش آمار ریاضی

خانم مهسا عابدینی

تحت عنوان

کاربرد توزیع های آمیخته – مقیاس نرمال – چوله در مدل ها ی رگرسیونی غیر خطی

در تاریخ۹۱/۸/۲۸ توسط هیأت داوران زیر بررسی با درجه **عالی** به تصویب نهایی رسید.

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

۲- استاد داخل پایان نامه دکتر محمد بهرامی با مرتبه‌ی علمی استادیار امضاء

۳- استاد داور خارج از گروه دکتر سروش علیمرادی .با مرتبه‌ی علمی استادیار امضاء

امضای مدیر گروه

خدای را بسی شاکرم که از روی کرم، پدر و مادری فداکار به من عطا فرمود تا در سایه درخت پربار وجودشان بیاسایم و از ریشه آنها شاخ و برگ گیرم و از سایه‌ی وجودشان در راه کسب علم و دانش قدم بردارم، والدینی که وجودشان تاج افتخاری است بر سرم و نامشان دلیلی است بر بودنم چرا که این دو وجود، پس از پروردگار مایه‌ی هستی‌ام بوده‌اند، دستم را گرفته‌اند و راه رفتن را در این وادی زندگی پر فراز و

نشیب به من آموخته‌اند؛ اساتیدی که برایم زندگی کردن، بودن و انسان بودن را معنا کردند و حال این برگ سبزی است تحفه‌ی درویش که تقدیم می‌کنم به

پدر و مادر عزیزم و خانواده‌ی مهربانم

به نام یگانه خالق هستی بخش جان آفرین که در ذات احدیتش شکی نیست، به عون و قوه‌ی الهی این پروژه را با راهنمایی‌های جناب آقای دکتر کاظمی به پایان رساندم و بنا به فرموده‌ی مولای متقیان علی(ع): «من علّمنی حرفاً فقد یسّرنی عبداً» همواره افتخار شاگردی ایشان را با خود خواهم داشت و کمال تشکر را از ایشان دارم.

چکیده

یکی از مباحث مهم در تحلیل رگرسیون غیرخطی، فرض درباره توزیع خطای مدل است که در چند دهه اخیر موضوع تحقیق بسیاری از محققان بوده است. پیش از این، فرض بر این بوده که خطا از توزیع نرمال پیروی می‌کند. بسیاری از محققان توزیع‌های جدیدی را با خواص منحصر به فردی برای توزیع خطا در نظر گرفته‌اند، از جمله می‌توان به خانواده توزیع آمیخته-مقیاس چوله-نرمال اشاره نمود. مبحث مهم دیگر در تحلیل رگرسیون غیرخطی، تعیین شکل تابع انتظار است که با توجه به ساختار داده‌های واقعی مشخص می‌شود.

هدف از انجام این پایان‌نامه، معرفی خانواده توزیع آمیخته مقیاس چوله-نرمال و توزیع‌های چوله-نرمال، چوله-تی و چوله-اسلش متعلق به آن و بررسی خواص مطلوب آن‌ها در حالت کلی و در رگرسیون غیرخطی است. از خواص مطلوب این خانواده از توزیع‌ها می‌توان به نمایش سلسله مراتبی آن که امکان محاسبه ساده‌تر برآورد پارامترها به روش الگوریتم EM، تعمیم‌های آن و رهیافت نمونه‌گیر گیبز را فراهم می‌کند، اشاره کرد. با برازش دادن مدل رگرسیون غیرخطی مناسب با توزیع خطای مختلف متعلق به خانواده توزیع آمیخته-مقیاس چوله-نرمال بر داده‌های واقعی به این نتیجه رسیدیم که توزیع‌های چوله در مواردی بهتر از توزیع نرمال به داده‌ها برازش می‌شود. با انجام مطالعات شبیه‌سازی از یک مدل رگرسیون غیرخطی با خطای چوله-تی و مقایسه با سایر توزیع‌های معرفی شده برای خطا، قابلیت توزیع آمیخته مقیاس چوله-نرمال در برازش مدل‌های رگرسیون غیرخطی مشخص می‌شود.

کلید واژه: شکل سلسله مراتبی، نمونه‌گیر گیبز، الگوریتم EM، توزیع‌های نامتقارن، توزیع‌های دم-سنگین.

