بسم الله
الرحمن الرحیم

تاییدیه اعضای هیات داوران حاضر در جلسه دفاع از رساله دکتری

آقای محمد غلامی فشارکی رشته آمارزیستی رساله دکتری خود را با عنوان « تحلیل بیزی چند سطحی با استفاده از توزیع t چوله و کاربرد آن در داده های پزشکی » در تاریخ ۱۳۹۲/۳/۸ ارائه کردند.

بدینوسیله اعضای هیات داوران نسخه نهایی این رساله را از نظر فرم و محتوا تایید کرده و پذیرش آنرا برای تکمیل درجه دکتری پیشنهاد می‌کنند.

اعضای هیات داوران	نام و نام خانوادگی	امضاء
استاد راهنما	دکتر انوشیروان کاظم نژاد	
استاد مشاور	دکتر فرید زایری	
استاد ناظر	دکتر فاطمه الحانی	
استاد ناظر	دکتر یداله محرابی	
استاد ناظر	دکتر محمود رضا گوهری	
استاد ناظر و نماینده تحصیلات تکمیلی	دکتر ابراهیم حاجی زاده	

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 ارایه در مجامع علمی باید به نام دانشگاه بوده و با تایید استاد راهنمای اصلی، یکی از اساتید راهنما، مشاور و یا دانشجوی 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 به صورت ترکیبی از اطلاعات جدید و نتایج حاصل از پایان‌نامه/ رساله نیز منتشر می‌شود نیز باید نام دانشگاه درج شود.

ماده ۳- انتشار کتاب و یا 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 آئین‌نامه‌های مصوب انجام شود.

ماده ٤- ثبت اختراع و تدوین دانش فنی و یا ارایه 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 ٥- این آیین‌نامه در ۵ ماده و یک تبصره در تاریخ ۸۷/۴/۱ 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 **محمد غلامی فشارکی** دانشجوی رشته **آمار زیستی** ورودی سال تحصیلی **۱۳۸۸** مقطع **دکتری** دانشکده **علوم پزشکی** متعهد می شوم کلیه نکات مندرج در آیین نامه حق مالکیت مادی و معنوی در مورد نتایج پژوهش های علمی دانشگاه تربیت مدرس را در انتشار یافته های علمی مستخرج از پایان نامه / رساله تحصیلی خود رعایت نمایم. در صورت تخلف از مفاد آیین نامه فوق الاشعار به دانشگاه وکالت و نمایندگی می دهم که از طرف اینجانب نسبت به لغو امتیاز اختراع بنام بنده و یا هرگونه امتیاز دیگر و تغییر آن به نام دانشگاه اقدام نماید. ضمناً نسبت به جبران فوری ضرر و زیان حاصله براساس برآورد دانشگاه اقدام خواهم نمود و بدینوسیله حق هرگونه اعتراض را از خود سلب نمودم».

امضاء

تاریخ ۹۲/۳/۲۲

آئین‌نامه پایان‌نامه (رساله)های دانشجویان دانشگاه تربیت مدرس

نظر به اینکه چاپ و انتشار پایان‌نامه (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 پایان‌نامه (رساله)ی خود، مراتب را قبلاً به‌طور کتبی به دفتر "دفتر نشر آثار علمی" 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 زیستی** است که در سال **۱۳۹۱** در دانشکده علوم پزشکی دانشگاه تربیت مدرس به راهنمایی **آقای دکترانوشیروان کاظم نژاد** ، مشاوره **آقای دکتر فرید زایری** از آن دفاع شده است.

ماده ۳ : به منظور جبران بخشی از هزینه‌های انتشارات دانشگاه، تعداد یک درصد شمارگان کتاب (در هر نوبت چاپ) را به "دفتر نشر آثار علمی" دانشگاه اهداء کند. دانشگاه می‌تواند مازاد نیاز خود را به نفع مرکز نشر در معرض فروش قرار دهد.

ماده ۴ : در صورت عدم رعایت ماده ۳، ۵۰٪ بهای شمارگان چاپ شده را به‌عنوان خسارت به دانشگاه تربیت مدرس، تادیه کند.

