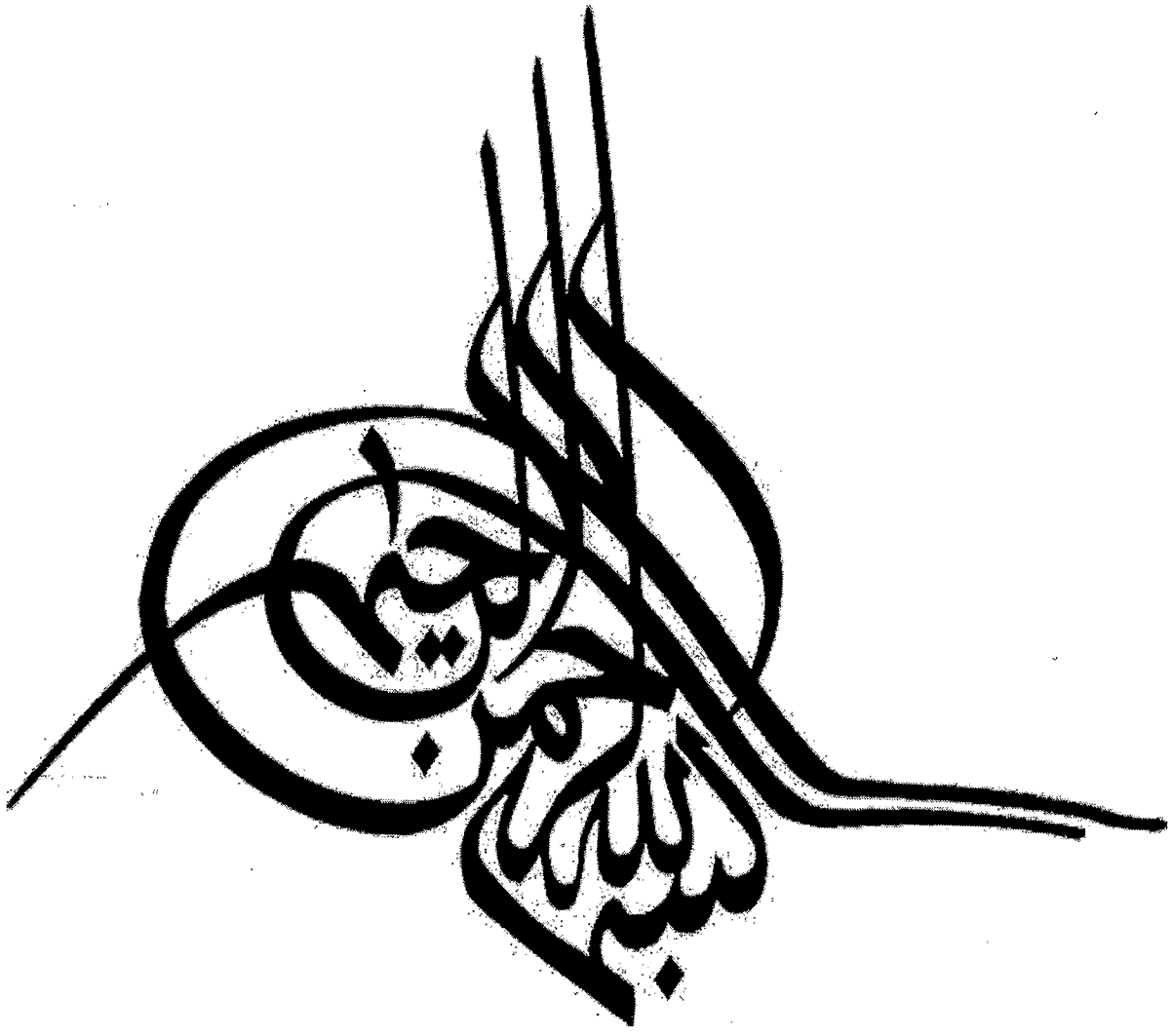

٢٠١٢٥١٦ — ١٣٤٧٠٠

دانشگاه اصفهان

دانشکده علوم

گروه آمار

پایان‌نامه‌ی کارشناسی ارشد رشته‌ی آمار گرایش آمار ریاضی

مطالعه‌ای بر مدل های رگرسیونی با اثرات آمیخته با توزیع‌های غیرنرمال

مرکز اطلاعات و مدارک علمی ایران
تاریخ ثبت مدارک
۱۳۸۹/۲/۶-

استاد راهنما:

دکتر ایرج کاظمی

پژوهشگر:

زهرا مهدیه

بهمن‌ماه ۱۳۸۸

۱۳۴۷۰۰

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

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

دانشگاه اصفهان

دانشکده علوم

گروه آمار

پایان‌نامه‌ی کارشناسی ارشد رشته‌ی آمارگرایش آمار ریاضی

خانم زهرا مهدیه

تحت عنوان

برازش مدل های رگرسیونی با اثرات آمیخته با توزیع های غیر نرمال

در تاریخ۲۱ /۸۸/۱۱ توسط هیأت داوران زیر بررسی با درجه **عالی** به تصویب نهایی رسید.

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

۲- استاد داور داخل گروه پایان‌نامه دکتر مجید اسدی با مرتبه‌ی علمی استاد — امضاء

۳- استاد داورخارج از گروه دکتر مجتبی گنجعلی با مرتبه‌ی علمی دانشیار — امضاء

امضای مدیر گروه

سپاسگزاری

ستایش خداوند لطیف را که بی لطف او هیچ نمی‌توان کرد و سپاس او را که همواره و در همه حال با ماست.

حمد و ثنا می‌گویم او را در خور زبان قاصر خویش نه در شأن و مقام پروردگاریش.

