
بسم الله الرحمن الرحيم

دانشکده

رشته آمار ریاضی ، گرایش آمار ریاضی

پایان‌نامه کارشناسی ارشد

مدل‌های اثرات تصادفی فضایی و غیرفضایی بیزی برای برآورد نواحی کوچک

نگارنده: مطهره یوسفی

استادان راهنما

دکتر محمدرضا ربیعی
دکتر حسین باغیشنی

شهریور ۱۳۹۷

شماره:
تاریخ:

باسمه‌تعالی

مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم مطهره یوسفی با شماره دانشجویی ۹۴۱۹۳۰۴ رشته آمار گرایش آمارریاضی تحت عنوان مدل‌های اثرات تصادفی فضایی و غیرفضایی بیزی برای برآورد نواحی کوچک که در تاریخ ۹۷/۶/۱۳ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

قبول (با درجه: بسیار خوب.....) ☐ **مردود** ☐

نوع تحقیق: **نظری** ☑ **عملی** ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنمای اول	دکتر محمدرضا ربیعی	استادیار	
۲- استادراهنمای دوم	دکتر حسین باغیشنی	استادیار	
۳- استاد مشاور			
۴- نماینده تحصیلات تکمیلی	دکتر احمد معتمدنژاد	دانشیار	
۵- استاد ممتحن اول	دکتر محمد آرشی	دانشیار	
۶ـ استاد ممتحن دوم	دکتر نگار اقبال	استادیار	

نام و نام خانوادگی رئیس دانشکده: دکتر ابراهیم هاشمی

تاریخ و امضاء و مهر دانشکده:

تبصره: در صورتی که کسی مردود شود حداکثر یکبار دیگر (در مدت مجاز تحصیل) می تواند از پایان نامه خود دفاع نماید (دفاع مجدد نباید زودتر از ۴ ماه برگزار شود).

الهی

جان ما را صفای خود ده و دل ما را هوای خود ده، و چشم ما را ضیای خود ده، و ما را از فضل و کرم خود آن ده که آن به.

یارب دل ما را تو به رحمت جان ده درد همه را به صابری درمان ده

این بنده چه داند که چه می باید جست دانندهٔ تویی هر آنچه دانی آن ده

تقدیم به پدر و مادرم عزیزم

که از نگاهشان صلابت

از رفتارشان محبت

و از صبرشان ایستادگی را آموختم

سپاس گزاری

سپاس خدای را که سخنوران، در ستودن او بمانند و شمارندگان، شمردن نعمت های او ندانند و کوشندگان، حق او را گزاردن نتوانند. سلام و درود بر محمّد و خاندان پاک او، طاهران معصوم، هم آنان که وجودمان وامدار وجودشان است.

بدون شک جایگاه و منزلت معلم، اجلّ از آن است که در مقام قدردانی از زحمات بی شائبه ی او، با زبان قاصر و دست ناتوان، چیزی بنگاریم. اما از آنجایی که تجلیل از معلم، سپاس از انسانی است که هدف و غایت آفرینش را تامین می کند و سلامت امانت هایی را که به دستش سپرده اند، تضمین؛ بر حسب وظیفه و از باب „ من لم یشکر المنعم من المخلوقین لم یشکر اللهِ عزّ و جلّ „ ؛ از پدر و مادر عزیزم این دو معلم بزرگوارم که همواره بر کوتاهی و درشتی من، قلم عفو کشیده و کریمانه از کنار غفلت هایم گذشته اند و در تمام عرصه های زندگی یار و یاوری بی چشم داشت برای من بوده اند؛ از اساتید با کمالات و شایسته؛ جناب آقای دکتر ربیعی و دکتر باغیشنی که در کمال سعه صدر، با حسن خلق و فروتنی، از هیچ لگی در این عرصه بر من دریغ ننمودند و زحمت راهنمایی این رساله را بر عهده گرفتند؛ تشکر می کنم. باشد که این خردترین، بخشی از زحمات آنان را سپاس گوید. از استادان دلسوز گروه آمار دانشگاه صنعتی شاهرود؛ سرکار خانم دکتر اقبال و جناب آقای دکتر آرشی که زحمت داوری این پایان نامه را متقبل شدند؛ کمال تشکر و قدردانی را دارم.

مطهره یوسفی
شهریور ۱۳۹۷

تعهد نامه

این‌جانب **مطهره یوسفی** دانشجوی کارشناسی ارشد رشته **آمار ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **مدل‌های اثرات تصادفی فضایی و غیرفضایی بیزی برای برآورد نواحی کوچک** ، تحت راهنمایی دکتر **محمدرضا ربیعی** و دکتر **حسین باغیشنی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط این‌جانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام ‘‘ دانشگاه صنعتی شاهرود ’’ یا “ Shahrood University of Technology ” به چاپ خواهند رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌شود.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده‌اند.

مطهره یوسفی
شهریور ۱۳۹۷

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته‌شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان‌نامه بدون ذکر منبع مجاز نیست.

چکیده

در سال‌های اخیر مساله برآورد کوچک ناحیه‌ای به‌دلیل نیاز به آمارهای قابل اعتماد، بسیار مورد توجه قرار گرفته است. مشکل عمده در این زمینه، ناممکن بودن اندازه‌گیری متغیر هدف برای تک تک افراد حاضر در نواحی مورد نظر است. حتی، نمونه‌گیری از همه نواحی تحت مطالعه می‌تواند هزینه‌های مالی و زمانی سنگینی در پی داشته باشد. مدل‌های آماری متعددی برای نیل به این هدف پیشنهاد شده‌اند. هدف اصلی این مدل‌ها، به‌کارگیری اطلاعات کمکی برای ارتقا برآوردهای مستقیم است. به‌همین منظور، اهمیت استفاده از اطلاعات مکانی نواحی در چارچوب مدل‌های فضایی، نقش موثری در تحلیل نواحی کوچک دارند. مشابهت فضایی بین نواحی همجوار از جمله اطلاعات مفید پرکاربرد است که مدل‌های کوچک ناحیه‌ای فضایی برای استفاده از این اطلاعات، پیشنهاد و توسعه یافته‌اند. با توجه به این‌که در بسیاری از تحلیل‌های نواحی کوچک با پاسخ‌های غیر نرمال مانند پاسخ‌های شمارشی مواجه هستیم، در این پایان‌نامه رده مدل‌های آمیخته خطی تعمیم‌یافته فضایی را در چارچوب استنباط بیزی، برای تحلیل داده‌های نواحی کوچک مورد بررسی قرار می‌دهیم. با تحلیل مجموعه داده‌های بیمه استان گیلان، عملکرد این مدل‌ها را ارزیابی می‌کنیم.

