دانشگاه تربیت مدرس

دانشکده علوم ریاضی

رساله دوره دکتری آمار

تحلیل بیزی با تاکید بر آزمون همگنی مدل‌های مارکوف برای داده‌های طولی

توسط

سجاد نوریان

استاد راهنما

دکتر مجتبی گنجعلی

استادان مشاور

دکتر خلیل شفیعی هولیقی

دکتر موسی گل‌علی‌زاده لهی

آبان ۱۳۹۲

بسم الله الرحمن الرحيم

دانشگاه تربیت مدرس

دانشکده علوم ریاضی

بسمه تعالی

تأییدیه اعضای هیأت داوران حاضر در جلسه دفاع از رساله دکتری

آقای سجاد نوریان به شماره دانشجویی ۸۷۵۷۰۲۰۰۳ رساله واحدی خود را با عنوان:

«تحلیل بیزی با تأکید بر آزمون همگنی مدل‌های مارکوف برای داده‌های طولی» در تاریخ ۹۲/۸/۲۰ ارائه کردند.

اعضای هیأت داوران نسخه نهایی این رساله را از نظر فرم و محتوا تأیید کرده است و پذیرش آن را برای تکمیل درجه دکتری پیشنهاد می‌کند.

اعضای هیأت داوران	نام و نام‌خانوادگی	رتبه علمی	امضاء
۱- استاد راهنما	آقای دکتر مجتبی گنجعلی	استاد	
۲- استاد مشاور	آقای دکتر خلیل شفیعی	استادیار	
۳- استاد مشاور	آقای دکتر موسی گل‌علی‌زاده	استادیار	
۴- استاد ناظر داخلی	آقای دکتر محسن محمدزاده	استاد	
۵- استاد ناظر داخلی	آقای دکتر مجید جعفری‌خالدی	استادیار	
۶- استاد ناظر خارجی	آقای دکتر حمید پزشک	استاد	
۷- استاد ناظر خارجی	آقای دکتر فرزاد اسکندری	دانشیار	
۸- نماینده شورای تحصیلات تکمیلی	آقای دکتر محسن محمدزاده	استاد	

آیین نامه چاپ پایان‌نامه (رساله)های دانشجویان دانشگاه تربیت مدرس

نظر به اینکه چاپ و انتشار پایان نامه (رساله)های تحصیلی دانشجویان دانشگاه تربیت مدرس، مبین بخشی از فعالیتهای علمی - پژوهشی دانشگاه است بنابراین به منظور آگاهی و رعایت حقوق دانشگاه،دانش آموختگان این دانشگاه نسبت به رعایت موارد ذیل متعهد می‌شوند:

ماده ۱: در صورت اقدام به چاپ پایان نامه (رساله)ی خود، مراتب را قبلاً به طور کتبی به «دفتر نشر آثارعلمی» دانشگاه اطلاع دهد.

ماده ۲: در صفحه سوم کتاب (پس از برگ شناسنامه) عبارت ذیل را چاپ کند:

«کتاب حاضر، حاصل رساله دکتری نگارنده در رشته **آمار** است که در سال **۱۳۹۲** در دانشکده **علوم ریاضی** دانشگاه تربیت مدرس به راهنمایی جناب آقای دکتر **مجتبی گنجعلی**، مشاوره جناب آقای دکتر **خلیل شفیعی هولیقی** و مشاوره جناب آقای دکتر **موسی گل‌علی‌زاده لهی** از آن دفاع شده است.»

ماده ۳: به منظور جبران بخشی از هزینه‌های انتشارات دانشگاه، تعداد یک درصد شمارگان کتاب (در هر نوبت‌چاپ) را به «دفتر نشر آثارعلمی» دانشگاه اهدا کند. دانشگاه می‌تواند مازاد نیاز خود را به نفع مرکز نشر درمعرض فروش قرار دهد.

ماده ۴: در صورت عدم رعایت ماده ۳، ۵۰٪ بهای شمارگان چاپ شده را به عنوان خسارت به دانشگاه تربیت‌مدرس، تأدیه کند.

ماده ۵: دانشجو تعهد و قبول می کند در صورت خودداری از پرداخت بهای خسارت، دانشگاه می‌تواند خسارت مذکور را از طریق مراجع قضایی مطالبه و وصول کند؛ به علاوه به دانشگاه حق می‌دهد به منظور استیفای حقوق خود، از طریق دادگاه، معادل وجه مذکور در ماده ۴ را از محل توقیف کتابهای عرضه شده نگارنده برای فروش، تامین نماید.

ماده ۶: اینجانب **سجاد نوریان** دانشجوی رشته **آمار** مقطع **دکتری**

تعهد فوق وضمانت اجرایی آن را قبول کرده، به آن ملتزم می شوم.

نام و نام خانوادگی: **سجاد نوریان**

تاریخ و امضا: ۹۲/۸/۲۰

آیین‌نامه حق مالکیت مادی و معنوی در مورد نتایج پژوهشهای علمی دانشگاه تربیت مدرس

مقدمه: با عنایت به سیاست‌های پژوهشی و فناوری دانشگاه در راستای تحقق عدالت و کرامت انسانها که لازمه شکوفایی علمی و فنی است و رعایت حقوق مادی و معنوی دانشگاه و پژوهشگران، لازم است اعضای هیأت علمی، دانشجویان، دانش‌آموختگان و دیگر همکاران طرح، در مورد نتایج پژوهشهای علمی که تحت عناوین پایان‌نامه، رساله و طرحهای تحقیقاتی با هماهنگی دانشگاه انجام شده است، موارد زیر را رعایت نمایند:

ماده ۱- حق نشر و تکثیر پایان نامه/ رساله و درآمدهای حاصل از آنها متعلق به دانشگاه می باشد ولی حقوق معنوی پدید آورندگان محفوظ خواهد بود.