ماده ۵ : دانشجو تعهد و قبول می‌کند در صورت خودداری از پرداخت‌های بهای خسارت، دانشگاه 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 **محمد غلامی فشارکی** دانشجوی رشته **آمار زیستی** مقطع **دکتری** تعهد فوق و ضمانت اجرایی آن را قبول کرده، به آن ملتزم می‌شوم.

نام و نام خانوادگی

تاریخ و امضا

۹۲/۳/۲۲

دانشگاه تربیت مدرس
دانشکده علوم پزشکی

رساله

دوره دکتری تخصصی (Ph.D.) در رشته آمار زیستی

عنوان

تحلیل بیزی چندسطحی با استفاده از توزیع t چوله و کاربرد آن در داده های پزشکی

نگارش

محمد غلامی فشارکی

استاد راهنما

دکتر انوشیروان کاظم نژاد

استاد مشاور

دکتر فرید زایری

خرداد 1392

تقدیم به :

ایثار مادر مرحومم

و گذشت همسر فداکارم

و

مهربانی پدرم

و به خواهر و برادرانم.

تشکر و قدردانی

از جناب آقای دکتر انوشیروان کاظم نژاد (استاد راهنما) و آقای دکتر فرید زایری (استاد مشاور) به خاطر راهنمایی های ارزنده و مفیدشان در انجام این پایان نامه کمال تشکر و قدردانی را می نمایم.

همچنین از اساتید گروه آمار زیستی جناب آقای دکتر ابراهیم حاجی‌زاده و جناب آقایان دکتر سقـراط فقیه‌زاده که از محضرشان بهره‌ی علمی و اخلاقی برده‌ام تشکر و قدردانی نمایم که سپاس از آن‌ها همانا سپاس از خالق است.

همچنین در اینجا جا دارد که از خانواده خود و همسرم که مانند همیشه؛ مهربانانه و بی‌منت پشتیبان و مشوق من در طول تحصیل بوده‌اند و علی الخصوص از همسر صبور، مهربان و فداکارم که همواره در زندگی مشترکم احساس رضایتی دو چندان را نموده اند کمال تشکر و قدردانی را می نمایم.

و

سپاس از همه‌ی شما عزیزان

چکیده

آنالیز چند سطحی(آشیانه ای) روشی کارا برای تجزیه و تحلیل داده های اجتماعی و پزشکی است که در بیش از یک سطح قرار گرفته اند، و در واقع حالت بسط داده شده از مدل های خطی تعمیم یافته می باشد که در آن علاوه بر مدل بندی متغیر پاسخ ضرایب رگرسیونی نیز مدل بندی می شوند

یکی از کاربردهای این روش در تحلیل داده های طولی با ساختار چندسطحی می باشد. متداول‌ترین فرض‌ها دربکارگیری تحلیل چند سطحی، براساس نرمال بودن خطا و اثرات تصادفی سطوح بنا شده است که گاهی به دلیل مشاهدات دور افتاده و یا فرم نامتقارن توزیعی، برآورد درستی از ضرایب را در اختیار محقق قرار نمی‌دهد.

از این رو، پژوهش حاضر با هدف بررسی جنبه‌های نظری مدل‌های بیزی چند سطحی با اثرات تصادفی غیر نرمال انجام پذیرفته است. از طرف دیگر در این رساله با استفاده از مطالعات شبیه سازی صحت تحلیل بیزی چند سطحی با توزیع غیرنرمال مورد بررسی قرار گرفته است.

در پایان تحلیل بیزی چندسطحی با استفاده از توزیع های غیر متقارن بر روی داده های گذشته نگر که هدف آن بررسی رابطه نوبت کاری با کلسترول خون با کنترل متغیرهای مخدوشگر بود برازش داده شده که نتایج این برازش حاکی از عدم مشاهده رابطه معنی دار بین نوبت کاری و کلسترول خون بوده است.