کلمات کلیدی: برآورد کوچک ناحیه‌ای، مدل‌های فضایی، مدل‌های آمیخته خطی تعمیم‌یافته فضایی، استنباط بیزی.

پیش‌گفتار

روش‌های برآورد کوچک ناحیه‌ای به منظور جبران نقص مربوط به خطاهایی که از برآوردهای حاصل از داده‌های نمونه‌ای با تعدادی محدود به وجود آمده است، به‌کار گرفته می‌شود (رائو، ۲۰۰۳). در این روش سعی بر این است، تا با به‌کارگیری از مدل‌های آماری مناسب برای استفاده از سایر اطلاعات موجود در هر ناحیه که با متغیر تحت بررسی مرتبط هستند، برای بهبود این نوع از برآوردگرها استفاده شود. استفاده روزافزون از این برآوردها باعث شده است که قابل اطمینان بودن آن‌ها حائز اهمیت قرار گیرد. از آن‌جا که اندازه نمونه در این نواحی کوچک یا بسیار کوچک است، برآوردگرهای مستقیم با خطای استاندارد و ضریب تغییرات بزرگ همراه هستند. بنابراین لزوم استفاده از برآوردگرهای غیرمستقیم در ناحیه کوچک برای دسترسی به آماره‌های قابل اطمینان و دقیق اهمیت پیدا می‌کند. در عمل نواحی جغرافیایی، معمولا، از یکدیگر مستقل نیستند و ویژگی‌های مورد مطالعه در هر ناحیه مشابه نواحی همسایه هستند. به‌کارگیری این وابستگی بین نواحی، که به آن وابستگی فضایی گویند، می‌تواند به‌عنوان اطلاعات کمکی در برآورد نواحی کوچک موثر باشد و به برآوردگرهای دقیق‌تری منجر شود.

هدف اصلی در این پایان‌نامه به‌کارگیری دیدگاه بیزی در برآورد کوچک ناحیه‌ای است. آرورا و لاهیری (۱۹۹۷) طبق قضیه‌ای نشان دادند که برآوردهای بیزی برابر بهترین پیشگوی نااریب خطی پارامترها هستند و آن‌ها نشان دادند که مزیت روش‌های بیزی، در کوچک بودن میانگین توان دوم خطاها است. برای تعیین برآوردهای بیزی، با در نظر گرفتن پیشین مناسب و استفاده از نمونه‌گیری گیبز، توزیع پسین پارامترها به‌دست می‌آید و پارامتر مورد علاقه به روش بیزی برآورد می‌شود. به‌عبارت دیگر با به‌دست آوردن توزیع پسین مورد نظر، برآوردگر بهینه به‌دست می‌آید. همچنین برای بیان اثرات تصادفی فضایی حاصل از مجاورت نواحی، به علت بهینگی، معمولا، از مدل اتورگرسیو شرطی استفاده می‌شود.

با این مقدمه، تمرکز این پایان‌نامه به استنباط بیزی در مدل‌های کوچک ناحیه‌ای (فضایی) و کاربرد آن برای پیش‌گویی برآورد کوچک ناحیه‌ای است. در فصل اول، برخی از مفاهیم پایه‌ای مرتبط با بر برآوردهای نواحی کوچک معرفی می‌شود. در فصل دوم، مدل‌های کوچک ناحیه‌ای را معرفی می‌کنیم. در فصل سوم، مدل‌بندی بیزی را بر روی نواحی کوچک بیان خواهیم کرد و در فصل چهارم نیز کاربرد مدل فضایی بیزی را برای برآورد کوچک ناحیه‌ای داده‌های بیمه استان گیلان نشان خواهیم داد. همچنین با ارائه پیشنهاداتی و یک نتیجه‌گیری کلی، ادامه روند پژوهش را در آینده تحقیق، مطرح می‌نماییم.

لیست مقالات مستخرج از پایان‌نامه

۱. یوسفی، م.، ربیعی، م.ر. و باغیشنی، ح.(۱۳۹۷).''استنباط بیزی فضایی نواحی کوچک برای بیمه‌شدگان اجباری استان گیلان''، مجموعه مقالات چهاردهمین کنفرانس آمار ایران، شاهرود (۶۸۴).

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مفاهیم پایه برآورد کوچک ناحیه‌ای

با توجه به محدود بودن منابع، استفاده از داده‌ها در سطوح پایین جغرافیایی الزامی است، که در این صورت برآوردگرهای مستقیم[۱] با خطای استاندارد و ضریب تغییرات بزرگ همراه می‌شوند. لذا برای بالا بردن دقت برآوردگرها و کاهش واریانس آن‌ها از برآوردگرهای غیرمستقیم[۲] کوچک ناحیه‌ای استفاده می‌شود. بنابراین در برآورد نواحی کوچک[۳] روش‌هایی پیشنهاد شده‌اند. که هر کدام از آن‌ها حائز اهمیت هستند. در این فصل به تعاریف و مقدمات لازم برای ورود به فصل‌های بعدی خواهیم پرداخت. عمده از مطالب استفاده شده در این فصل برگرفته از کتاب برآورد کوچک ناحیه‌ای رائو (۲۰۰۳) و کتاب آمار فضایی محمدزاده (۱۳۹۴) است.