ماده ۲- انتشار مقاله یا مقالات مستخرج از پایان‌نامه/ رساله به صورت چاپ در نشریات علمی و یا ارائه در مجامع علمی باید به نام دانشگاه بوده و با تایید استاد راهنمای اصلی، یکی از اساتید راهنما، مشاور و یا دانشجو مسئول مکاتبات مقاله باشد. ولی مسئولیت علمی مقاله مستخرج از پایان نامه و رساله به عهده اساتید راهنما و دانشجو می باشد.

تبصره: در مقالاتی که پس از دانش‌آموختگی بصورت ترکیبی از اطلاعات جدید و نتایج حاصل از پایان‌نامه/ رساله نیز منتشر می‌شود نیز باید نام دانشگاه درج شود.

ماده ۳- انتشار کتاب، نرم افزار و یا آثار ویژه (اثری هنری مانند فیلم، عکس، نقاشی و نمایشنامه) حاصل از نتایج پایان‌نامه/ رساله و تمامی طرحهای تحقیقاتی کلیه واحدهای دانشگاه اعم از دانشکده ها، مراکز تحقیقاتی، پژوهشکده ها، پارک علم و فناوری و دیگر واحدها باید با مجوز کتبی صادره از معاونت پژوهشی دانشگاه و براساس آئین نامه های مصوب انجام شود.

ماده ۴- ثبت اختراع و تدوین دانش فنی و یا ارائه یافته ها در جشنواره‌های ملی، منطقه‌ای و بین‌المللی که حاصل نتایج مستخرج از پایان‌نامه/ رساله و تمامی طرح‌های تحقیقاتی دانشگاه باید با هماهنگی استاد راهنما یا مجری طرح از طریق معاونت پژوهشی دانشگاه انجام گیرد.

ماده ۵- این آیین‌نامه در ۵ ماده و یک تبصره در تاریخ۸۷/۴/۱ در شورای پژوهشی و در تاریخ ۸۷/۴/۲۳ در هیأت رئیسه دانشگاه به تایید رسید و در جلسه مورخ ۸۷/۷/۱۵ شورای دانشگاه به تصویب رسیده و از تاریخ تصویب در شورای دانشگاه لازم‌الاجرا است.

«اینجانب **سجاد نوریان** دانشجوی رشته **آمار** ورودی سال تحصیلی **۱۳۸۷** مقطع **دکتری** دانشکده **علوم ریاضی** متعهد می شوم کلِیه نکات مندرج در آئین نامه حق مالکیت مادی و معنوی در مورد نتایج پژوهش های علمی دانشگاه تربیت مدرس را در انتشار یافته های علمی مستخرج از رساله تحصیلی خود رعایت نمایم. در صورت تخلف از مفاد آئین نامه فوق الاشعار به دانشگاه وکالت و نمایندگی می دهم که از طرف اینجانب نسبت به لغو امتیاز اختراع بنام بنده و یا هر گونه امتیاز دیگر و تغییر آن به نام دانشگاه اقدام نماید. ضمناً نسبت به جبران فوری ضرر و زیان حاصله بر اساس برآورد دانشگاه اقدام خواهم نمود و بدینوسیله حق هر گونه اعتراض را از خود سلب نمودم»

امضا:

تاریخ: **۹۲/۸/۲۰**

تقدیم به پدر بزرگوارم که

همچون اشک از چشمم رفت و باز نمی‌آید

و تقدیم به همسر مهربانم، فرشته، و دختر دلبندمان کیمیا

سپاس گزاری...

وگویند سپاس خدای راکه مارابدین راه رهنمون شدو اگر وی راهبر نبود، هرگز راه نمی یافتیم (اعراف، ۴۳)

ستایش و سپاس معبود یگانه را که در پرتو الطاف بی‌شمارش، بدین راه آمدیم و آمدنمان نه از پی حشمت و جاه بود و نه از بد حادثه. رهرو منزل عشق بودیم و این راه تنها با لطف و مددش طی شد و بس. خدایا خود در ادامه این راه یاورمان باش.

در آغاز وظیفه‌ی خود می‌دانم از راهنمایی‌ها و زحمات استاد گرانقدرم جناب آقای دکتر مجتبی گنجعلی در به ثمر رساندن این رساله کمال تشکر و قدردانی را داشته باشم، چرا که با حمیت و جدیت، مرا به دقت، اندیشه، درک و تعمق وامی‌داشتند.

همچنین از اساتید بزرگوار جناب آقایان دکتر خلیل شفیعی هولیقی و دکتر موسی گل علیزاده لهی که مشاوره رساله را عهده‌دار بودند و همواره از رهنمودهای ایشان بهره‌مند بودم، کمال تشکر را دارم.