فهرست مطالب

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

فهرست شکل‌ها

مخفف یا کوتاه نوشت

امیدریاضی	E
برآورد خطای استاندارد مجانبی	$ease$
تقریب	$\approx$
تناسب	$\propto$
توزیع	$\sim$
خطای استاندارد مجانبی	ase
ماتریس قطری	$diag$
مشاهده	obs
میانگین توان دوم خطا	MSE
واریانس	Var
هم‌توزیع	$\underline{\underline{d}}$

فصل اول

مقدمه

۱-۱ موضوع و پیشینه تحقیق

در بسیاری از آزمایش‌های بالینی در علوم پزشکی و یا تحلیل داده‌های وابسته و پانلی که عموماً از مطالعه‌های اقتصادی گردآوری شده‌اند مشاهده می‌شود که مدل رگرسیونی مناسب برای داده‌ها مدلی غیرخطی است. بنابراین در این موارد اهمیت مدل رگرسیونی غیرخطی آشکار می‌شود. گاهی نیز در تحلیل رگرسیونی غیرخطی، محقق با این مسئله مواجه می‌شود که ساختار داده‌ها متقارن نیست و یا دارای مشاهده‌های دورافتاده است. بنابراین برای این مدل‌های غیرخطی در نظر گرفتن توزیع خطای نرمال مناسب نیست و توزیع‌های متعلق به خانواده توزیع $SMSN$ انتخابی می‌تواند مناسب باشد. با توجه به خواص مطلوب این خانواده از توزیع‌ها، پژوهش‌های جدید بسیاری بر مبنای خواص این خانواده بنا شده‌اند. در این پایان‌نامه خانواده توزیع آمیخته-مقیاس چوله-نرمال[1] ($SMSN$) که تعمیمی از توزیع چوله-نرمال[2] (SN) (آزالینی[3]، ۱۹۸۵) است و دو روش برآوردیابی

[1] Scale mixture of skew normal
[2] Skew normal
[3] Azzalini

تکراری-عددی تحت عنوان‌های الگوریتم ماکسیمم سازی امید[۱] (EM) (دمپستر و همکاران[۲]، ۱۹۷۷) و رهیافت مونت کارلوی زنجیر مارکفی[۳] ($McMC$) برای پارامترهای آن معرفی می‌شود. همچنین پس از یادآوری مختصر مدل رگرسیون خطی، به بررسی مدل رگرسیون غیرخطی پرداخته می‌شود. در نهایت نیز با در نظر گرفتن توزیع خطای $SMSN$ در مدل رگرسیون غیرخطی بخشی از کاربرد این خانواده توزیع نشان داده می‌شود.

در سال ۲۰۰۱ برانکو و دی[۴] استفاده از خانواده توزیع $SMSN$ را برای تحلیل مشاهده‌های نامتقارن[۵] پیشنهاد دادند. حالت خاصی از توزیع‌های فوق، برای مدل‌های غیرخطی توسط لین و همکاران[۶] (۲۰۰۹) ارائه شد. لاچس و همکاران[۷] (۲۰۱۰) با شبیه‌سازی داده‌های چوله از توزیع‌های دم-سنگین[۸]، کاربرد این توزیع‌ها را نشان داده‌اند. باسو و همکاران[۹] (۲۰۱۰) نیز توزیع‌های $SMSN$ را در مدل‌های آمیخته پیشنهاد دادند. مدل پیشنهادی آن‌ها به طور همزمان عدم تقارن توزیع داده‌ها و تأثیر مشاهده‌های دورافتاده[۱۰] را در نظر می‌گیرد. گری و همکاران[۱۱] (۲۰۱۱) برازش مدل‌های رگرسیونی غیرخطی را توسط توزیع‌های $SMSN$ و با به کارگیری الگوریتم EM انجام دادند. آنان همچنین با مطالعه‌های شبیه‌سازی این مدل‌ها برای داده‌هایی با ساختار پخش نامتقارن نتیجه گرفتند که استنباط آماری پارامترها توسط این توزیع‌های آمیخته-مقیاس در برابر مشاهدات دورافتاده نسبت به چوله-نرمال استوارتر است. همچنین استفاده از روش‌های بیزی مانند $McMC$ را در برازش مدل‌های فوق پیشنهاد دادند. کانچو و همکاران[۱۲] (۲۰۱۱) با استفاده از رهیافت $McMC$، کاربردی از خانواده توزیع $SMSN$ را در مدل‌های غیر خطی ارائه دادند. این مدل‌ها با لحاظ کردن چولگی و دم-سنگین بودن در تحلیل داده‌ها مدل‌های بسیاری، از جمله چوله-تی[۱۳] (ST) و چوله-اسلش[۱۴] (SSL)، را پوشش می‌دهند. آن‌ها با استفاده از روش‌های بیزی بر اساس معیار کولبک-لیبر به تشخیص داده‌های پرنفوذ پرداختند. با تحلیل داده‌های واقعی نشان دادند که مدل ST متعلق به خانواده توزیع $SMSN$ برازنده‌تر از مدل SN است.