۱.۱ مقدمه

دانشواژه «**کوچک ناحیه**[۴]» بیانگر هر زیر جامعه‌ای است که برای آن نتوان برآوردهای مستقیم را با دقت کافی تولید کرد (رائو، ۲۰۰۳). دیر زمانی است که تقاضا برای برآوردهای کوچک ناحیه‌ای به طور قابل توجهی افزایش یافته است. این توجه به دلیل افزایش تقاضای برآوردهای معتبر برای نواحی کوچک است. زیرا تصمیم‌های مربوط به کسب و کار، به‌خصوص تصمیم‌های

[۱] Direct estimators
[۲] Indirect estimators
[۳] Small area estimation
[۴] Small area

مربوط به بخش‌های خصوصی و دولتی که مرتبط با بنگاه‌های کوچک‌اند، به شدت متکی بر شرایط محلی هستند. در حال حاضر روش برآورد کوچک ناحیه‌ای برای بررسی مسائل اقتصادی آن‌هم درکشورهای اروپایی و کشورهای اتحاد شوروی اهمیت بسزایی دارد، زیرا این کشورها در حال دورشدن از تصمیم‌گیری‌های متمرکز می‌باشند.

عموما بررسی‌ها به گونه‌ای طراحی شده‌اند که برآورد پارامترهای مورد نظر، در سطوح ملی یا حوزه‌های کشوری حاصل می‌شوند. بنابراین این برآوردها در سطوح پایین‌تر نیز لازم هستند. از طرفی اندازه‌ی نمونه‌ها در این سطوح کوچک‌اند، پس، نمی‌توان این برآوردها را به‌طور قابل قبولی با داده‌های بررسی به‌دست آورد. اغلب داده‌های ثبتی در دسترس‌اند، اما تنها اطلاعات بسیار کمی را در رابطه با متغیرهای مورد نظر خواهند داشت.

در خصوص برآورد کوچک ناحیه‌ای، رائو (۲۰۰۳)، جیانگ و همکاران (۲۰۰۶)، پففرمن (۲۰۱۳) و رائو و مولینا (۲۰۱۵) تاریخچه‌ای را از شکل‌گیری آن را شرح داده‌اند و روش‌ها و مطالعات موجود در این زمینه را دسته‌بندی کرده‌اند (آرورا و لاهیری، ۱۹۹۷).

پرسش کلیدی که در برآورد نواحی کوچک مطرح خواهد شد این است که، وقتی اندازه‌ی نمونه شامل مشاهدات بسیار اندک و اطلاعات کم باشد، چگونه می‌توان برآوردهای بسیار معتبر و دقیقی را به‌دست آورد؟ برای پاسخ به این پرسش، دانستن این موضوع کافی است که زمانی که اندازه‌ی نمونه‌ی به‌دست آمده از یک ناحیه، کوچک باشد (حتی ممکن است گاهی نیز صفر باشد) برآوردهای مستقیم از یک انحراف معیار بزرگ و غیرقابل قبول برخوردار می‌شوند (سالواتی، ۲۰۰۴). به همین منظور برآوردهای به‌دست آمده از این داده‌ها نامعتبراند. لذا فنون برآورد نواحی کوچک را برای تحلیل و بررسی این مشکلات به‌کار می‌برند.

مشکل عمده‌ای که در برآورد نواحی کوچک وجود دارد، ناممکن بودن اندازه‌گیری متغیر هدف[۵] برای تک تک افراد حاضر در نواحی مورد نظر و استفاده کردن از یک نمونه است. حتی ممکن است، نمونه‌گیری از همه نواحی تحت مطالعه هزینه‌های مالی و نیز زمانی سنگینی را در پی داشته باشد (سارندال و همکاران، ۱۹۹۲).

کافی نبودن اندازه داده‌ها در آمارگیری از نواحی کوچک، موجب کم دقتی برآوردهای مستقیم در قسمت‌های مختلف نواحی می‌شود. لذا برای تولید آماره‌های معتبر و دقیق در نواحی کوچک، مطالعات زیادی انجام شده‌اند که با ارائه رهیافت‌های مناسب به دنبال حل این مشکلات هستند. یکی از این رهیافت‌ها را می‌توان به‌کارگیری منابع مختلف و اطلاعات کمکی در نظر گرفت تا با دادن قدرت قرضی به برآوردگر، دقت آن راافزایش دهند. به این نوع از برآوردگرها که باعث افزایش دادن دقت نواحی کوچک می‌شوند، برآوردگرهای غیرمستقیم می‌گویند. به‌طور کلی باید قدرت متغیر پاسخ‌مان را با استفاده از اطلاعات کمکی افزایش دهیم چرا که، موجب افزایش دقت برآوردگر مورد نظر می‌شود. اغلب از یک مدل رگرسیونی مناسب و برازش آن به نمونه منتخب و بهره‌گیری از اطلاعات کمکی اضافی، می‌توان نواحی خارج از نمونه‌گیری را برآورد کرد (بانرجی و همکاران، ۲۰۱۴).

[۵] Target variable

در حال حاضر، با توجه به پیشرفت‌های محاسباتی و توان رایانه‌های امروزی، در دیدگاه کلاسیک و بیزی، به‌دست آوردن یک آماره معتبر در مدل‌های خطی کار پیچیده‌ای نیست. از این رو، در دیدگاه کلاسیک برای برازش مدل‌های رگرسیونی با اثرات تصادفی، در برآورد نواحی کوچک، معمولا از روش بهترین پیشگوی نااریب خطی تجربی[۶] (EBLUP) استفاده می‌شود (رابینسون، ۱۹۹۱؛ جیانگ و لاهیری، ۲۰۰۶). اما هنگامی که متغیر هدف غیرنرمال باشد، برازش این مدل‌های مناسب و محاسبه EBLUP کار پیچیده‌ای خواهد شد. با توجه به مشکلات محاسباتی دیدگاه کلاسیک در مدل‌های نواحی کوچک، رهیافت مدل‌بندی بیزی آن‌ها، به‌دلیل وجود الگوریتم‌های نمونه‌گیری زنجیر مارکوف مونت کارلویی[۷]MCMC، طرفداران بیشتری پیدا کرده است (گوش و همکاران، ۱۹۹۴).

افزون براین، دیدگاه بیزی مدل‌های نواحی کوچک مزایای ارزشمند دیگری نیز دارد؛ به‌طور جزئی‌تر، این دیدگاه یک چارچوب منسجم را فراهم می‌آورد که بر اساس آن می‌توان انواع متغیرهای هدف شامل پیوسته، دودویی، دسته‌ای و انواع ساختارهای وابستگی شامل مستقل، وابستگی فضایی[۸] و فضایی‌ـ‌زمانی [۹] نواحی فاقد اطلاعات مستقیم سرشماری شده را در نظر گرفت، تا با روش‌های محاسباتی یکسان، مدل را برازش داد.