از اساتید گرانبها و عزیز جناب آقای دکتر محسن محمدزاده، جناب آقای دکتر حمید پزشک، جناب آقای دکتر مجید جعفری خالدی و جناب آقای دکتر فرزاد اسکندری که داوری رساله را برعهده داشتند و توصیه‌های ارزشمندی جهت بهبود آن داشتند، سپاسگزاری می‌نمایم.

تشکری ویژه دارم از همسر مهربانم، سرکار خانم فرشته حق شناس، که همواره همراه و پشتیبان من بوده است و همه‌ی سختی‌های این راه را به جان خریده است. از کیمیای عزیزم به خاطر شوری که با آمدنش به ما داد هم ممنونم.

شکرگزار خدای مهربان هستم برای داشتن پدری که سال‌ها از نعمت وجودش بهره‌مند بودم و رهنمودهایشان همواره چراغ راه من بوده است، همچنین برای داشتن مادر عزیزی که مهر مادرانه و دعاهای مهربانانه‌اش همیشه همراهمان بوده و خواهد بود، انشاالله، و نیز وجود برادرانی همیشه همراه.

سجاد نوریان
آبان ۱۳۹۲

چکیده

در این رساله، یک چارچوب بیزی برای تحلیل داده‌های پاسخ ترتیبی طولی ارائه شده است. در تحلیل داده‌های طولی، باید همبستگی‌های بین پاسخ‌های مرتبط با هر آزمودنی درنظر گرفته شود. برای مدل‌بندی این همبستگی‌ها مدل‌های مختلفی را می‌توان به کار برد که عبارتند از مدل‌بندی حاشیه‌ای، مدل‌بندی اثرهای تصادفی و مدل‌بندی انتقالی (مارکوف). به منظور به دست آوردن برآورد پارامترها، از یک مدل رگرسیون لوژستیک تجمعی و رهیافت بیزی، استفاده شده است. برآورد بیزی پارامتر گاما در سطوح مختلف متغیّرهای تبیینی به دست آمده است. سپس، حساسیت خلاصه‌های پسین نسبت به تغییر ابرپارامترهای توزیع پیشین مورد ارزیابی قرار گرفته است.

همچنین، یک آزمون بیزی برای فرضیه‌ی همگنی مدل انتقالی برای تحلیل داده‌های پاسخ ترتیبی طولی ارائه شده است. به عبارت دیگر، فرضیه‌ی ثابت بودن احتمال‌های انتقال آزمودنی mاُم از موقعیت a به موقعیت b در زمان t، (π_{mabt})، برای هر مقدار t، آزمون شده است. همچنین، همگنی برای یک متغیّر تبیینی خاص (یا مجموعه‌ای از متغیّرهای تبیینی) تعریف شده است.

به علاوه، یک چارچوب بیزی برای تحلیل یک مدل انتقالی برای داده‌های پاسخ ترتیبی و پیوسته‌ی آمیخته‌ی طولی ارائه شده است. مدل پیشنهادی، یک مدل آمیخته‌ی چند متغیّره را برای پاسخ‌ها در نظر می‌گیرد. برای مدل‌بندی پاسخ‌های ترتیبی و پیوسته‌ی هر آزمودنی به ترتیب از یک مدل رگرسیون لوژستیک تجمعی انتقالی و یک مدل اتورگرسیو استفاده می‌شود. همچنین به‌منظور در نظر گرفتن ارتباط بین پاسخ‌های ترتیبی و پیوسته طولی در هر زمان از یک پارامتر پیوند استفاده می‌شود. آزمونی برای فرضیه‌ی همگن بودن و همچنین آزمونی برای فرضیه‌ی صفر بودن این پارامتر و در نتیجه ناهمبسته بودن پاسخ‌های ترتیبی و طولی ارائه شده است.

واژه‌های کلیدی: تحلیل بیزی، عامل بیزی، عرض پیشگوی شرطی، همگنی پارامتر پیوند، مطالعه‌های طولی، پاسخ‌های آمیخته‌ی ترتیبی و پیوسته، مدل انتقالی.

فهرست مندرجات

لیست جداول د

لیست اشکال و

۱ مفهوم‌های اولیه ۱

۱.۱ معرفی داده‌های طولی و مقایسه‌ی آن‌ها با داده‌های مقطعی ۱

۲.۱ معیار پیوند گاما برای داده‌های گسسته ۴

۳.۱ مدل‌های مختلف برای تحلیل داده‌های طولی ۶

۴.۱ مفهوم گم‌شدگی و الگوهای گم‌شدگی ۸

۱.۴.۱ انصراف به‌طور کامل تصادفی ۹

۲.۴.۱ انصراف تصادفی ۱۰

۳.۴.۱ انصراف غیرتصادفی ۱۱

۵.۱ سابقه تحقیق ۱۱

۲ مدل‌هایی برای تحلیل داده‌های طولی با تأکید بر مدل‌های انتقالی ۲۱

۱.۲ مدل‌های مختلف تحلیل داده‌های طولی ۲۱

۱.۱.۲ مدل‌های حاشیه‌ای ۲۲

۲.۱.۲ مدل‌های اثرهای تصادفی ۲۵

۳.۱.۲ مدل‌های انتقالی (مارکوف) ۲۷

۴.۱.۲ مقایسه رهیافت‌ها ۲۹

۲.۲ مدل‌های انتقالی و کاربرد آن‌ها در تحلیل داده‌های طولی ۳۴

۱.۲.۲ مدل انتقالی و رَویه‌های برازش آن ۳۷
۲.۲.۲ مدل‌های انتقالی برای داده‌های گسسته ۴۱
۳.۲.۲ مدل‌های انتقالی برای داده‌های ترتیبی ۴۴
۳.۲ تحلیل داده‌های طولی در حضور داده‌های گم‌شده ۴۹
۱.۳.۲ مدل‌بندی الگوی انصراف ۵۴
۲.۳.۲ ساختار کوواریانس داده‌ها ۵۸
۳.۳.۲ تابع درستنمایی ۶۲
۴.۲ بحث و نتیجه‌گیری ۶۴