کلمات کلیدی

تحلیل چندسطحی، مدل آشیانه ای، داده های طولی، تحلیل بیزی، الگوریتم مونت کارلوی زنجیر مارکفی، کلسترول، توزیع های نرمال چوله مستقل

فهرست مطالب

فصل اول : مقدمه و مروری بر مطالعات گذشته 1

1-1 موضوع تحقیق 2

1-2 پیشینه نظری 12

1-3 ضرورت مطالعه از دیدگاه پزشکی 16

1-4 اهداف تحقیق 18

1-5 جنبه نوآوری رساله 18

فصل دوم: مواد و روش ها 19

2-1 مروری بر فرآیندهای تولید توزیع های چوله 20

2-1-1 توزیع نرمال چوله 30

2-1-2 توزیع تی چوله 33

2-2 تحلیل سه سطحی 35

2-2-1 محاسبه تابع چگالی به روش توزیع توام 37

2-2-2 محاسبه تابع چگالی به روش توزیع شرطی 38

2-2-3 روش های برازش مدل 41

2-2-4 برآورد اثرات تصادفی مدل 45

2-2-5 روش های آزمون پارامترها و شاخص های کفایت مدل 45

2-3 تحلیل بیزی 46

2-3-1 مزایای تحلیل بیزی 47

2-3-2 معایب تحلیل بیزی 50

2-3-3 استنباط بیزی 51

2-3-4 انتگرالگیری مونت کارلو، الگوریتم متروپولیس-هستینگز و نمونه گیری گیبس 52

2-3-5 شاخص انتخاب مدل 54

فصل سوم :تحلیل بیزی چندسطحی با استفاده از توزیع تی چوله 56

3-1 تحلیل بیزی چندسطحی با استفاده از خانواده توزیع های چوله 57

3-2 مطالعات شبیه سازی 65

3-2-1 مطالعه شبیه سازی اول 65

3-2-2 مطالعه شبیه سازی دوم 71

3-1 روش جمع آوری داده های پزشکی 72

3-1-1 نوع مطالعه، نحوه نمونه گیری 72

2-1-3 شرایط ورود و خروج از مطالعه 73
3-1-3 ملاحظه‌های اخلاقی 73
4-1-3 نحوهٔ تعریف متغیر نوبت کاری 74
5-1-3 محاسبات آماری 74
فصل چهارم : نتایج 76
1-4 نتایج حاصل از مطالعه شبیه سازی اول 77
2-4 نتایج حاصل از مطالعه شبیه سازی دوم 77
3-4 نتایج حاصل از داده های واقعی 84
1-3-4 نتایج توصیفی داده ها 84
2-3-4 تحلیل داده ها 86
فصل پنجم :بحث، نتیجه‌گیری و پیشنهادها 93
1-5 بحث از دیدگاه آماری 94
2-5 بحث از دیدگاه پزشکی 97
3-5 پیشنهاد ها 99
فهرست منابع 101
ضمائم 109
چکیده انگلیسی 112

فهرست جداول

جدول 2-1 : برخی از توزیع های چوله معروف براساس نوع انتخاب g و H (معادله 2-1) 28
جدول 2-2 : برخی از توزیع های چوله معروف براساس نوع انتخاب $\boldsymbol{U}$ (معادله 2-11)..... 29
جدول 2-3: تابع برخی از توزیع های چوله معروف دیگر 29
جدول 2-4: دامنه پوشش چولگی و کشیدگی برخی از توزیع های چوله معروف 30
جدول 3-1 : توزیع های پیشین برای ضرایب بتا، پارامتر چولگی و پراکندگی و پارامترهای دیگر مدل 61
جدول 3-2 : چهار استراتژی برای مولفه تصادفی $u(j)$ و $\epsilon i(j)$ 72
جدول 4-1: پارامترها و میزان خطای آن برای سه روش برآورد با حجم نمونه های متفاوت با فرض میانگین 3..... 79
جدول 4-2 : پارامترها و میزان خطای آن برای سه روش برآورد با حجم نمونه های متفاوت با فرض میانگین 3-..... 80
جدول 4-3: مقادیر پارامترهای محاسبه شده و میزان خطای آن برای مطالعه شبیه سازی دوم 82
جدول 4-4 : توزیع فراوانی وضعیت سیگار، فشار خون بالا، تاهل و تحصیلات به تفکیک الگوی کاری 85
جدول 4-5 : متغیرهای دموگرافیک افراد شرکت کننده در مطالعه به تفکیک الگوی کاری در اولین زمان مشاهده..... 85
جدول 4-6 : میزان میانگین، میانه، انحراف معیار ، چولگی و کشیدگی سه مولفه تصادفی خطا، فرد و ناحیه 90
جدول 4-7 : خلاصه اطلاعات DIC برای مدل های برازش داده شده 90
جدول 4-8 : خلاصه اطلاعات حاصل از برازش مدل شماره 4..... 91