با توجه به این‌که اشاره‌ای به وابستگی فضایی شده است، کافی است بدانید، نواحی جغرافیایی معمولا از یکدیگر مستقل نیستند و ویژگی‌های مورد مطالعه در هر ناحیه مشابه نواحی همسایه هستند. به‌کارگیری این وابستگی بین نواحی، که به آن وابستگی فضایی گویند، می‌تواند به عنوان اطلاعات کمکی در برآوردهای نواحی کوچک موثر واقع شود تا برآوردهای دقیق و با کیفیت بهتری نتیجه شود. گوش و همکاران (۱۹۹۸)، رائو (۲۰۰۳)، پراتسی و سالواتی (۲۰۰۸) و جیانگ و لاهیری (۲۰۰۶) با استفاده از مدل‌های مختلف فضایی اثرات نواحی همسایه‌ها را در پیش‌گویی نواحی کوچک لحاظ کردند. یک ویژگی مطلوب اطلاعات مکانی آن است که همیشه و بدون تحمیل هزینه‌ای، در دسترس هستند. با توجه به این‌که استنباط آماری مدل‌ها در دیدگاه بیزی مبتنی بر توزیع پسین است و معمولا، صورت تحلیلی خاصی ندارد، به‌کارگیری روش‌های MCMC پرکاربرد خواهد بود و برخلاف روش‌های کلاسیک که معمولا بر گزاره‌های مجانبی متکی هستند، روش‌های بیزی با کمک الگوریتم‌های MCMC یک پاسخ کلی برای تقریب توزیع پسین فراهم می‌آورند. روش‌های برآوردیابی در رهیافت بیزی با وجود پیشرفت‌های محاسباتی از سختی محاسباتی کم‌تری برخوردار هستند. در این پایان‌نامه نواحی کوچک با استفاده از مدل‌های فضایی و غیرفضایی[۱۰] و نیز به‌کارگیری از مدل‌های سلسه‌مراتبی بیزی [۱۱](HB) و الگوریتم‌هایMCMC مد نظر قرار گرفته است. برخی از مقالات مروری مرتبط

[۶] Emprical best linear unbiased prediction
[۷] Markov chain monte carlo
[۸] Spatial dependence
[۹] Spatial-time
[۱۰] Spatial and non-spatial models
[۱۱] Bayesian hierarchical models

با برآورد کوچک ناحیه‌ای منتشر که کمک شایانی برای آشنایی بیشتر در این‌باره می‌کنند، عبارت‌اند از رائو (۱۹۸۶)، چاودری (۱۹۹۲)، گوش و رائو (۱۹۹۴)، مارکر (۱۹۹۹)، رائو (۲۰۰۱)، و پففرمن (۲۰۰۲). همچنین، کتاب‌هایی نیز درباره‌ی نظریه برآورد کوچک ناحیه‌ای منتشر شده‌اند، که می‌توان به رائو (۲۰۰۳) و موخوپادیای (۱۹۹۸) اشاره کرد (دیگل و همکاران، ۲۰۰۲).

۲.۱ علل تقاضا برای برآورد کوچک ناحیه‌ای

آمارهای کوچک ناحیه‌ای در انگلستان در سده‌ی یازدهم و در کانادا در سده‌ی هفدهم، به دو صورت مبتنی بر سرشماری یا اطلاعات ثبتی مورد استفاده قرار می‌گرفت (براکستون، ۱۹۸۷). در سال‌های اخیر آمارهای کوچک ناحیه‌ای به‌طور چشم‌گیری مورد توجه قرار گرفته‌اند. علت به‌کارگیری این آمارها نیز آن است، که در گذشته سیاست‌گذاران به آن دسته از برآوردهای تولیدشده‌ای که برای ناحیه‌های جغرافیایی بزرگ و وسیع بود، قانع بودند. اما در حال حاضر سیاست گذاران برای تصمیم‌گیری‌های خود نیازمند برآوردهایی در ناحیه‌های کوچکتر همانند استان و شهرستان هستند. به‌عنوان مثال، دولت اگر بخواهد برای رفع محرومیت در منطقه برنامه‌ریزی کند قبل از هر چیز باید برآوردهایی مانند نرخ بیکاری و تعداد افراد زیر خط فقر در آن منطقه را داشته باشد و صرفا دانستن آمارها در سطح کشوری نمی‌تواند کافی باشد. به‌عنوان مثالی دیگر، وقتی یک بیماری در سطح کشور شیوع یافته است، برای کنترل سریع‌تر آن توجه به شرایط مناطق مختلف بسیار مهم است.

از دلایل دیگر نیاز به آمارهای کوچک ناحیه‌ای، می‌توان به فرمول‌بندی سیاست‌ها و برنامه‌ها در تخصیص منابع دولتی به کمک دولت، برنامه‌ریزی منطقه‌ای، قانون‌گذاری از طریق حکومت‌های ملی برای ناحیه‌های مختلف، تقاضا از جانب بخش خصوصی و تصمیم‌های بنگاه‌های کوچک اشاره کرد.

۳.۱ آشنایی با مفاهیم لازم در برآورد کوچک ناحیه‌ای

از دیرباز آمارگیری نمونه‌ای به‌عنوان وسیله‌ای کم هزینه برای دست‌یابی به برآوردهایی از کل جامعه یا زیربخش‌هایی از جامعه به‌کار می‌رفته است. در این راستا، دو نوع برآوردگر به نام‌های مستقیم و غیرمستقیم مورد توجه قرار گرفته‌اند. در آمارگیری نمونه‌ای، برآوردگری را مستقیم گویند، هرگاه بر داده‌های نمونه‌ای ناحیه مبتنی باشد. این نوع از برآوردگرهای مستقیم عموما «**طرح پایه**[۱۲]» هستند اما، با این حال می‌توانند از مدلی نشات گرفته باشند و طبق همان مدل نیز توجیه شوند.