در طول دوران تحصیلم افتخار شاگردی محضر اساتید بزرگواری نصیبم شد که جبران قطره‌ای از زحمات و محبت‌هایشان برایم ممکن نیست. لذا در این مجال از کلیه این عزیزان تشکر و قدردانی نموده و خالصانه‌ترین سپاس‌ها را به محضر استاد راهنمای عزیزم دکتر کاظمی که با صبوری کاستی‌های این شاگرد را تحمل نموده و بی‌دریغ از دریای ژرف علم و ادبشان به من آموختند، تقدیم می‌دارم و امیدوارم سلامتی و موفقیت چاشنی همیشگی زندگیشان باشد.

در پایان نهایت سپاس و تشکر را به حضور پدر و مادر مهربانم که با محبت‌های بی‌دریغشان همواره یار و یاورم بوده‌اند تقدیم می‌دارم و از زحمات کلیه عزیزانی که در دانشکده علوم نهایت کمک و همکاری را با من نمودند و نیز همه دوستانی که در طول دوران تحصیل لحظات تکرار نشدنی را در کنارشان گذراندم، قدردانی نموده و روزهای خوشی را توأم با کامیابی و مسرت برایشان آرزومندم.

تقدیم به حضرت دوست

که هرچه داریم از اوست

تقدیم به پدر و مادر عزیزم

که عاشقانه سوختند تا روشن‌گر راهم باشند و گرمابخش وجودم

تقدیم به استاد بزرگوارم دکتر کاظمی

به پاس آنچه به من آموختند.

چکیده

در بسیاری از تحقیقات کاربردی بخصوص در علوم پزشکی، کشاورزی و مطالعات اقتصادی، مدل‌های رگرسیونی با اثرات آمیخته برای تحلیل داده‌های طولی و مشاهدات وابسته به کار می‌روند. در این مدل‌ها وجود اثرهای تصادفی باعث تبیین تغییرپذیری درمشاهدات تکراری وتعدیل ساختارهمبستگی بین اندازه‌های مکرر مربوط به واحدها و یا مشاهدات وابسته درون طبقات شده که این موجب افزایش دقت وکارایی بیشتر برآورد پارامترها می‌شود. فرض‌های معمول در برازش این مدل‌ها نرمال بودن توزیع مؤلفه‌های خطا و اثرهای تصادفی است. اما در برازش مدل به داده‌های واقعی ممکن است فرآیند تولید داده‌ها به گونه‌ای باشد که به کار بردن توزیع نرمال برای مؤلفه‌های خطا و اثرات تصادفی منجر به نتایج نادرست و غیر قابل اعتبارگردد.

استفاده از روش ماکسیمم‌درستنمایی جهت برآوردیابی پارامترهای مجهول منجر به حل انتگرال‌های پیچیده در تابع درستنمایی حاشیه‌ای بخصوص در حالت غیرنرمال بودن توزیع اثرات تصادفی شده که نیاز به روش‌های تکراری عددی پیشرفته دارد. در مطالعات اخیر روش مربع‌بندی گوسی به‌عنوان یک روش مؤثر جهت تقریب عددی انتگرال‌های پیچیده برای یافتن برآورد ماکسیمم‌درستنمایی پارامترها در مدل‌های آمیخته با اثرات تصادفی نرمال ارائه شده است. اما در حالت غیرنرمال بودن توزیع اثرات تصادفی نیاز به روش مناسبی است که تقریب خوبی از انتگرال‌ها را ارائه دهد.

این پایان نامه با ارائه روش‌های جدید برآوردیابی در مدل‌های خطی و تعمیم یافته خطی با اثرات تصادفی غیرنرمال، به مطالعه این موضوع می‌پردازد. همچنین کاربرد روش‌های معرفی شده را با استفاده از داده‌های واقعی مورد بررسی قرار می‌دهد.

این پایان‌نامه به دلیل ویژگی‌های منحصربه فرد توابع مفصل ازجمله قابلیت تولید ساختارهای مختلف همبستگی غیرخطی و ایجاد توزیع‌های حاشیه‌ای متفاوت، این توابع را به عنوان روشی معقول برای ساخت توزیع‌های چندمتغیره براساس توزیع‌های حاشیه‌ای و ساختار همبستگی بین متغیرها معرفی می‌نماید. در این راستا از این توابع به عنوان ابزاری برای ایجاد ارتباط بین توزیع حاشیه‌ای اثرات تصادفی در مدل‌های شامل دو و یا بیش از دو اثر تصادفی و همچنین برای برقراری ارتباط بین متغیرهای پاسخ در مدل‌های چندمتغیره با اثرات تصادفی استفاده شده و در آخر کاربرد مدل‌های مذکور با استفاده از داده‌های شبیه سازی و تجربی نشان داده شده است.

کلیدواژه‌ها: توزیع آمیخته، مدل‌های تعمیم یافته با اثرات آمیخته ، مدل‌های چندمتغیره آمیخته، مربع‌بندی گوسی، مفصل‌های ارشمیدسی، مفصل‌های بیضوی.