---

[1] Expectation maximization algorithm
[2] Dempster et al.
[3] Markov chain
[4] Branco and Dey
[5] Asymmetric
[6] Lin et al.
[7] Lachos et al.
[8] Heavy-tailed
[9] Basso et al.
[10] Outlier
[11] Garay et al.
[12] Cancho et al.
[13] Skew t
[14] Skew slash

۱-۲ اهمیت و اهداف تحقیق

در تحلیل داده‌هایی که مدل‌های رگرسیون با فرض نرمال بودن مؤلفه‌های خطا به آن‌ها خوب برازش نمی‌شوند اهمیت استفاده از توزیع‌های دیگر، از جمله توزیع‌های $SMSN$، روشن است. اهمیت ویژه خانواده توزیع‌های معرفی شده در کاربرد آن‌ها جهت مدل‌سازی داده‌هایی با ساختار نامتقارن و دارای مشاهده‌های دورافتاده یا پرنفوذ است. نمایش سلسله مراتبی مطلوب توزیع‌های $SMSN$ امکان برآوردیابی توسط روش‌های EM و $McMC$ برای پارامترهای این توزیع‌ها را فراهم می‌کند.

هدف از انجام این پایان‌نامه، بررسی توزیع‌های $SMSN$ و به دست آوردن خواص احتمالی و استنباطی آن‌ها در مدل‌سازی آماری است. با توجه به خواص مطلوب این توزیع‌ها توجه بیشتری به کاربرد آنها در برازش مدل‌های خطی و غیر خطی جهت بر طرف کردن مشکلات مطرح شده خواهد شد. در این رابطه مدل‌های فوق بر مجموعه داده‌های واقعی برازش داده می‌شوند و با استفاده از معیارهای انتخاب مدل مناسب، درستنمایی توزیع‌های متعلق به خانواده توزیع $SMSN$ جهت حصول نتایج معتبر مقایسه می‌شوند. در این زمینه جهت برآورد پارامترها روش‌های تحلیل بیزی، به ویژه رهیافت نمونه‌گیرگیبز[1]، استفاده خواهد شد. به علاوه برای شناخت بهتر خواص این توزیع‌ها در مدل‌های غیرخطی، شبیه‌سازی از آن‌ها صورت می‌گیرد و نتایج به دست آمده تحلیل می‌شوند.

۱-۳ روش‌های محاسباتی

برازش مدل‌های رگرسیونی و تحلیل آن‌ها با استفاده از نرم‌افزارهای نسخه sas 9.2 و اُپن‌باگز[2] نسخه 3.2.1 انجام شده است. با استفاده از نرم‌افزار $maple$ نمودار توزیع‌ها رسم شده‌اند. برای برآورد بیزی پارامترهای مدل از نرم‌افزار اُپن‌باگز و برای سایر تحلیل‌ها از دستورهای$proc\ nlin$ و یا $proc\ nlmixed$ از نرم‌افزار sas استفاده شده است.

۱-۴ ساختار پایان‌نامه

ساختار این پایان‌نامه به این صورت است که در فصل دوم مباحثی که در فصل‌های بعد مورد استفاده قرار می‌گیرند مطرح می‌شوند. در فصل سوم خانواده توزیع $SMSN$ و توزیع‌های SSL و ST متعلق به آن معرفی و

[1] Gibbs sampler
[2] Open Bugs

پارامترهای آن‌ها از طریق روش‌های گیبز مبتنی بر $McMC$ و تعمیم‌های الگوریتم EM برآورد می‌شوند. فصل چهارم در بردارنده مطالبی در خصوص مدل رگرسیون غیرخطی است. در نهایت در فصل پنجم کاربرد توزیع‌های متعلق به خانواده توزیع $SMSN$ در مدل غیرخطی بررسی می‌شود. در انتهای این فصل نیز داده‌های واقعی را تحلیل خواهیم کرد.

فصل دوم

مفاهیم و تعاریف مقدماتی

۲-۱ مقدمه

در این فصل تعاریف و مفاهیم به‌کارگرفته شده در این پایان‌نامه را شرح می‌دهیم. خانواده توزیع $SMSN$ (برانکو و دی، ۲۰۰۱) که در این پایان‌نامه مورد بررسی قرار می‌گیرد آمیخته-مقیاسی از چوله-نرمال است که برای معرفی این خانواده توزیع ابتدا آشنایی مختصری با خانواده توزیع‌های چوله-متقارن (ونگ و همکاران[1]، ۲۰۰۴) و آمیخته-مقیاس نرمال (SMN) (آندره و مالوز[2]، ۱۹۷۴) ارائه می‌شود. از طرف دیگر، برای محاسبه برآورد پارامترها نیاز به معرفی روش‌های تکراری-عددی مانند الگوریتم EM (دمپستر و همکاران، ۱۹۷۷) مبتنی بر رهیافت فراوانی‌گرا و روش‌های $McMC$ از جمله متروپلیس-هستینگز[3] (هستینگز، ۱۹۷۰) و گیبز (گمن و گمن[4]، ۱۹۸۰) بر اساس رهیافت بیز است.

[1] Wang et al.
[2] Andrew and Mallows
[3] Metropolis-Hastings
[4] Geman and Geman