[۱۲] Design based

برآوردگرهای طرح پایه از وزن‌های نمونه‌گیری استفاده می‌کنند، و استنباط‌هایی که انجام می‌دهند مبتنی بر توزیع‌های احتمالی هستند که توسط طرح نمونه‌گیری و ثابت فرض کردن مقادیر جامعه تولید می‌شوند. برآوردگرهای مستقیم «**مدل یار**[۱۳]» که از مدل‌های «**کاری**[۱۴]» استفاده می‌کنند نیز طرح پایه‌اند، و اهدافشان «**استوار** [۱۵]» سازی استنباط‌ها در مقابل تعیین نادرست (بد مشخص‌سازی) مدل است. برآوردگرهای مدل‌یار مبنی بر نمونه‌ی به‌دست آمده از جامعه‌ای است که دارای مدل است که آن مدل از یک مدل کلی به‌دست آمده است. در نتیجه به آن «**مدل یار**» می‌گویند (برسلو، ۱۹۹۳).

در ادامه برخی از مفاهیم لازم در برآورد کوچک ناحیه‌ای را بیان می‌کنیم.

تعریف ۱.۳.۱. (برآوردگر غیرمستقیم). به برآوردگر، غیرمستقیمی گفته می‌شود که علاوه بر استفاده از داده‌های مشاهده شده در نمونه، از مقدارهای مورد مطالعه در ناحیه‌ها و زمان‌های مرتبط و داده‌های کمکی در سطح جامعه نیز استفاده می‌کند.

تعریف ۲.۳.۱. (ناحیه‌ی بزرگ[۱۶]). به ناحیه‌ای اطلاق می‌شود که اندازه‌ی نمونه‌ی آن به اندازه‌ی کافی بزرگ باشد که برآوردگر مستقیم حاصل از آن از دقت کافی برخوردار باشد.

تعریف ۳.۳.۱. (ناحیه‌ی کوچک). به ناحیه‌ای اطلاق می‌شود که اندازه‌ی نمونه‌ی آن به اندازه‌ی کافی بزرگ نیست که برآوردگر مستقیم حاصل از آن از دقت کافی برخوردار باشد.

تعریف ۴.۳.۱. (برآوردگر مدل پایه[۱۷]). برآوردگرهای غیرمستقیمی که مبتنی بر مدل‌های کوچک ناحیه‌ای هستند را برآوردگرهای مدل پایه می‌نامند (گلمن، ۲۰۰۶).

سه نوع برآوردگر «**غیرمستقیم**» معرفی شده‌اند که عبارت‌اند از « **غیرمستقیم حوزه‌ای** [۱۸]»، «**غیرمستقیم زمانی** [۱۹]» و «**غیرمستقیم حوزه‌ای و زمانی**[۲۰]» (رائو، ۲۰۰۳).

تعریف ۵.۳.۱. (برآوردگر غیرمستقیم حوزه‌ای). برآوردگری است که از مقادیر پاسخ و سایر متغیرهای مربوط با حوزه‌ای دیگر، اما نه از دوره‌ی زمانی دیگر، بهره می‌گیرد.

تعریف ۶.۳.۱. (برآوردگر غیرمستقیم زمانی). برآوردگری است که ازمقادیر پاسخ و سایر متغیرهای مربوط با دوره‌ی زمانی دیگر برای حوزه‌ی مورد نظر اما نه از حوزه‌ای دیگر، استفاده می‌کند.

---

[۱۳] Model assisted
[۱۴] Working
[۱۵] Robust
[۱۶] Large area
[۱۷] Model based estimator
[۱۸] Indirect domain
[۱۹] Indirect time
[۲۰] Indirect domain and time

تعریف ۷.۳.۱. (برآوردگر غیرمستقیم حوزه‌ای و زمانی). برآوردگری است که از مقادیر پاسخ و سایر متغیرهای مربوط با حوزه‌ای دیگر و نیز دوره‌ی زمانی دیگر بهره می‌گیرد (ابرلی، ۲۰۰۰).

بقیه مدل‌ها را می‌توان به دو دسته مدل‌های پیونددهنده‌ی ضمنی[۲۱] و صریح[۲۲] تقسیم نمود که در بخش ۱.۲ به‌طور کامل توضیح داده می‌شوند. همچنین دو نوع از برآوردگرهای غیرمستقیم سنتی[۲۳] وجود دارند که از مدل‌های ضمنی استفاده می‌کنند. این برآوردگرها عموما طرح پایه‌اند و واریانس‌های طرحی آن‌ها (یعنی، واریانس‌های متناظر با توزیع احتمال القا شده توسط طرح نمونه‌گیری) نسبت به واریانس‌های طرحی برآوردگرهای مستقیم کوچک می‌باشند. با این تفاوت که برآوردگرهای غیرمستقیم عموما طرح اریب[۲۴]اند و وقتی اندازه‌ی نمونه کلی افزایش یابد، مقدار طرح اریبی کاهش نخواهد یافت. اگر فرض بر این باشد که مدل پیونددهنده‌ی ضمنی به‌طور تقریبی درست باشد، آن‌گاه مقدار طرح اریبی کوچک می‌شود، و به مقداری از میانگین توان دوم خطای طرحی منجر می‌شود که به‌طور معنی‌داری در قیاس با میانگین توان دوم خطای برآوردگر مستقیم کوچک‌تر است. کاهش در میانگین تنها دلیل عمده برای بهره‌گیری از برآوردگرهای غیرمستقیم در میانگین است (رائو، ۲۰۰۳). برآوردگرهای غیرمستقیم مبتنی بر مدل‌های کوچک ناحیه‌ای را «**برآوردگرهای مدل پایه**» می‌نامند.

استنباط‌های ناشی از برآوردگرهای مدل پایه به توزیعی اشاره دارند که از مدل مفروض نتیجه می‌شود. بنابراین، مدل‌گزینی و اعتبارسنجی نقشی حیاتی در روش برآورد مدل پایه ایفا می‌کنند. اگر مدل‌های مفروض برازش خوبی را به داده‌ها فراهم نسازند، برآوردگرهای مدل پایه، مدل اریب خواهند بود، که به نوبه خود می‌توانند به استنباط‌های ناکارآمد منجر شوند. به دلیل برتری‌های استفاده از مدل‌های صریح کوچک ناحیه‌ای به‌طور کلی هرگاه بخواهیم برآوردگرهای غیرمستقیم را به‌دست آوریم، بایستی بر اساس مدل‌های صریح کوچک ناحیه‌ای باشند؛ زیرا

۱. تحت مدل مفروض، برآوردگرهای «**بهینه**[۲۵]» را می‌توان به‌دست آورد.