فهرست مطالب

عنوان **صفحه**

فصل اول: مقدمه

۱-۱- موضوع و پیشینه تحقیق ۱

۱-۲- اهداف تحقیق ۶

۱-۳- اهمیت و کاربرد تحقیق ۷

۱-۴- محاسبات آماری ۸

۱-۵- ساختار پایان نامه ۹

فصل دوم: مدل‌سازی مشاهدات با اندازه‌های مکرر

۲-۱- ساختار اندازه‌های مکرر ۱۱

۲-۲- تحلیل اندازه‌های مکرر ۱۳

۲-۳- تعمیم مدل‌های خطی برای اندازه‌های مکرر ۱۷

۲-۴- کاربرد مدل‌های خطی با اثرهای آمیخته در تحلیل اندازه‌های مکرر ۲۱

فصل سوم: برازش مدل‌های رگرسیونی با اثرهای آمیخته

۳-۱- مشخصات مدل‌های خطی با اثرهای آمیخته ۲۸

۳-۲- مثال‌هایی از مدل‌های خطی با اثرهای آمیخته ۲۹

۳-۲-۱- مدل تحلیل واریانس یک راهه با اثرهای تصادفی ۲۹

۳-۲-۲- مدل با ضرایب تصادفی ۳۲

۳-۳- تحلیل مدل با فرض نرمال بودن توزیع اثرهای تصادفی و باقیمانده‌ها ۳۳

۳-۴- تحلیل مدل با اثرهای تصادفی و باقیمانده‌های غیرنرمال ۳۵

۳-۵- روش مربع بندی گوسی ۳۷

۳-۶- روش تبدیل احتمال انتگرال ۳۹

۳-۶-۱- روش تبدیل احتمال انتگرال برای یک اثرتصادفی ۳۹

۳-۶-۲- روش تبدیل احتمال انتگرال برای چند اثرتصادفی مستقل ۴۱

۳-۷- روشی جایگزین برای فرمول‌بندی تابع درستنمایی با اثرهای تصادفی غیرنرمال ۴۲

۳-۸- تحلیل داده‌های بقا ۴۳

عنوان | صفحه

۳-۸-۱- مفاهیم مقدماتی ۴۳
۳-۸-۲- مدل رگرسیونی خطرات متناسب ۴۵
۳-۸-۳- برازش مدل رگرسیونی خطرات متناسب ۴۶
۳-۸-۴- کاربرد روش LR در برازش مدل رگرسیونی خطرات متناسب ۴۸
۳-۸-۵- مثال کاربردی: تأثیراشعه لیزر در به تأخیر انداختن شروع کم‌بینی ناشی از دیابت ۴۹
۳-۹- توزیع‌های آمیخته ۵۱
۳-۹-۱- توزیع آمیخته متناهی ۵۱
۳-۹-۲- توزیع آمیخته متناهی با مؤلفه‌های نرمال ۵۲
۳-۹-۳- کاربرد توزیع‌های آمیخته در مدل‌های رگرسیونی ۵۳
۳-۹-۴- کاربرد روش LR در برازش مدل آمیخته ۵۴
۳-۹-۵- مطالعه شبیه سازی ۵۵
۳-۱۰- نتیجه‌گیری ۵۶

فصل چهارم: مدل‌های خطی و تعمیم‌یافته خطی چندمتغیره با اثرهای آمیخته

۴-۱- تحلیل توأم پاسخ‌های وابسته ۵۸
۴-۲- مدل‌های چندمتغیره خطی با اثرهای آمیخته ۶۲
۴-۳- کاربرد مدل‌های خطی تعمیم‌یافته آمیخته در تحلیل همزمان پاسخ‌های غیرنرمال ۶۴
۴-۳-۱- تعمیم مدل‌های خطی آمیخته برای تحلیل پاسخ‌های غیرنرمال ۶۴
۴-۳-۲- کاربرد مدل‌های خطی تعمیم یافته آمیخته درمدل‌سازی پاسخ‌های وابسته ۶۷
۴-۴- مثال کاربردی: فعالیت مایوالکتریک اسب‌های غیرمسابقه ۷۱
۴-۵- نتیجه‌گیری ۷۴

فصل پنجم: توابع مفصل و کاربرد آن‌ها در مدل‌های خطی آمیخته

۵-۱- معرفی مفصل ۷۶
۵-۱-۱- تعریف مفصل دوبعدی ۷۶
۵-۱-۲- تابع چگالی مفصل ۷۷
۵-۱-۳- تعمیم مفصل به حالت چندمتغیره ۷۸

عنوان | صفحه

۵-۱-۴- کران‌های فرشت ۷۹
۵-۲- مفصل و چند اندازه ارتباط ۸۰
۵-۲-۱- τ کندال ۸۲
۵-۲-۲- ρ اسپیرمن ۸۲
۵-۲-۳- ویژگی‌های همبستگی τ کندال و ρ اسپیرمن ۸۲
۵-۳- معرفی انواع مفصل ۸۳
۵-۳-۱- مفصل‌های بیضوی ۸۶
۵-۳-۲- مفصل‌های ارشمیدسی ۸۸
۵-۴- برآوردیابی مفصل ۹۱
۵-۵- ارزیابی و انتخاب مفصل ۹۲
۵-۶- کاربرد توابع مفصل در مدل‌های رگرسیونی با اثرهای آمیخته ۹۵
۵-۶-۱- کاربرد روش LR در برازش مدل تعمیم یافته خطی با شیب و عرض از مبدأتصادفی ۹۵
۵-۶-۲- مطالعه شبیه سازی ۱۰۱
۵-۶-۳- مثال کاربردی: میزان برداشت گندم ۱۰۵
۵-۷- کاربرد توابع مفصل در مدل‌سازی پاسخ‌های چندمتغیره ۱۰۷
۵-۷-۱- مشخصات مدل ۱۰۸
۵-۷-۲- برآوردیابی پارامترها ۱۰۸
۵-۷-۳- مثال کاربردی: جهانی شدن و قدرت حزب کارگر ۱۰۹
۵-۸- نتیجه گیری ۱۱۵