۳ تحلیل بیزی مدل‌های انتقالی ۶۵
۱.۳ تحلیل بیزی مدل‌های انتقالی بدون مقادیر گم‌شده ۶۵
۱.۱.۳ مدل‌های انتقالی و تابع درستنمایی ۶۶
۲.۱.۳ تحلیل بیزی مدل‌های انتقالی ۶۹
۳.۱.۳ برآورد بیزی پارامترهای گاما و گامای جزئی ۷۱
۲.۳ تحلیل بیزی مدل‌های انتقالی با مقادیر گم‌شده ۷۲
۳.۳ آزمون فرضیه ۷۴
۴.۳ مثال کاربردی: تحلیل داده‌های بی‌خوابی ۷۶
۱.۴.۳ تحلیل اکتشافی داده‌ها ۷۶
۲.۴.۳ مدل انتقالی و تابع درستنمایی برای داده‌های بی‌خوابی ۸۲
۳.۴.۳ استنباط بیزی برای داده‌های بی‌خوابی ۸۴
۴.۴.۳ آزمون فرضیه ۸۷
۵.۴.۳ عرض پیشگوی شرطی ۹۰
۶.۴.۳ تحلیل حساسیت ۹۲
۵.۳ مثال کاربردی: تحلیل داده‌های فلووکسامین ۹۵
۱.۵.۳ تحلیل اکتشافی داده‌ها ۹۶
۲.۵.۳ مدل‌های انتقالی و تابع درستنمایی برای داده‌های فلووکسامین ۹۹
۳.۵.۳ استنباط بیزی برای داده‌های فلووکسامین ۱۰۰
۶.۳ بحث و نتیجه‌گیری ۱۰۸

۴ همگنی پارامتر پیوند در داده‌های با پاسخ‌های آمیخته‌ی ترتیبی و پیوسته ۱۰۹
۱.۴ مقدمه ۱۰۹
۲.۴ مدل انتقالی توأم و تابع درستنمایی ۱۱۰
۳.۴ تحلیل بیزی ۱۱۵
۴.۴ مثال کاربردی: تحلیل داده‌های آزمون انفرادی پیشرفت پی‌بادی ۱۱۶
۱.۴.۴ آزمون فرضیه ۱۱۹
۵.۴ نتیجه‌گیری و پیشنهادها ۱۲۳

کتاب‌نامه ۱۲۵

نام‌نامه ۱۳۴

واژه‌نامه فارسی به انگلیسی ۱۳۸

واژه‌نامه انگلیسی به فارسی ۱۴۰

لیست جداول

۱.۱.۲ خلاصه ۱۲۰۰ مشاهده بیماری تنفسی، رمد چشم و سن برای ۲۷۵ کودک در مطالعه‌ی سلامت کودکان اندونزیایی (سامر، ۱۹۸۲) ۲۴

۱.۴.۳ زمان به خواب رفتن که با توجه به پاسخ بیماران به این سؤال که "چند دقیقه بعد از رفتن به رختخواب، می‌خوابید؟"، بدست آمده است. پاسخ زمان دوم (Y_2) با توجه به نوع درمان، پاسخ در زمان اول (Y_1)، تعداد مشاهده و درصد در هر سطر بیان شده است. ۷۸

۲.۴.۳ توزیع حاشیه‌ای برای پاسخ‌های زمان اول و دوم با توجه به روش درمانی. ۷۹

۳.۴.۳ انحراف استاندارد و میانگین پسین: $\gamma(0)$ برای گروه دارونما و $\gamma(1)$ برای گروه داروی مؤثر. ۸۲

۴.۴.۳ برآوردهای بیزی و فراوانی‌گرای پارامترها با مدل‌بندی حاشیه‌ای پاسخ‌ها، بدون در نظر گرفتن پیوند بین پاسخ‌ها (پارامترهای معنی‌دار در سطح ۵٪ با * نشان داده شده‌اند). ۸۵

۵.۴.۳ برآوردهای بیزی و فراوانی‌گرای پارامترهای مدل انتقالی (پارامترهای معنی‌دار در سطح ۵٪ با * نشان داده شده‌اند). ۸۷

۶.۴.۳ عامل بیزی برای مدل‌های ارائه شده در داده‌های بی‌خوابی. در هر خانه، عامل بیزی برای مقایسه فرضیه داده شده در ستون مرتبط در مقابل فرضیه داده شده در سطر مرتبط داده شده است. ۸۹

۷.۴.۳ میانگین پسین ضریب‌های رگرسیونی که پس از حذف مشاهده یا مشاهده‌های غیرمعمول بدست آمده‌اند (پارامترهای معنی‌دار در سطح ۵٪ با * نشان داده شده‌اند). ۹۴