فهرست اشکال

شکل 2-1: توزیع نرمال چوله با پارامترهای مختلف 31
شکل 2-2 : توزیع تی چوله و مقایسه آن با توزیع نرمال چوله متناظر با آن 34
شکل 2-3 : الگوی شماتیک ساختار داده های آشیانه ای با ساختار چندسطحی 36
شکل 2-4 : میزان ICC و ارتباط آن با انتخاب مدل 46
شکل 4-1 : نمودار متوسط خطا در برآورد پارامترهای توزیع نرمال چوله با استفاده از سه روش بیزی مستقیم، غیرمستقیم و درستنمایی براساس حجم نمونه های مختلف.....81
شکل 4-2 : هیستوگرام میانگین کلسترول مشاهدات برای نواحی مختلف کارخانه فولاد مبارکه 87
شکل 4-3 : هیستوگرام میانگین کلسترول مشاهدات برای مشاهدات برای هر فرد.....88
شکل 4-4 : هیستوگرام میانگین کلسترول مشاهدات برای مشاهدات برای کلیه مشاهدات 88
شکل 4-5 : نمودار چگالی پسین برای پارامترهای برآورد شده مدل شماره 4 92

فصل اول : مقدمه و مروری بر مطالعات گذشته

1-1 موضوع تحقیق

بسیاری از داده ها، از جمله داده های برخی از مشاهدات مربوط به انسان و علوم بیولوژیکی، دارای ساختار خوشه ای یا سلسله مراتبی هستند. به طور مثال، فرزندان یک خانواده از لحاظ مشخصات فیزیکی و روحی، دارای شباهت بیشتری نسبت به دیگر افراد جامعه می باشند. از جنبه آماری شباهت افراد به یکدیگر موید مستقل نبودن داده ها است. دونفری که متعلق به یک پدر و مادر هستند از لحاظ بسیاری از شرایط مانند یکدیگرند [1]. برای مثال اگر فرزند یک خانواده دارای بیماری خاصی مانند هپاتیت باشد امکان داشتن هپاتیت در برادر و خواهر وی افزایش می یابد و یا در مطالعات اندازه های مکرر شده بر روی متغیر های اثرگذار بر فشار خون، فشار خون نمونه ها در زمان های مختلف اندازه گرفته می شود. مسلما فشار خون هر فرد به خاطر وجود شرایط ذاتیش دارای تغییرات مخصوص به خود در طول زمان می‌باشد. همبستگی هایی از این نوع در طرح های مقطعی در ساختارهای آشیانه ای دیده می شوند. آنالیز چند سطحی معمولا برای تجزیه و تحلیل داده های اجتماعی و زیستی که در بیش از یک سطح قرار گرفته اند، به کار می رود، سطح جزیی از داده های

آشیانه ای بوده و توسط نحوه جمع آوری داده ها مشخص می شود. در واقع سطح، مقطع جمع آوری داده ها است [1-3]. تحلیل چندسطحی در واقع حالت بسط داده شده از مدل های خطی تعمیم یافته است که در آن علاوه بر مدل بندی متغیر پاسخ ضرایب رگرسیونی نیز مدل بندی می گردد. این روش، روشی کارا در مدل سازی داده هایی با ساختار آشیانه ای بوده و هدف آن مدل بندی متغیر وابسته براساس تابعی از متغیرهای پیشگو (مستقل) در بیشتر از یک سطح می باشد [2]. آنالیز چندسطحی با اسامی دیگری نظیر مدل‌های خطی آشیانه ای[1]، مدل های ضرایب تصادفی[2] و مدل های اثرات تصادفی[3] نیز شناخته می شود[3].