پیوست ۱۱۶
منابع و مآخذ ۱۲۱

فهرست شکل‌ها

عنوان **صفحه**

شکل۳-۱- توزیع آمیخته نرمال ۵۳

شکل۵-۱- نمودارهای کانتور بعضی مفصل‌های ارشمیدسی و بیضوی با حاشیه‌های نرمال ۸۵

شکل۵-۲- نمودارهای پراکندگی بعضی مفصل‌های ارشمیدسی و بیضوی با حاشیه‌های نرمال ۸۵

شکل۵-۳- کانتورهای توزیع نرمال و لجستیک استاندارد ۹۷

شکل۵-۴- نمودار پراکندگی و QQ اثرات تصادفی در مطالعه شبیه‌سازی ۱۰۲

شکل۵-۵- نمودار پراکندگی اثرهای تصادفی در مثال میزان برداشت گندم ۱۰۶

شکل۵-۶- نمودار پراکندگی دو متغیر *civilian* و *transfer* ۱۱۱

فهرست جدول‌ها

عنوان **صفحه**

جدول۲-۱- ساختار کلی اندازه‌های مکرر ۱۲
جدول۲-۲- ساختار یک حالت خاص از نمونه‌های چندگانه ۱۳
جدول۲-۳- ساختارهای مختلف ماتریس کوواریانس ۲۵
جدول۳-۱- مقادیر w_k و z_k در روش مربع بندی گوسی ۳۸
جدول۳-۲- برازش مدل خطرات متناسب با شکستگی گاما و نرمال ۵۰
جدول۳-۳- برآورد پارامترها: مدل آمیخته ۵۶
جدول۴-۱- برآورد پارامترها: مدل چندمتغیره آمیخته ۷۴
جدول۵-۱- معرفی چند مفصل ارشمیدسی و بیضوی دومتغیره ۸۴
جدول۵-۲- برآورد پارامترها: مفصل کلایتون ۱۰۲
جدول۵-۳- برآورد پارامترها: مفصل گوسین ۱۰۳
جدول۵-۴- برآورد پارامترها: مفصل t استیودنت ۱۰۴
جدول۵-۵ نتایج برازش مدل‌های مختلف با استفاده از روش LR در مثال میزان برداشت گندم ۱۰۶
جدول۵-۶- برآوردضرایب رگرسیونی برای پاسخ اول در مثال جهانی‌سازی ۱۱۲
جدول۵-۷- برآوردضرایب رگرسیونی برای پاسخ دوم در مثال جهانی‌سازی ۱۱۳
جدول۵-۸- برآوردمؤلفه‌های واریانس در مثال جهانی‌سازی ۱۱۴
جدول۵-۹- مقایسه AIC مدل‌های مختلف در مثال جهانی‌سازی ۱۱۴

فصل اول

مقدمه

۱-۱ موضوع و پیشینه تحقیق

دراکثر تحقیقات کاربردی، بخصوص درمباحث مربوط‌به‌علوم زیستی‌ومطالعات اقتصادی، عموماً به‌دلیل تفاوت‌های خاصی‌که بین واحدهای آزمایشی وجود دارد، تغییرپذیری پاسخ‌های مربوط به واحدها نیاز به مدل‌سازی ویژه‌ای دارد. درصد عمده‌ای ازاین تغییرپذیری غالباً به سبب وجود تفاوت‌هایی است که قبل ویا در حین آزمایش‌ایجاد شده است. اگراین منبع تغییرپذیری قابل تفکیک ازخطا و اثرهای آزمایشی باشد آنگاه می‌توان دقت آزمایشها را با برازش مدل‌های تحلیل واریانس افزایش داد و نیز توزیع‌هایی‌که بهتر برازنده داده‌ها هستند را انتخاب نمود (دیویس[1]، ۲۰۰۲). امروزه بخش قابل توجهی از طرح‌های مطالعاتی درعلوم مختلف، ازجمله آزمایش‌های کلینیکی و اطلاعات وابسته به موجودات زنده، ایجاب می‌کندکه واحدهای آزمایشی در زمان‌ها و یا موقعیت‌های مشخصی، به‌منظورکنترل تفاوت‌های بین آزمودنی‌ها، اندازه‌گیری شوند. دراین نوع آزمایش‌ها که هرواحد آزمایشی بطورمتناوب اندازه‌گیری می‌شود معمولاً از اصطلاح اندازه‌های مکرر[2] و در برخی حالات خاص از واژه داده‌های طولی[3] استفاده می‌گردد (دیگل و همکاران[4]، ۲۰۰۲؛ فیتزموریس و همکاران[5]، ۲۰۰۹).

---

1. Davis
2. Repeated Measurements
3. Longitudinal Data
4. Diggle et al.
5. Fitzmaurice et al.

ویژگی‌های منحصر به فرد هر یک از واحدهای آزمایشی غالباً وابستگی اندازه‌های مکرر مربوط به هر واحد آزمایشی را به همراه خواهد داشت. به‌این‌ترتیب برخلاف سایر آزمایش‌ها بجای مشاهدات مستقل با مجموعه‌ای از مشاهدات وابسته سروکار خواهیم داشت. به دلیل آنکه اندازه‌های مکرر از دو بعد زمان و واحدهای آزمایشی تشکیل می‌شوند استفاده از مدل‌های مناسب که اثر مؤلفه‌های مذکور را در توصیف این داده‌ها به درستی لحاظ کند، ضروری است. یکی از پرکاربردترین مدل‌هایی که در تحلیل این نوع داده‌ها به کار می‌رود مدل‌های خطی با اثرهای آمیخته (LMM[1]) است (لرد و ویر[2]، ۱۹۸۲ ؛ مولر و استوارت[3]، ۲۰۰۶) که به مدل با ضرایب تصادفی[4] و یا چندسطحی[5] نیز معروف است.