۸.۴.۳ میانگین‌های پسین پارامترهای گامای جزئی و ضریب‌های رگرسیونی برای مقادیر مختلف واریانس توزیع‌های پیشین. ۹۵

۹.۵.۳ احتمال‌های نمونه‌ای برای هر سطح از پاسخ‌های Y_1، Y_2 و Y_3. . . . . ۹۸

۱۰.۵.۳احتمال نمونه‌ای مشاهده‌شدن پاسخ و تعداد پاسخ‌های مشاهده شده برای زمان‌های دوم و سوم به شرط سطوح متغیّر پاسخ قبلی. ۹۹

۱۱.۵.۳میانگین پسین نقاط برش و ضریب‌های رگرسیونی برای مدل ناهمگن ممسک [مدل (I)]. ۱۰۳
۱۲.۵.۳میانگین پسین نقاط برش و ضریب‌های رگرسیونی برای مدل همگن جزئی ممسک [مدل (II)]. ۱۰۴
۱۳.۵.۳میانگین پسین نقاط برش و ضریب‌های رگرسیونی برای مدل همگن ممسک [مدل (III)]. ۱۰۵
۱۴.۵.۳عامل بیزی برای مدل‌های ارائه شده در داده‌های فلووکسامین. در هر خانه، عامل بیزی برای مقایسه مدل داده شده در ستون مرتبط در مقابل مدل داده شده در سطر مرتبط داده شده است. ۱۰۵

۱.۴.۴ میانگین پسین و انحراف معیار (در داخل پرانتز) پارامترهای نقطه‌ی برش، ضریب‌های رگرسیونی و پیوند برای مدل (I) (پارامترهای معنی‌دار در سطح ۵٪ با * نشان داده شده‌اند). ۱۲۰
۲.۴.۴ میانگین پسین و انحراف معیار (در داخل پرانتز) پارامترهای نقطه‌ی برش، ضریب‌های رگرسیونی و پیوند برای مدل (II) (پارامترهای معنی‌دار در سطح ۵٪ با * نشان داده شده‌اند). ۱۲۱
۳.۴.۴ میانگین پسین و انحراف معیار (در داخل پرانتز) پارامترهای نقطه‌ی برش، ضریب‌های رگرسیونی و پیوند برای مدل (III) (پارامترهای معنی‌دار در سطح ۵٪ با * نشان داده شده‌اند). ۱۲۲
۴.۴.۴ عامل بیزی برای مدل‌های ارائه شده در داده‌های آزمون انفرادی پیشرفت پی‌بادی. در هر خانه، عامل بیزی برای مقایسه مدل داده شده در ستون مرتبط در مقابل مدل داده شده در سطر مرتبط داده شده است. . . ۱۲۳
۵.۴.۴ ملاک اطلاع انحرافه (DIC) برای سه مدل ارائه شده در داده‌های آزمون انفرادی پیشرفت پی‌بادی. . . ۱۲۳

لیست اشکال

۱.۱.۱ داده‌های فرضی در مورد رابطه‌ی بین سن (محور افقی) و توانایی خواندن (محور عمودی) (دیگل و همکاران، ۲۰۰۲، ص ۲) ۳

۱.۴.۳ تابع چگالی پسین برای پارامترهای β_1^1، β_1^2، β_1^3 و β_1^4. . . . . ۸۶

۲.۴.۳ عرض پیشگوی شرطی برای پاسخ همه آزمودنی‌ها در زمان دوم با توجه به مقدار پاسخ در زمان اول، (الف) برای $Y_{m1} = 1$، (ب) برای $Y_{m1} = 2$، (ج) برای $Y_{m1} = 3$ و (د) برای $Y_{m1} = 4$. ۹۲

۳.۴.۳ تابع چگالی پسین $\gamma(0)$ (راست) و $\gamma(1)$ (چپ) برای مقادیر مختلف واریانس توزیع‌های پیشین ضریب‌های رگرسیونی. ۹۳

فصل ۱

مفهوم‌های اولیه

هدف از این فصل معرفی مفهوم‌های اولیه داده‌های طولی و داده‌های گم‌شده است. در بخش اول تعریف‌ها و مفهوم‌های اساسی داده‌های طولی بیان می‌شود. سپس ویژگی‌های این داده‌ها با داده‌های مقطعی مقایسه می‌شوند. در بخش دوم پس از معرفی داده‌های گم‌شده، انواع الگوهای گم‌شدن (انصراف و عام) در داده‌های طولی و همچنین مدل‌بندی هم‌زمان پاسخ و انصراف بیان می‌شود. مدل‌های مختلف برای تحلیل داده‌های طولی، در بخش سوم ارائه می‌شود.