ویژگی اصلی داده های چند سطحی خصوصیت گروه بندی آنهاست. معمولا گروه های مورد مطالعه به صورت تصادفی انتخاب می شوند و از این رو علاوه بر خطای ناشی از اندازه گیری مشاهدات درون هر گروه، خطای دیگری مربوط به نمونه گیری از گروه ها نیز در تحلیل داده های چندسطحی دخالت دارد. روشهای سنتی مدل‌بندی رگرسیونی این دومین خطا را نادیده می گیرند. در نظر نگرفتن همبستگی بین مشاهدات منجر به کم برآوردی خطای برآورد ضرایب رگرسیونی و معنی داری نادرست ضرایب و به تناسب آن افزایش خطای نوع اول می گردد[4]. علاوه بر این می توان به عدم امکان تعمیم نتایج راجع به گروه بندی به کل گروه و عدم امکان کشف تغییرپذیری منتسب به گروه به عنوان معایب دیگر مدل های مرسوم رگرسیونی اشاره کرد. در عوض مدل های تحلیل چند سطحی این مشکلات را برطرف می نمایند. به اضافه این که با استفاده از مدل‌های تحلیل چند سطحی امکان برآورد تعداد زیاد پارامترها و تقلیل خطای اندازه گیری نیز وجود دارد.

1- Hierarchical Linear Models
2- Random Coefficients Models
3- Random Effects Models

یکی دیگر از کاربردهای تحلیل چند سطحی استفاده از این روش در تحلیل داده های طولی با ساختار آشیانه‌ای می باشد [1, 2]. برای مثال فرض کنید محققی می خواهد رابطه بین تعداد ساعات کار بر فشار خون کارگران شاغل در یک کارخانه فولاد سازی را وقتی که مشاهدات به صورت طولی آشیانه ای جمع آوری شده است اندازه گیری نماید. از آنجا که افراد در زمانهای مختلفی اندازه گیری شده اند، از این رو داده های این مطالعه طولی و به یکدیگر وابسته هستند. از طرف دیگر چون فشار خون به عوامل محیطی چون دما، نوع فعالیت، استرس و سایر عوامل محیطی وابسته است، از این رو فشار خون افرادی که متعلق به یک ناحیه کاری مثلا فولاد گرم هستند بخاطر شرایط کاری یکسان به یکدیگر وابسته اند. در اینجا **مجموعه اندازه های گرفته شده در زمان های مختلف برای هر یک از اعضای نمونه، به عنوان واحدهای سطح اول، هر یک از افراد انتخاب شده به عنوان سطح دوم و تعلق هریک از اعضای نمونه با توجه به ساختار آشیانه ای داده ها (در این مثال محیط کاری) به عنوان سطح سوم در نظر گرفته می‌شود.**

البته در متون آماری بیشتر از سه روش کلی، مدل های حاشیه‌ای[1]، مدل های اثرات تصادفی و یا آمیخته[2] و مدل‌های انتقال[3] به عنوان سه روش متداول در تحلیل داده های طولی سخن به میان آمده است. به طور خلاصه در مدل‌های حاشیه ای تاثیر همبستگی داده ها در طول زمان با استفاده از ماتریس ساختار همبستگی در مدل سازی حذف شده و مدل‌بندی بر حسب میانگین پاسخ، فقط به متغیرهای کمکی وابسته است[5]. این در حالی است که در مدل‌های آماری اثرات تصادفی، عموماً فرض بر آن است که برخی ضرایب رگرسیونی برای آزمودنی‌های مختلف به دلایل متفاوت، به خود آزمودنی وابسته است و نیز

1- Marginal Models
2- Mixed Effect Models
3-Transition Models

دارای توزیعی جداگانه هستند. به عبارت دیگر، همبستگی میان مشاهدات در داده‌های تکراری یا داده‌های موجود در یک خوشه را می‌توان ناشی از وجود یک اثر تصادفی یکسان در آن‌ها دانست، به طوری‌که در مورد یک فرد یا یک خوشه این اثر تصادفی یکسان و در فرد یا خوشه‌ی دیگر این تاثیر متفاوت بوده و در حقیقت این ویژگی باعث ایجاد همبستگی می‌گردد [6]. همچنین در مدل‌های انتقال که در تحلیل داده‌های طولی کاربرد دارند، فرض بر آن است که متغیر پاسخ در زمان اندازه گیری به متغیرهای پاسخ در زمان‌های قبلی وابسته است. این مدل‌سازی توسط اصول حاکم بر زنجیرهای مارکف[1] انجام می‌گیرد و همبستگی‌ها از طریق ارتباط با پاسخ‌های قبلی مورد بررسی قرار می‌گیرند [7].