مدل با اثرهای آمیخته معمولاً برای‌اندازه‌های مکرر، داده‌های وابسته و برخی داده‌های طبقه‌بندی شده به خوبی بکار می‌رود. چراکه دراین مشاهدات به علت نقض فرض استقلال مشاهدات استفاده از مدل‌های خطی معمولی مناسب نیست. در مدل‌های خطی با اثرهای آمیخته وجود اثرهای تصادفی باعث تبیین تغییرپذیری درمشاهدات تکراری و تعدیل ساختار همبستگی بین‌اندازه‌های مکرر مربوط به واحدها و یا مشاهدات وابسته درون طبقات شده واین امر موجب افزایش دقت وکارایی برآورد پارامترها می‌گردد (فیتزموریس و همکاران، ۲۰۰۹).

در حالت کلی یک مدل خطی با اثرهای آمیخته بصورت زیر معرفی می‌شود

$$\mathbf{Y}_{N\times 1} = \mathbf{X}_{N\times p}\boldsymbol{\beta}_{p\times 1} + \mathbf{Z}_{N\times q}\mathbf{u}_{q\times 1} + \mathbf{e}_{N\times 1},$$

که درآن $\boldsymbol{Y}$ بردار مقادیر مشاهده شده متغیرپاسخ، ماتریس پررتبه ستونی $\boldsymbol{X}$ شامل متغیرهای توضیحی، $\boldsymbol{\beta}$ بردار پارامترهای مجهول اثرهای ثابت، بردار $\boldsymbol{u}$ نمایانگر اثرهای تصادفی و ماتریس $\boldsymbol{Z}$ شامل برخی متغیرهای توضیحی و نشانگراست که با توجه به فرض‌های موجود درباره اثرهای تصادفی $\boldsymbol{u}$ در مدل ساخته می‌شوند. $\boldsymbol{e}$ نیز برداری از مؤلفه‌های خطای تصادفی مدل است که مستقل از $\boldsymbol{U}$ توزیع می‌گردد. استنباط آماری پارامترهای مدل بر اساس تابع درستنمایی حاشیه‌ای

$$L(\boldsymbol{\beta},\boldsymbol{\theta}\mid \mathbf{x},\mathbf{y}) = \int_{R^q} f(\mathbf{y}\mid \mathbf{x},\mathbf{u},\boldsymbol{\beta})g(\mathbf{u}\mid\boldsymbol{\theta})d\mathbf{u}$$

---

[1]. Linear Mixel Model
[2]. Laird and Ware
[3]. Muller and Stewart
[4]. Random Coefficient Model
[5]. Multilevel

انجام می‌شود که معمولاً درآن تابع چگالی اثرهای تصادفی $g(\boldsymbol{u}|\theta)$ و چگالی شرطی پاسخ $f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{u},\beta)$ نرمال فرض می‌شوند. یکی از متداول‌ترین روش‌های برآوردیابی پارامترهای مجهول در مدل با اثرهای آمیخته، روش حداکثر درستنمایی (لی‌ونلدر[1]، ۱۹۹۶) است. استفاده از این روش منجربه حل انتگرال‌های پیچیده در تابع درستنمایی حاشیه‌ای، بخصوص درحالت غیرنرمال بودن توزیع اثرهای تصادفی شده که به روش‌های تکراری عددی پیشرفته نیاز دارد. در مطالعات اخیر روش مربع‌بندی گوسی سازوار (مک‌کولا و سرل[2]، ۲۰۰۱؛ پین‌هیرو و چاو[3]، ۲۰۰۶) بعنوان یک روش مؤثر برای تقریب عددی انتگرال‌های پیچیده به‌منظور یافتن برآورد ML پارامترها در مدل با اثرهای تصادفی نرمال ارائه شده است.

در برازش مدل به داده‌های واقعی ممکن است فرآیند توزیع مشاهدات به‌گونه‌ای باشد که استفاده از توزیع نرمال برای مؤلفه‌های خطا و اثرهای تصادفی منتهی به نتایج نادرست و غیرقابل اعتبارگردد. اما اگر توزیع اثرهای تصادفی نرمال نباشد باید روشی را بکار برد که تقریب خوبی از انتگرال‌های موجود در تابع درستنمایی حاشیه‌ای را ارائه دهد.

پین هیرو و همکاران (۲۰۰۱) برازش مدل‌های خطی با اثرهای تصادفی غیرنرمال بر مشاهدات طولی که دارای نقاط پرت هستند را با فرض توزیع T چند متغیره برای اثرهای تصادفی و خطاها، مورد بررسی قرار داده‌اند. آنان با در نظر گرفتن یک ساختار سلسله مراتبی نرمال- گاما برای اثرهای تصادفی به برازش مدل پرداخته و نتایج آن را با نتایج حاصل از برازش مدل با اثرهای تصادفی نرمال مقایسه کردند. همچنین نشان دادند که با وجود مشاهدات پرت توزیع T در مقایسه با توزیع نرمال عموماً برازش بهتری را از مدل به همراه دارد. ژنگ و دیویدیان[4] (۲۰۰۱) نیز یک توزیع قابل انعطاف برای اثرهای تصادفی معرفی کرده‌اند که یک حالت خاص آن نرمال است.

نلسون و همکاران[5] (۲۰۰۶) برآورد ML پارامترهای مجهول در مدل‌های خطی وتعمیم یافته خطی با اثرهای تصادفی غیر نرمال را با استفاده از یک روش محاسباتی بر مبنای قضیه تبدیل احتمال انتگرال و روش مربع‌بندی گوسی بدست آورده‌اند. همچنین لیو وهانگ[6] (۲۰۰۷) ازاین روش جهت برازش مدل‌های شکننده گاما استفاده کردند.