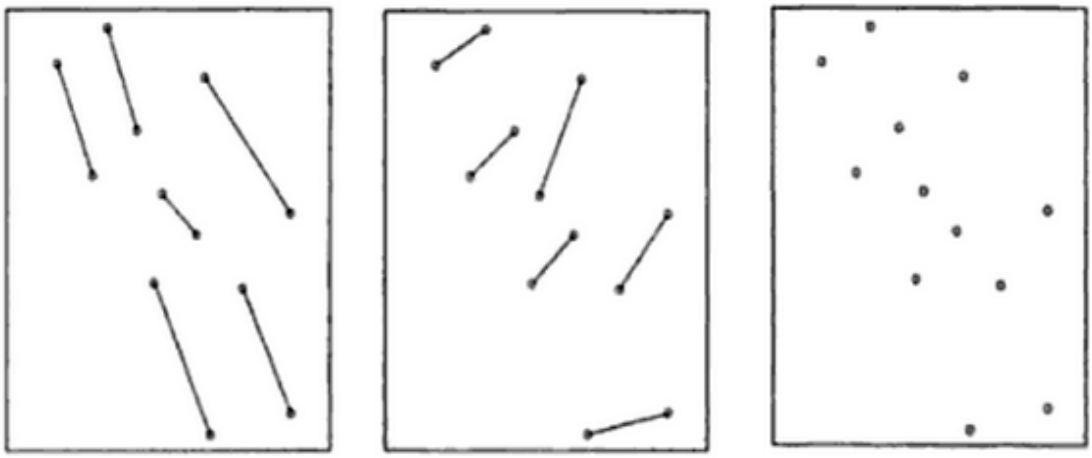
۱.۱ معرفی داده‌های طولی و مقایسه‌ی آن‌ها با داده‌های مقطعی

هدف اصلی مطالعه‌های طولی، مشخص‌سازی تغییرات متغیّر پاسخ در طول زمان و بررسی عامل‌های تأثیرگذار بر این تغییرات است. مهم‌ترین مشخصه‌ی مطالعه‌های طولی اندازه‌گیری تکراری واحدها در طول زمان است. به این معنا که برای یک واحد چندین پاسخ (مانند فشار خون و میزان کبد چرب) در طول زمان ثبت می‌گردد. عموماً پاسخ‌های هر فرد دارای همبستگی مثبتی هستند. همچنین با استفاده از اندازه‌گیری‌های تکراری بر

روی هر واحد، دست‌یابی به تغییرات درون هر واحد نیز ممکن می‌شود. امکان بررسی تغییرات پاسخ با متغیّرهای تبیینی در طول زمان و همچنین امکان انجام استنباط آماری در پدیده‌های نادری که حجم نمونه کوچک است از مهم‌ترین مزایای مطالعه‌های طولی هستند. در مقابل مطالعه‌های طولی، مطالعه‌های مقطعی قرار دارد که در آن‌ها پاسخ تنها در یک زمان اندازه‌گیری می‌شود و تنها می‌توان تفاوت‌های میان واحدهای نمونه (یا خوشه، در صورت وجود) در همان زمان را بررسی کرد. برای مثال مطالعه‌ی مقطعی بر روی خوشه‌هایی متشکل از دانشجویان یک کلاس، می‌تواند به بررسی تفاوت‌ها میان کلاس‌های مختلف بپردازد، اما هیچ اطلاعی را در مورد تغییرات ویژگی‌های مورد بررسی برای افراد داخل خوشه (دانشجویان کلاس)، مهیا نمی‌کند.

مثال ۱.۱.۱. در شکل ۱.۱.۱ (دیگل و همکاران، ۲۰۰۲) در سه قسمت، توانایی خواندن (محور عمودی) در مقابل سن (محور افقی) رسم شده است. اگر قسمت (الف) را برای یک مطالعه‌ی مقطعی در کودکان به منظور بررسی تأثیر سن بر توانایی خواندن در نظر بگیریم، به‌نظر می‌رسد که توانایی خواندن در کودکان با افزایش سن آن‌ها کاهش می‌یابد. در قسمت (ب) این شکل، فرض کنید که داده‌های مشابه از یک مطالعه‌ی طولی جمع‌آوری شده باشند که در آن توانایی خواندن هر کودک دو بار اندازه‌گیری شده است. حال با وجود این‌که کودکان خردسال‌تر در ابتدا دارای سطح خواندن بالاتری هستند اما واضح است که توانایی خواندن هر شخص در طول زمان افزایش می‌یابد. اگر مجموعه داده‌ها همانند قسمت (ج) شکل ۱.۱.۱ باشند آن‌گاه می‌توان تفسیری متفاوت از مشاهده‌ها بدست آورد. اکنون در قسمت‌های (الف) و (ج)، مطالعه‌های مقطعی و طولی یک موضوع غیرمعمول را بیان می‌کنند و آن این است که توانایی خواندن با افزایش سن کاهش می‌یابد!. نکته این مثال این است که مطالعه‌های طولی [شکل‌های (ب) و (ج)] می‌توانند تغییرهای در طول زمان در هر آزمودنی (تأثیر سن) را مشخص سازند، در حالی‌که در مطالعه‌های مقطعی این کار امکان‌پذیر نیست.