۲. برخلاف معیارهای فراموضعی (که نسبت به ناحیه‌های کوچک گرفته شده‌اند) و اغلب با برآوردگرهای غیرمستقیم سنتی به‌کار می‌روند، معیارهای تغییرپذیری مختص ناحیه را می‌توان با هر برآوردگر همراه ساخت.

۳. اعتبار مدل‌ها را می‌توان از روی داده‌های نمونه‌ای سنجید.

۴. بسته به طبیعت متغیرهای پاسخ و پیچیدگی ساختارهای داده‌ها (مانند ساختارهای فضایی و سری زمانی)، انواع گوناگون مدل‌ها را می‌توان آزمود.

---

[۲۱] Implicit

[۲۲] explicit connective

[۲۳] Indirect traditional estimators

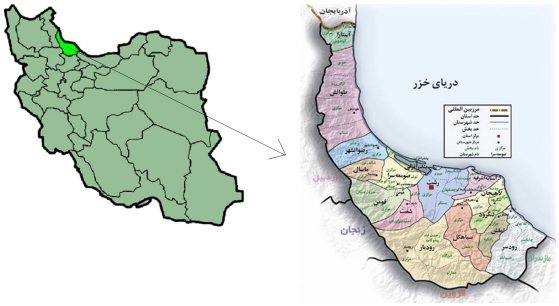
[۲۴] Design bias

[۲۵] Optimal

۱.۳.۱ برآورد کوچک ناحیه‌ای

اغلب آمارگیری‌های نمونه‌ای برای فراهم آوردن برآوردهای طرح پایه‌اند که از لحاظ آماری قابل اعتماد باشند، و همچنین از لحاظ جغرافیایی در سطح‌های ملی، استانی و ناحیه‌ای طراحی می‌شوند. به‌طور کلی این آمارگیری‌ها برای فراهم آوردن برآوردهایی برای انواع زیرجامعه‌ها و حوزه‌ها به‌کار برده می‌شوند.

برای بیان کردن حوزه‌ها همانند شکل ۱.۱ می‌توان این گونه تعریف کرد که، حوزه‌ها شامل ناحیه‌های جغرافیایی همانند ایالت یا استان، شهرستان، دهستان، یا ناحیه‌های شهرداری می‌باشند.



شکل ۱.۱: نقشه حوزه‌های استان گیلان

برای غلبه بر مشکل اندازه کم نمونه، روش‌های برآورد کوچک ناحیه‌ای مورد استفاده قرار می‌گیرد چرا که برآوردهایی تولید می‌کنند که بهتر از برآورد مستقیمی است که از نمونه در هر ناحیه کوچک به‌دست می‌آید. در حالت کلی به عنوان تعریفی از روش‌های برآورد کوچک ناحیه‌ای می‌توان گفت: ”روش‌های ایجاد برآودهای قابل اعتماد و مناسب برای نواحی هستند، که در آن‌ها استفاده از روش‌های برآورد مستقیم، دقت مورد نیاز را فراهم نمی‌کند.“

استفاده از برآورد مستقیم، به معنی به‌کارگیری از روش‌های برآورد نمونه‌ای طرح پایه‌ی کلاسیک است و از واحدهای نمونه‌ی موجود در همان ناحیه استفاده می‌کند. بنابراین اگر نمونه ناحیه به اندازه‌ی کافی بزرگ نباشد که بتواند برآورد مستقیم و قابل اعتماد و همچنین با دقت کافی تولید کند آن ناحیه به عنوان ناحیه‌ی کوچک تلقی نمی‌شود. به این منظور، بسیاری از حوزه‌های مورد نظر (همانند شهرستان) ممکن است حتی دارای اندازه‌ی نمونه‌ی صفر باشند. به این دلیل از روش‌های برآورد کوچک ناحیه‌ای می‌توان برای تهیه‌ی آماره‌های قابل اعتماد برای ناحیه‌های کم نمونه یا بدون نمونه استفاده کرد.

تفاوتی که بین برآوردگرهای غیرمستقیم و مستقیم وجود دارد آن است که، برآوردگرهای غیرمستقیم برخلاف برآوردگرهای مستقیم تنها از واحدهای نمونه یا مقدار متغیر مورد نظر همان ناحیه استفاده نمی‌کنند، بلکه این برآوردگرها برای فراهم آوردن برآوردهایی با سطح دقت کافی از مقدارهای متغیر مورد نظر ناحیه‌ها یا دوره‌های زمانی مرتبط دیگر نیز استفاده

می‌کنند. لذا در این برآوردگرها از اطلاعات کمکی متغیر مورد نظر مانند اطلاعات آخرین سرشماری واطلاعات ثبتی جاری نیز استفاده می‌شود. داده‌های متغیرهای مورد نظر ناحیه‌ها و دوره‌های زمانی مرتبط و اطلاعات کمکی از طریق یک مدل مناسب به فرآیند برآورد انتقال داده می‌شود. بنابراین در حالت کلی می‌توان گفت در دسترس بودن داده‌های کمکی خوب و تعیین مدل‌های پیونددهنده‌ی مناسب برای ساختن برآوردگرهای غیر مستقیم عامل‌های اساسی هستند. برای اطلاعات بیشتر راجع به برآوردهای کوچک ناحیه‌ای به پلاتگ و سینگ (۱۹۸۶)، پلاتک و همکاران (۱۹۸۷)، چاودری (۱۹۹۴)، مارکر (۱۹۹۹)، رائو (۲۰۰۱)، پففرمن (۲۰۰۲) و رائو (۲۰۰۳) مراجعه کنید.