---

1. Lee and Nelder
2. McCulloch and Searle
3. Pinheiro and Chao
4. Zhang and Davidian
5. Nelson et al.
6. Liu and Huang

از آنجاکه روش نلسون و همکاران (۲۰۰۶) تنها در مواردی که شکل معکوس تابع‌توزیع اثرهای تصادفی صریح باشد، کاربرد دارد لیو و یو[1] (۲۰۰۷) یک روش ابتکاری برای یافتن برآورد ML پارامترها ارائه دادند. دراین روش با گنجاندن عبارت کسری ضرب وتقسیم تابع چگالی نرمال استاندارد در عبارت تحت انتگرال و فرمول‌بندی مجدد تابع درستنمایی شرطی بر اساس اثرهای تصادفی نرمال، شرایط استفاده از روش مربع‌بندی گوسی برای حل انتگرال‌های پیچیده امکان‌پذیر شده است. آنها این روش را برای برازش مدل‌های خطی با اثرهای آمیخته‌ای که از یک توزیع آمیخته متناهی پیروی می‌کند و همچنین مدل‌های شکننده گاما با اثرهای تصادفی غیر نرمال ، بکار برده‌اند.

در برخی از مدل‌های آماری با اثرهای آمیخته ممکن است با مدل‌هایی که حاوی بیش ازیک متغیر پاسخ هستند مواجه شویم. در شرایطی که متغیرهای پاسخ، مستقل از یکدیگر باشند می‌توان مدل‌های مجزایی را به هریک از آنها برازش داد. اما انجام این عمل و تحلیل داده‌ها بر اساس آن مناسب نیست. زیرا دراین مدل‌ها علاوه بر تحلیل همبستگی بین‌اندازه‌های مکرر مربوط به هر یک از متغیرهای پاسخ همبستگی بین متغیرهای پاسخ نیز می‌بایست مدل شود.

تاکنون روش‌های متعددی برای مدل‌سازی دو پاسخ پیوسته باتوزیع‌های نرمال و همچنین دو پاسخ پیوسته و دودویی با توزیع‌های نرمال و دوجمله‌ای ارائه شده است. بعنوان مثال ثیابت و همکاران[2] (۲۰۰۱) روشی عملی را با استفاده از ستور $Proc\ MIXED$ در نرم افزار SAS برای مدل‌سازی پاسخ‌های چندمتغیره پیوسته ارائه داده‌اند. آنان ابتدا برای هر یک از متغیرهای پاسخ یک مدل خطی شامل یک مؤلفه تصادفی، مستقل از مؤلفه خطا، برازش داده و سپس با درنظر گرفتن یک توزیع نرمال چندمتغیره برای مؤلفه‌های تصادفی به تحلیل ساختار همبستگی متغیرهای پاسخ براساس ماتریس همبستگی توزیع نرمال پرداخته‌اند. همچنین برآورد پارامترهای مجهول را تحت فرض نرمال بودن توزیع حاشیه‌ای و توأم متغیرهای پاسخ و توزیع مؤلفه‌های تصادفی بدست آوردند.

همانگونه که در تحلیل پاسخ‌های یک متغیره، پاسخ می‌تواند پیوسته، گسسته و یا دودویی باشد در مطالعات مربوط به پاسخ‌های چندمتغیره نیز پاسخ‌هاهمواره رفتار نرمال و یکسانی نخواهند داشت و ممکن است از مقیاس‌ها و توزیع‌های یکسان و یا متفاوتی انتخاب شوند. جورجیوا[3] (۲۰۰۲) برای تحلیل آمیخته‌ای از پاسخ‌های گسسته و

---

1. Liu and Yu
2. Thiebaut et al.
3. Gueorguieva

پیوسته غیرنرمال به تعمیم مدل‌های غیرخطی برای حالات چندمتغیره پرداخت و سپس روش برآوردیابی پارامترهای مجهول را مورد بررسی قرار داد.

موضوع مهمی که در تحلیل داده‌های وابسته مطرح می‌شود تعیین ساختار کوواریانس بین اثرهای تصادفی است که بطور مناسبی همبستگی بین اثرها را نشان دهد. در تحقیقات کاربردی توزیع نرمال چندمتغیره با یک ماتریس کوواریانس معلوم که اغلب دارای خاصیت تقارن مرکب است، استفاده می‌شود. اما پارامترهای وابستگی دراین توزیع تنها قادر به‌اندازه‌گیری درجه همبستگی خطی بین متغیرها بوده و وقتی که همبستگی بین متغیرها از نوع غیرخطی و یا غیریکنواخت باشد، نمی‌توان با استفاده ازاین پارامترها ساختار همبستگی بین متغیرها را به طور دقیق تعیین کرد. علاوه براین، در توزیع نرمال و سایر توزیع‌های بیضوی همواره حاشیه‌ای‌ها یکسان فرض می‌شوند. لذا ممکن است درمواقعی که رفتار فرآیند تولید مشاهدات مختلف یکسان نیست، توزیع‌های بیضوی برای مدل‌سازی و ایجاد ارتباط بین آنها مناسب نباشند. برای حل این مشکل می‌توان از توابع مفصل در مدل سازی مشاهدات پاسخ‌های وابسته بهره گرفت.

واژه مفصل[1] در آمار برای اولین بار توسط اسکلار[2] (۱۹۵۹)، به منظور معرفی توابعی که توزیع‌های چندبعدی را به توزیع‌های حاشیه‌ای با ابعاد کمتر ارتباط می‌دهند، به کار رفت. سپس مفصل بعنوان یک فن مناسب برای فرمول‌بندی و ساخت خانواده توزیع‌های چند متغیره و نیز روشی برای مطالعه‌اندازه‌های آزاد از مقیاس وابستگی، مورد توجه محققان قرار گرفت.