هر چند می‌توان داده‌های طولی را به دو صورت آینده‌نگر (با دنبال‌کردن آزمودنی‌ها در طول زمان) یا



(ج) (ب) (الف)

شکل ۱.۱.۱: داده‌های فرضی در مورد رابطه‌ی بین سن (محور افقی) و توانایی خواندن (محور عمودی) (دیگل و همکاران، ۲۰۰۲، ص ۲)

گذشته‌نگر (با استخراج اندازه‌گیری‌های مختلف هر آزمودنی از داده‌های ثبت‌شده در گذشته) جمع‌آوری کرد، اما داده‌های طولی به‌طور معمول به‌صورت آینده‌نگر جمع‌آوری می‌شوند. زیرا کیفیت اندازه‌گیری‌های مکرر که از اطلاعات ثبت‌شده آزمودنی در گذشته بدست می‌آید ممکن است پایین باشد. چون مجموعه مشاهده‌های یک آزمودنی دارای نوعی درون‌همبستگی هستند باید این همبستگی را به‌منظور استخراج استنباط‌های آماری معتبر در نظر گرفت. بنابراین تحلیل داده‌های طولی نیازمند روش‌های آماری مخصوص به‌خود است. مسئله در نظر گرفتن همبستگی بین پاسخ‌ها در طول زمان در تحلیل سری‌های زمانی نیز وجود دارد، با این تفاوت که در مطالعه‌های طولی برای تعداد زیادی آزمودنی، در طول دوره‌ی زمانی به‌طور معمول کوتاه مدت، چندین پاسخ اندازه‌گیری می‌شود، در حالی که در سری‌های زمانی تنها پاسخ برای یک آزمودنی در طول دوره‌ی زمانی بلندمدت در نظر گرفته می‌شود. دیگل (۱۹۹۰) تحلیل سری‌های زمانی را در علوم زیستی مورد بحث قرار داد. به دلیل آن‌که می‌توان در داده‌های طولی آزمودنی‌ها را مستقل در نظر گرفت تحلیل این‌گونه داده‌ها ساده‌تر از تحلیل سری‌های زمانی به‌نظر می‌رسد. در تحلیل داده‌های طولی سازگاری یک الگو در میان همه آزمودنی‌ها، به عنوان پایه و اساسی برای نتایج به‌شمار می‌رود. به همین دلیل استنباط‌های بدست آمده از

مطالعه‌های طولی نسبت به تغییر در فرض‌های مدل، به‌ویژه نسبت به تغییر در فرض‌هایی در مورد ساختار همبستگی، استوارتر از آن‌ها در تحلیل سری‌های زمانی هستند (دیگل و همکاران، ۲۰۰۲).

به‌طور خلاصه می‌توان گفت که تحلیل داده‌های طولی نمایشی از ادغام دو موضوع تحلیل رگرسیونی و تحلیل سری‌های زمانی را ارائه می‌کند. همانند بسیاری از مجموعه داده‌های رگرسیونی، داده‌های طولی از یک نمونه از آزمودنی‌ها بدست آمده است. بر خلاف داده‌های رگرسیونی، در داده‌های طولی می‌توان پاسخ هر آزمودنی را در طول زمان مشاهده کرد.

تفاوت مطالعه‌های طولی با اندازه‌گیری‌های مکرر این است که در اندازه‌گیری‌های مکرر فاصله زمانی بین اندازه‌گیری‌ها ممکن است مورد توجه نباشد در حالی‌که این موضوع در مطالعه‌های طولی مهم است.

در جامعه‌شناسی و اقتصاد اغلب مطالعه‌های طولی را مطالعه‌های پانلی می‌نامند. در داده‌های پانلی، یک "پانل" گروهی از افراد است که پاسخ‌های مرتبط با آن‌ها در طول زمان به‌صورت مکرر جمع‌آوری شده است.

در این رساله داده‌های طولی با پاسخ‌های گسسته ترتیبی مدنظر قرار گرفته شده است.

۲.۱ معیار پیوند گاما برای داده‌های گسسته

برای اندازه‌گیری پیوند در داده‌های طولی با پاسخ ترتیبی می‌توان از معیارهای گاما و τ کندال استفاده کرد، اما چون معیار گاما از لحاظ محاسباتی ساده‌تر است (اگرستی، ۱۹۸۴) در این رساله از معیار گاما استفاده می‌شود که توسط گودمن و کروسکال (۱۹۵۴) مطرح گردید.

با توجه به این‌که معیار گاما تابعی از جفت‌های هماهنگ و ناهماهنگ است، ابتدا این دو مفهوم تعریف می‌شوند. اگر ترتیب رده‌های پاسخ دو آزمودنی یا یک آزمودنی در دو زمان مختلف روی دو متغیّر ترتیبی X و Y را در نظر گیریم، یک جفت آزمودنی را هماهنگ می‌نامند هرگاه با افزایش X، Y افزایش و با کاهش X، Y کاهش یابد و یک جفت آزمودنی را ناهماهنگ می‌نامند هرگاه با افزایش X، Y کاهش و با کاهش X، Y افزایش یابد. همچنین یک جفت آزمودنی را گره می‌نامند هرگاه حداقل یکی از دو متغیّر X و Y ثابت

و دیگری به‌طور یکنوا تغییر کند. اگر تعداد جفت‌های هماهنگ را با C و تعداد جفت‌های ناهماهنگ را با D نمایش دهیم، آن‌گاه اگر $C - D > ۰$، پیوند را مثبت و اگر $C - D < ۰$، پیوند را منفی می‌نامند.