باید گفت هیچ یک از دو روش مدل سازی حاشیه ای و انتقالی توانایی تحلیل داده های طولی با ساختار آشیانه ای را دارا نبوده و تنها روش در تحلیل چنین داده هایی استفاده از روش مدل سازی اثرات تصادفی است. از این منظر می توان تحلیل چند سطحی را به نوعی زیرمجموعه مدل سازی اثرات تصادفی دانست. در واقع نحوه برآورد ضرایب مدل چند سطحی زیر مجموعه ای از روش برآورد مدل های اثرات تصادفی یا آمیخته می باشد. به همین علت تمام نرم افزارهای برازش دهنده مدل های آمیخته توانایی برازش مدل بندی چند سطحی را دارا هستند. البته روش تحلیل چندسطحی دارای مزایایی نسبت به مدل های تصادفی است که از آن جمله می توان به ملموس بودن روش انجام تحلیل برای محققان، امکان همگرایی و برآورد پذیری و نیز حضور محیط به عنوان یک پارامتر تعیین کننده در مدل سازی و بالاخره استفاده از یک الگوریتم قدم به قدم برای جلوگیری از ورود متغیرهای مزاحم (ورود متغیر به ترتیب سطوح آشیانه ای آن در مدل سازی) اشاره نمود [3].

1- Markov Chain

در حالت ساده فرض کنید هدف بررسی که می خواهیم اثر یک متغیر مستقل مانند نوبت کاری (X) بر کلسترول خون (Y)، وقتی که داده ها به صورت طولی آشیانه ای جمع آوری شده اند، در حضور متغیر کمکی سطح 1 مانند درجه حرارت محیط (W) باشد. در این حالت معادله برازش به فرم زیر نمایش داده می شود.

(1-1)

$Level\ 1$:

$$Y_{ijk} = \beta_{0ij} + \beta_{1ij}X_{ijk} + \varepsilon_{ijk}$$

$Level\ 2$:

$$\beta_{0ij} = \alpha_{00j} + \alpha_{01j}W_{ij} + u_{0ij}$$
$$\beta_{1ij} = \alpha_{10j} + \alpha_{11j}W_{ij} + u_{1ij}$$

$Level\ 3$:

$$\alpha_{00j} = \gamma_{00} + v_{00j}$$
$$\alpha_{01j} = \gamma_{01} + v_{01j}$$
$$\alpha_{10j} = \gamma_{10} + v_{10j}$$
$$\alpha_{11j} = \gamma_{11} + v_{11j}$$

در معادله بالا، k : نشان دهنده ناحیه k ام، j : نشان دهنده فرد j ام در ناحیه k ام و i : نشان دهنده مشاهده i ام برای فرد j ام در ناحیه k ام می باشد.

فرضیات معادله 1-1 به فرم معادله 2-1 می باشد.

(2-1)

$$\epsilon \sim N(0, \sigma^2) \qquad u = \begin{bmatrix} u_0 \\ u_1 \end{bmatrix} \sim N(0, \Omega_u) \qquad v = \begin{bmatrix} v_{00} \\ v_{01} \\ v_{10} \\ v_{11} \end{bmatrix} \sim N(0, \Omega_v)$$
$$Cov(\epsilon, u) = Cov(\epsilon, v) = Cov(u, v) = 0$$

ارایه مدل به فرم معادلات (1-1) علاوه بر بیان همه متغیرهای مستقل و وابسته، نشان دادن ماهیت چندسطحی مدل می باشد. همان گونه که در این معادله مشخص است سطح یک مدل ($Level1$)

شبیه نوعی از مدل رگرسیون چندگانه و قسمت دوم آن (*Level2*) نشان دهنده نحوه ارتباط ضرایب رگرسیونی با متغیرهای سطح دوی مدل می باشد و این بدان معنا است که اینکه شیب ها و عرض از مبداها خود تابعی از متغیر سطح دوم یعنی W ها می باشند. همان گونه که در این مدل مشخص است علاوه بر منبع خطا ε، دو منبع خطای u و چهار منبع خطای v نیز موجود است. به جای استفاده از چنین سیستم معادلاتی، می توانیم بخش های سطح سوم را در سطح دوم و سطح دوم را در داخل سطح یک جایگذاری کرده و پس از مرتب سازی به معادله ای به فرم معادله (3-1) دست یابیم.