معرفی توابع مفصل مختلف، شرح خواص و ویژگی‌های آن‌ها به طور مبسوط در نلسن[3] (۲۰۰۶) آمده است. از آنجایی‌که براساس هر یک از توابع مفصل می‌توان توزیع‌های چندمتغیره مختلفی را ساخت، جی‌نست و مکای[4] (۱۹۸۶) تابع چگالی احتمال چند خانواده از توابع مفصل را در حالت کلی بدست آورده‌اند. (تریودی و زیمر[5]، ۲۰۰۷) نیز روش‌هایی را به‌منظور یافتن مفصلی که بهترین برازش را به مشاهدات داشته باشد، معرفی کرده‌اند.

در دو دهه اخیر تحقیقات وسیعی در زمینه تئوری و کاربرد توابع مفصل در مدل‌سازی‌های رگرسیونی انجام شده است. ازجمله بریکن و همکاران[6] (۲۰۰۷) توابع مفصل‌رابرای مدل‌سازی جداگانه ساختار وابستگی بین مانده‌ها و توزیع‌های حاشیه‌ای آنها در مدل‌های خطی به کار برده‌اند. همینطور سان و همکاران[7] (۲۰۰۸) کاربرد توابع مفصل

---

1. Copula
2. Sklar
3. Nelsen
4. Genest and Mackay
5. Trivedi and Zimmer
6. Braeken et al.
7. Sun et al.

را در تحلیل وابستگی بین مشاهدات طولی دم‌سنگین مورد بررسی قرار داده‌اند. اخیراً یک دیدگاه جدید برمبنای استفاده از مفصل‌ها برای تحلیل همزمان آمیخته‌ای از پاسخ‌های وابسته‌ی گسسته و پیوسته توسط سانگ و همکاران[1] (۲۰۰۹) ارائه شده است.

انتخاب ضریب همبستگی ρ برای بررسی ساختار همبستگی بین متغیرها به دلیل وجود برخی ویژگی‌های خاص آن ممکن است مشکلاتی به همراه داشته باشد (یان،۲۰۰۶). برای مثال، ضریب ρ تنها درجه وابستگی خطی[2] بین متغیرها را اندازه‌گیری می‌کند وتحت تبدیلات غیرخطی اکیداً یکنوا، پایا[3] نیست و همین ویژگی باعث می‌شود که از صفر بودن این ضریب همبستگی همواره نتیجه استقلال متغیرهای غیرنرمال حاصل نگردد. همچنین محاسبه این ضریب تنها در حالت متناهی بودن گشتاور مرتبه دوم متغیرها امکان‌پذیراست و برای برخی توزیع‌های دم‌سنگین مانند توزیع کوشی کاربرد ندارد. یکی‌دیگراز اشکالات ρ این است که براساس‌این ضریب همبستگی و اطلاعات حاشیه‌ها نمی‌توان توزیع توأم را بصورت دقیق تعیین نمود. به همین دلیل به ازای انتخاب‌های مناسب توزیع توأم دومتغیره برای دو توزیع حاشیه‌ای موردنظر نیز، نمی‌توان تمامی‌مقادیر بین ۱- تا ۱+ را برای ρ به‌دست آورد. ازاین‌رو آماردانان به معرفی اندازه‌های ارتباط دیگری مانند ρ_sاسپیرمن‌و τ کندال(امبرچتز وهمکاران[4]،۲۰۰۱) پرداخته و ارتباط آنها را با توابع مفصل بررسی کرده‌اند. یان (۲۰۰۶) نیزساختارهای مختلف همبستگی بین متغیرها را برای خانواده‌های مختلف توابع مفصل بررسی کرده است.

از آنجا که توابع مفصل را می‌توان برای بررسی ساختارهای غیرخطی همبستگی بین متغیرها بکار برد، دراین پایان نامه این توابع به‌منظور ایجاد ساختارهای مختلف همبستگی برای اثرهای تصادفی در مدل‌های آمیخته و همچنین برقراری ارتباط بین متغیرهای پاسخ وابسته در مدل‌های چندمتغیره استفاده شده‌اند.

۱-۲ اهداف تحقیق

هدف از انجام این تحقیق بررسی روش‌های مدل‌سازی و تحلیل مشاهدات وابسته است. بدین منظور با توجه به اثر واحدهای آزمایشی بصورت تصادفی، برازش مدل‌های خطی و تعمیم یافته خطی با اثرهای آمیخته را مدنظر قرار داده‌ایم. سپس به معرفی و مقایسه دو روش فرمول‌بندی مجدد درستنمایی(LR[5]) و تبدیل احتمال انتگرال (PIT[6])

---

[1]. Song et al.
[2]. Linear Dependence Measure
[3]. Invariant
[4]. Embrechts et al.
[5]. Likelihood Reformulation
[6]. Probability Integral Transformation

که برای برازش‌این مدل‌ها در حالت غیرنرمال بودن توزیع اثرهای تصادفی کاربرد دارد، پرداخته و روش‌های فوق را در مسائل مربوط به تحلیل بقا و همچنین برازش انواع مدل‌های آمیخته بکار گرفته‌ایم. سپس با مطالعات شبیه سازی به مقایسه مدل‌های مختلف پرداخته ونشان داده‌ایم که نتایج حاصل، به فرض‌های در نظر گرفته شده برای توزیع داده‌ها و نیز توزیع اثرهای تصادفی حساس است. همچنین داده‌های واقعی مختلفی را تحلیل کرده و توسط معیارهای انتخاب مدل به انتخاب برازنده‌ترین مدل به داده‌ها پرداخته‌ایم. در این راستا نشان داده‌ایم که توزیع نرمال همواره توزیع مناسبی برای اثرهای تصادفی نیست.