اگر در یک جدول پیشایندی $I \times J$، π_{ij} احتمال توأم برای سطر iاُم و ستون jاُم باشد، آن‌گاه برای دو مشاهده‌ی مستقل از این جدول احتمال هماهنگ‌بودن و ناهماهنگ‌بودن یک جفت که به ترتیب با Π_C و Π_D نمایش داده می‌شوند برابر است با

$$\Pi_C = ۲ \sum_i \sum_j \pi_{ij} \left(\sum_{h>i} \sum_{t>j} \pi_{ht} \right)$$

و

$$\Pi_D = ۲ \sum_i \sum_j \pi_{ij} \left(\sum_{h>i} \sum_{t<j} \pi_{ht} \right).$$

حال با توجه به Π_C و Π_D، معیار گاما برابر است با

$$\gamma = \frac{\Pi_C - \Pi_D}{\Pi_C + \Pi_D}.$$

معیار گاما در بازه‌ی $[-۱, ۱]$ مقدار می‌گیرد و با نسبت اختلاف بین جفت‌های هماهنگ و ناهماهنگ بر مجموع جفت‌های هماهنگ و ناهماهنگ، یعنی $\hat{\gamma} = \frac{C-D}{C+D}$، برآورد می‌شود. گاما زمانی مقدار ۱ را می‌گیرد که $\Pi_D = ۰$ و زمانی مقدار -۱ را می‌گیرد که $\Pi_C = ۰$. یک بودن مقدار قدرمطلق گاما، $(|\gamma| = ۱)$، به این معناست که ارتباط بین متغیّرهای X و Y یکنوا است اما لزوماً اکیداً یکنوا نیست (اگرستی، ۲۰۰۲، فصل ۲).

اگر علاوه بر متغیّرهای ترتیبی، یک متغیّر کنترل نیز وجود داشته باشد، آن‌گاه در هر سطح متغیّر کنترل می‌توان معیار گاما را برای اندازه‌گیری پیوند بین دو پاسخ ترتیبی مورد استفاده قرار داد. برای انجام این‌کار فرض کنید که متغیّر کنترل با Z نمایش داده شود و π_{ijk} نشان‌دهنده توزیع احتمال توأم در یک جدول سه بعدی $I \times J \times K$ باشد که در آن K نشان‌دهنده‌ی تعداد سطوح متغیّر کنترل است. برای سطح ثابت k از متغیّر کنترل Z، کافی است به جای π_{ij} در رابطه‌های Π_C و Π_D از π_{ijk} استفاده شود و سپس مقدار گاما را محاسبه کرد. این حالت از معیار گاما را گامای جزیی می‌نامند. به عبارت دیگر گامای جزیی برای بررسی

اینکه آیا دو متغیّر ترتیبی مورد نظر به شرط هر سطح متغیّر سوم (متغیّر کنترل) دارای پیوند هستند یا نه، استفاده می‌شود.

۳.۱ مدل‌های مختلف برای تحلیل داده‌های طولی

در این بخش مدل‌های مختلف برای تحلیل داده‌های طولی، به طور خلاصه معرفی خواهند شد، سپس در فصل ۲ این مدل‌ها بیشتر مورد بررسی قرار خواهند گرفت. در مطالعه‌های طولی یك دنباله از پاسخ‌ها برای هر آزمودنی وجود دارد، بنابراین باید همبستگی‌های بین داده‌های مرتبط با هر آزمودنی در نظر گرفته شود. برای درنظر گرفتن همبستگی بین پاسخ‌ها، مدل‌های مختلفی را می‌توان به‌کار برد. فهرمیر و تاتز (۱۹۹۴) و دیگل و همکاران (۲۰۰۲) این مدل‌ها را در سه رده مجزا در نظر گرفته‌اند که عبارتند از؛ مدل‌بندی حاشیه‌ای، مدل‌بندی اثرهای تصادفی و مدل‌بندی مارکوف. علیرغم این‌که این سه مدل، مدل‌هایی مهم برای اندازه‌گیری‌های تکراری هستند اما نویسندگان متعددی همچون دیگل و کنوارد (۱۹۹۴)، دیگل و همکاران (۲۰۰۲) و آئرتس و همکاران (۲۰۰۲) بین این مدل‌ها تفاوت قائل شده‌اند. این مدل‌ها می‌توانند برای پاسخ‌های با توزیع نرمال و نیز برای پاسخ‌های با توزیع غیرنرمال به‌کار روند. در حالت پاسخ‌های با توزیع نرمال، بین این سه مدل به‌خاطر ویژگی‌های خاص توزیع نرمال چند متغیّره می‌توان ارتباط برقرار کرد، اما هیچ ارتباط نزدیکی بین این مدل‌ها در حالتی که پاسخ‌ها غیرنرمال (دودویی یا رسته‌ای) هستند برقرار نمی‌باشد، (مولنبرگز و وربك، ۲۰۰۶). پاسخ به این مهم که کدام مدل باید انتخاب شود به‌طور عمده به سؤالات تحقیق بستگی دارد.

در مدل‌های انتقالی (مارکوف)، میانگین پاسخ در یك موقعیت یا زمان بر حسب پاسخ‌ها در موقعیت‌ها یا زمان‌های قبلی مدل‌بندی می‌شوند. در واقع این مدل‌بندی، همبستگی‌های دو یا چند متغیّره میان مشاهده‌ها در موقعیت‌ها یا زمان‌های متوالی را در نظر می‌گیرد.

مشهورترین مدل انتقالی، مدل انتقالی مرتبه اول است که در آن فرض می‌شود پاسخ Y_{it} تنها به پاسخ