۴.۱ برآوردهای غیرمستقیم سنتی

این دسته از برآوردگرها، مبنی بر مدل‌های پیونددهنده‌ی ضمنی، به دو صورت برآوردگرهای هم‌گذاشتی[۲۶] و ترکیبی[۲۷] دسته‌بندی شده‌اند. حالت عمومی این برآوردگرها طرح پایه است و واریانس طرحی آن‌ها معمولا نسبت به واریانس‌های طرحی برآوردگرهای مستقیم کوچک‌تر است. همچنین، برآوردگرهای غیرمستقیم سنتی درحالت کلی طرح اریب هستند و هنگامی که اندازه‌ی نمونه افزایش یابد مقدار طرح اریبی آن‌ها افزایش می‌یابد. بنابراین اگر مدل پیونددهنده‌ی ضمنی تقریبا درست باشد، مقدار طرح اریبی کوچک خواهد بود و این عمل منجر به مقداری از میانگین توان دوم خطای[۲۸] (MSE) طرحی خواهد شد که در قیاس با MSE برآوردگر مستقیم کوچک‌تر است. علت اصلی به‌کارگیری برآوردگرهای غیرمستقیم، کاهش در MSE است.

۵.۱ روش برآورد حوزه‌ای مستقیم

داده‌های آمارگیری نمونه‌ای که از طرح‌های نمونه‌گیری مطرح در روش‌های نمونه‌گیری به‌دست می‌آیند، برای تولید برآوردهای مستقیم و قابل اعتماد از مجموع‌ها و میانگین‌های مرتبط به کل جامعه و ناحیه‌ها یا حوزه‌های بزرگ استفاده می‌کنند. برآوردهای مستقیم هر ناحیه، تنها مقدارهای متغیر مورد نظر مربوط به واحدهایی را که در آن ناحیه واقع‌اند، به‌کار می‌گیرد. برای فراگیری بیشتر از روش برآورد مستقیم می‌توان به کتاب‌های کوکران (۱۹۷۷)، هدایت و سینها (۱۹۹۱)، سارندال، سوئنسون و رتمن (۱۹۹۲)، تامپسون (۱۹۹۷) و لهر (۱۹۹۹) مراجعه کنید.

---

[۲۶] Synthetic

[۲۷] combinational estimators

[۲۸] Mean squared error

از طرفی دیگر، روش‌های مدل پایه برای تولید برآوردگرهای مستقیم و استنباط‌های مربوط به آن به‌کار می‌روند، و استنباط در این روش‌ها، بر اساس نمونه‌ی خاص استخراج‌شده صرف نظر از طرح نمونه‌گیری به‌دست می‌آیند (بروئر ۱۹۶۳، رویال ۱۹۷۰؛ والیانت و همکاران، ۲۰۰۱). اما متاسفانه، راهبردهای مدل پایه تحت شرایط تعیین نادرست مدل وقتی اندازه‌ی نمونه در ناحیه افزایش یابد، ضعیف عمل می‌کند. برای مثال، هنسن و همکاران (۱۹۸۳) یک مورد از بد مشخص‌سازی مدل را معرفی کردند که از طریق آزمون‌های معنی‌داری در نمونه‌های تا اندازه‌ی ۴۰۰ قابل کشف نیست. آن‌ها سپس نشان دادند که احتمال‌های در برگرفتن میانگین، $\bar{Y}$، توسط بازه‌های اطمینان مدل پایه در نمونه‌گیری‌های مکرر به‌طور قابل توجهی کمتر از سطح مطلوب است و وقتی اندازه‌ی نمونه افزایش یابد، این کم‌پوشانی بدتر می‌شود. این عملکرد ضعیف تا حد قابل توجهی به دلیل ناسازگاری مجانبی برآوردگر مدل‌پایه نسبت به طرح نمونه‌گیری تصادفی طبقه‌ای است که توسط هنسن و همکاران (۱۹۸۳) به کار گرفته شد.

۱.۵.۱ رویکرد طرح پایه

ابتدا به بررسی روش طرح پایه به‌عنوان وسیله‌ی سنتی در روش‌های نمونه‌گیری برای به‌دست آوردن برآوردهایی از مجموع یا میانگین جامعه می‌پردازیم. برآوردهای مستقیم نواحی، با استفاده از نتایج مربوط به مجموع‌های جامعه، با تغییر دادن نمادها به سادگی به‌دست می‌آیند. فرض کنید جامعه‌ی تحت نمونه‌گیری U مبتنی بر N عنصر متفاوت است که با برچسب‌های $j = ۱, ..., N$ شناسایی می‌شوند. فرض کنید مشخصه مورد نظر مانند y مرتبط به عنصر j، به‌طور دقیق اندازه‌گیری شده باشد. بنابراین فرض می‌شود که خطاهای اندازه‌گیری وجود ندارند. در این‌جا، پارامتر مورد نظر، مجموع جامعه $Y = \sum_{\{j:j\epsilon U\}} y_j$ یا میانگین جامعه $\bar{Y} = Y/N$ است. برای استخراج نمونه s از U با احتمال $P(s)$، از طرح نمونه‌گیری استفاده می‌شود. احتمال استخراج مشاهده نمونه $P(s)$ می‌تواند متکی بر متغیرهای طرحی معلوم از قبیل متغیرهای نشانگر طبقه و معیارهای اندازه‌گیری خوشه‌ها باشد. در عمل، برای اجرای یک طرح نمونه‌گیری از یک برنامه‌ی نمونه‌گیری استفاده می‌شود. مثلا، نمونه تصادفی ساده با اندازه‌ی n را می‌توان با بیرون آوردن عدد تصادفی از ۱ تا N بدون جایگذاری به‌دست آورد. عمومی‌ترین طرح‌های نمونه‌گیری عبارت‌اند از؛ نمونه‌گیری تصادفی ساده طبقه‌ای و نمونه‌گیری چند مرحله‌ای طبقه‌ای.