از دیگر اهداف مهم این تحقیق، کاربرد توابع مفصل، بعنوان روشی مناسب برای تولید توزیع‌های چندمتغیره در مدل‌سازی‌های آماری است. به‌دلیل ویژگی‌های منحصربه‌فرد توابع مفصل ازجمله قابلیت تولید ساختارهای مختلف همبستگی غیرخطی و ایجاد توزیع‌های حاشیه‌ای متفاوت می‌توان این توابع را برای برقراری ارتباط معقول بین متغیرها در مدل‌های رگرسیونی بکار برد. این توابع وقتی متغیرها از توزیع‌های یکسان نبوده و یا ساختار همبستگی بین آنها خطی نباشد، مفید هستند. دراین تحقیق ما از توزیع‌های چندمتغیره متناظر با توابع مفصل برای اثرهای تصادفی در مدل‌های یک متغیره خطی و تعمیم یافته خطی با اثرهای آمیخته بهره برده‌ایم. همچنین در مدل‌های چندمتغیره نیز ما توابع مفصل را برای ایجاد ارتباط بین پاسخ‌های وابسته به کار برده‌ایم. در پایان با گنجاندن چند مثال تجربی با داده‌های واقعی کاربرد توابع مفصل را در برازش مدل‌های رگرسیونی آمیخته نشان داده‌ایم.

۱-۳ اهمیت و کاربرد تحقیق

دربسیاری از تحقیقات کاربردی داده‌های طولی روی مجموعه‌ای از واحدهای آزمایشی در دوره‌های زمانی مختلف جمع‌آوری می‌شوند. باوجود همبستگی بین این مشاهدات استفاده از مدل‌های خطی متداول برای تحلیل این داده‌ها منطقی نیست. لذا جایگزینی مدل‌های مناسب یکی از مباحث مهم است که بطور مناسب این همبستگی را در مدل‌سازی لحاظ کند.

یکی‌از فنون مناسب درتحلیل داده‌های طولی، واردکردن اثرهای تصادفی درمدل است تا ضمن کنترل تغییرپذیری بین واحدها، همبستگی بین مشاهدات را نیز درنظر بگیرد. اما انجام این عمل نیز اگر به درستی انجام نگیرد باعث برازش مدل نامناسب، نتایج نادرست و در نتیجه اتخاذ تصمیمات غیرمنطقی در زمینه‌های کاربردی می‌شود. پس مسئله انتخاب بهترین مدل برازنده به مشاهدات بسیار حائز اهمیت است. یکی از نکات مهم در این زمینه انتخاب

مناسب‌ترین توزیع برای اثرهای تصادفی و مؤلفه خطا است. از طرفی فرض توزیع‌های غیرنرمال به دلیل ایجاد ساختارهای پیچیده برای توابع درستنمایی، موجب بروز مشکل در روند برآوردیابی پارامترهای مدل می‌گردد. دراین پایان نامه با معرفی روش‌های جدید به حل این مشکل پرداخته‌ایم. سپس با درنظرگرفتن توزیع‌های متفاوت برای اثرهای تصادفی و مؤلفه‌های خطا در مدل‌های مختلف و با استفاده از معیارهای مناسب به انتخاب مدل برتر می‌پردازیم.

اهمیت دیگر این تحقیق، کاربرد توابع مفصل برای برقراری ارتباط بین متغیرهای وابسته است که بر این اساس خانواده‌های جدیدی از توزیع‌های چندمتغیره در مدل‌سازی معرفی می‌شوند. دراین‌راستا، علاوه بر بررسی ویژگی‌های منحصربه‌فرد توابع مفصل در ارتباط با تولید ساختارهای غیرخطی همبستگی، کاربرد آنها را در مدل-سازی مشاهدات درنظر گرفته‌ایم.

یکی دیگر از مباحث مهم آماری برازش مدل‌های رگرسیونی چندمتغیره با اثرهای آمیخته است. در این مدل‌ها استفاده از ساختار مناسب همبستگی بین متغیرهای پاسخ حائزاهمیت است. در این پایان‌نامه ضمن روشن کردن این موضوع، توابع مفصل را به‌منظور ایجاد ارتباط بین بردارهای پاسخ و ساختن توزیع‌های چندمتغیره جدید براساس آنها برای بردار اثرهای آمیخته بکار گرفته‌ایم. این موضوع نه تنها باعث‌ایجاد ساختارهای همبستگی انعطاف‌پذیرتر بین پاسخ‌های چندمتغیره می‌شود، بلکه امکان انتخاب توزیع‌های حاشیه‌ای متفاوت و غیرنرمال را برای هریک از متغیرهای پاسخ بوجود می‌آورد. این موضوع همواره در مدل‌های رگسیونی با اثرهای آمیخته نادیده گرفته شده است. بعبارت دیگر در این مدل‌ها برای توزیع‌های حاشیه‌ای تمامی‌متغیرهای پاسخ همواره توزیع‌هایی یکسان و بیضوی فرض شده است. یکی از ویژگی‌های این پایان‌نامه پیشنهاد استفاده از تابع مفصل جهت حل مشکل فوق و ارائه ساختاری مناسب برای همبستگی بین پاسخ‌ها می‌باشد.

۱-۴ محاسبات آماری

دراین تحقیق برای اجرای کلیه فرآیندهای مربوط به تولید داده‌ها و برآوردیابی پارامترهای مجهول در مدل‌های مختلف از نرم افزارهای آماری *SAS*، *R* و *Winbugs* استفاده شده است.

در نرم افزار *SAS* دستور *Proc MIXED* را برای برازش مدل‌های خطی با اثرهای آمیخته تحت فرض نرمال بودن توزیع خطا و اثرهای تصادفی، بکار گرفته‌ایم. همچنین از دستور *Proc NLMIXED* برای برازش مدل‌های خطی تعمیم یافته با اثرهای آمیخته وقتی توزیع خطا و یا اثرهای تصادفی نرمال هستند، بهره برده‌ایم.