(3-1)

$$Y_{\text{ijk}} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ijk} + \gamma_{11}W_jX_{ijk} + v_{00j} + v_{01j}W_j + v_{10j}X_{ijk} + v_{11j}W_jX_{ijk} + u_{0ij} + X_{ijk}u_{1ij} + \varepsilon_{ijk}$$

می توانیم ترکیب خطی بالا را به صورت ماتریسی و به فرم معادله (4-1) بنویسیم

(4-1)

$$Y_{ijk} = [\gamma_{00} \quad \gamma_{01} \quad \gamma_{10} \quad \gamma_{11}] \times \begin{bmatrix} 1 \\ W_j \\ X_{ijk} \\ W_jX_{ijk} \end{bmatrix} + [v_{00j} \quad v_{01j} \quad v_{10j} \quad v_{11j}] \times \begin{bmatrix} 1 \\ W_j \\ X_{ijk} \\ W_jX_{ijk} \end{bmatrix} + [1 \quad X_{ijk}] \times \begin{bmatrix} u_{0ij} \\ u_{1ij} \end{bmatrix} + \varepsilon_{ijk}$$

بنابراین در حالت کلی عبارت بالا را می توان به فرم ماتریسی و فرم معادله (5-1) نوشت.

(5-1)

$$Y = X\beta + Z_1v + Z_2u + \epsilon$$

در این معادله، به ترتیب X ، Z_1 و Z_2 بخشی از متغیرهای کمکی است که در پیش بینی اثرات ثابت، متغیرهای موجود در سطح دوم و سوم مدل سازی درگیر می باشد. ارایه مدل آماری به فرم معادله (5-1)

دارای مزایایی نسبت به فرم مبسوط(چندسطحی) می باشد. که از آن جمله می توان به فشرده تر بودن نحوه ارایه مدل، تعیین ساده تر اثرات ثابت و تصادفی و نیز نزدیک بودن این نمایش به خروجی نرم افزارهای متداول آماری اشاره نمود. اما علی رغم مزایای بیان شده تشخیص و تفسیر سطوح مدل در بیان به این فرم بسیار مشکل می باشد [1-3]. در این فرمول $X\beta$ بخش اثرات مدل و $Z_1v + Z_2u + \epsilon$ بخش اثرات تصادفی مدل می باشد. در این حالت، Z_1v اثر تصادفی مربوط به سطح سوم و Z_2u اثر تصادفی مربوط به سطح دوم و ϵ اثر تصادفی مربوط به سطح اول مدل می باشد. با در نظر گرفتن $Z = [Z_1 \quad Z_2]$ و $b = \begin{bmatrix} v \\ u \end{bmatrix}$ مدل را می توان به صورت کلی تر و به فرم معادله (1-6) نوشت.

$$Y = X\beta + Zb + \epsilon \tag{1-6}$$

این نحوه نگارش متناظر با نگارش مدل خطی آمیخته می باشد. از این رو یکی از اسامی تحلیل چندسطحی، تحلیل آمیخته و یا مدل اثرات تصادفی می باشد. همان گونه که در فرم معادلات (1-1) هم نمایش داده شده است، فرض رایج این مدل، نرمال بودن توزیع Z_1 ، Z_2 و ϵ است. با در نظر گرفتن این فرضیه استفاده روش‌های عددی، برآورد پارامترها و پیشگویی، با روش‌های معمول امکان‌پذیر می گردد[8].

متاسفانه این فرض ممکن است غیرواقعی بوده و ویژگی های مهم تغییرات درون و میان واحدها را به خوبی منعکس نساخته و در نتیجه استنباط ما را در مورد پارامترهای مورد بررسی نامطمئن و نادرست سازد. لانگ و رایان[1] و زانگ و داویدیان[2] با ذکر مثال هایی نشان داده اند که فرضیه نرمال بودن پارامتر

1-Lange and Ryan
2-Zhang and Davidian