برای انجام استنباط‌هایی درباره‌ی مجموع Y، مقادیر y مربوط به نمونه‌ی انتخاب‌شده‌ی s را مشاهده می‌کنیم. برای سادگی کار فرض می‌کنیم همه‌ی عناصر $j \in s$ را می‌توان مشاهده کرد و پاسخ کامل داشت. در رویکرد طرح پایه برآوردگر $\hat{Y}$ از Y را طرح نااریب گویند اگر امید ریاضی طرحی $\hat{Y}$ برابر Y باشد؛ یعنی،

$$E_P(\hat{Y}) = \sum P(s)\hat{Y}_s = Y \qquad (۱.۱)$$

که در آن جمع بندی روی همه‌ی نمونه‌های ممکن s تحت طرح مشخص و $\hat{Y}_s$ مقدار $\hat{Y}$ برای نمونه‌ی s است. دقت کنید منظور از زیرنویس P امید ریاضی نسبت به توزیع احتمال ایجادشده توسط طرح نمونه‌گیری است. واریانس طرحی $\hat{Y}$ به صورت

$$V_P(\hat{Y}) = E_P[\hat{Y} - E_P(\hat{Y})]^2$$

نشان داده می‌شود. برآوردگری از $V_P(\hat{Y})$ را به صورت $v(\hat{Y}) = s^2(\hat{Y})$ نشان می‌دهند، و اگر $E_P[v(\hat{Y})] = V_P(\hat{Y})$ ، $v(\hat{Y})$ را P نااریب گویند. برآوردگری مانند $\hat{Y}$ طرح سازگار است اگر $\hat{Y}$ برآوردگری P نااریب باشد و وقتی اندازه‌ی نمونه افزایش یابد $V(\hat{Y})$ به صفر میل می‌کند. به‌طور کلی لازم است P سازگاری در ساختار دنباله‌ای از جامعه‌های U_v در نظر گرفته شود، به‌طوری که وقتی $v \to \infty$، هم اندازه‌ی نمونه n_v و هم اندازه‌ی جامعه N_v به ∞ میل کنند. ویژگی P سازگاری برآوردگر واریانس $v(\hat{Y})$ به‌طور مشابه تعریف می‌شود. اگر برآوردگر $v(\hat{Y})$ هر دو P سازگار باشند، آن‌گاه رویکرد طرح پایه، به دور از مقادیر جامعه‌ای، استنباط‌های معتبر درباره‌ی Y را به‌دست می‌دهند به این معنا که وقتی اندازه‌ی نمونه افزایش یابد متغیر محوری $t = (\hat{Y} - Y)/s(\hat{Y})$ در توزیع به متغیر نرمال $N(0,1)$ می‌گراید. بدین ترتیب در نمونه‌گیری مکرر وقتی اندازه‌ی نمونه‌ی افزایش یابد، پس تقریبا $100(1-\alpha)$ درصد از بازه‌های اطمینان $[\hat{Y} - z_{\alpha/2}s(\hat{Y}), \hat{Y} + z_{\alpha/2}s(\hat{Y})]$ مقدار واقعی Y را در بر می‌گیرند. وزن‌های طرحی $w_j(s)$ نقشی مهم در ایجاد برآوردگرهای طرح‌پایه ایفا می‌کنند. این وزن‌های پایه‌ای ممکن است هم به s و هم به عنصر $j(j \in s)$ وابسته باشند. یک انتخاب مهم $w_j(s) = 1/\pi_j$ است که در آن به‌ازای $j = 1, ..., N$، کمیت‌های $\pi_j = \Sigma_{s:j\in s}P(s)$ همان احتمال‌های شمول[۲۹] اند.

تعریف ۱.۵.۱. (احتمال‌های شمول). مرتبه‌ی اول و دوم متناظر با طرح نمونه‌گیری $P(.)$ عبارت‌اند از:

$$\pi_i = \Sigma_{s\epsilon i}P(s), \pi_{ij} = \Sigma_{s\epsilon i,j}P(s)$$

که در آن $\Sigma_{s\epsilon i}$ جمع‌بندی روی تمام نمونه‌های s در بردارنده‌ی عنصر i و $\Sigma_{s\epsilon i,j}$ جمع‌بندی روی همه‌ی نمونه‌های s در بردارنده‌ی عنصر $i,j = 1, ..., N$ را نشان می‌دهد (سمپت، ۲۰۰۵).

در نبود اطلاعات کمکی جامعه، برآوردگر گسترشی

$$\hat{Y} = \Sigma_{j:j\in s}w_j y_j \qquad (۲.۱)$$

را به‌کار می‌برند. در این حالت، شرط (۱.۱) به شرط زیر تبدیل می‌شود:

$$\Sigma_{\{s:j\in s\}}P(s)w_j(s) = 1, \qquad j = 1, ..., N. \qquad (۳.۱)$$

گزینه‌ی $w_j(s) = 1/\pi_j$ در شرط نااریبی (۲.۱) صدق می‌کند، بنابراین به برآوردگر مشهور هوروتیز_تامپسون (H-T) [۳۰] می‌انجامد. و بر پایه‌ی هر طرح نمونه‌گیری، نااریب است که به

[۲۹] Inclusion probabilities

[۳۰] Horwitz-thompson esimator

اصطلاح طرح ناریب گفته می‌شود. برای جزئیات برآورد واریانس به کوکران (۱۹۷۷)، ولتر (۱۹۸۵) مراجعه کنید. رائو (۱۹۷۹) نشان داده است که برآوردگر ناریب نامنفی واریانس $\hat{Y}$ لزوما به‌صورت زیر است:

$$v(\hat{Y}) = v(y) = -\Sigma_{j<k}\Sigma_{j,k\in s} w_{jk}(s) b_j b_k (\frac{y_j}{b_j} - \frac{y_k}{b_k})^۲ \qquad (۴.۱)$$

که در آن وزن‌های $w_{jk}(s)$ در شرط نااریبی صدق می‌کنند و ثابت‌های ناصفر b_j چنانند که وقتی به‌ازای همه‌ی jها $y_j \propto b_j$ باشد، واریانس $\hat{Y}$ صفر می‌شود. برای مثال، در حالت خاص برای $w = ۱/\pi_j$ و طرح اندازه ثابت نمونه‌ای، داریم $b_j = \pi_j$ و $w_{jk}(s) = (\pi_{jk} - \pi_j\pi_k)/(\pi_{jk}\pi_j\pi_k)$ که در آن به‌ازای $j < k = ۱, ..., N$ کمیت‌های $\pi_{jk} = \Sigma_{s:(j,k)\in s} P(s)$ احتمال‌های شمول توام‌اند که فرض می‌شود مثبت هستند. برآوردگر واریانس (۴.۱) در این حالت به برآوردگر مشهور واریانس سن‌ـ ییتس‌ـ گروندی (S-Y-G)[۳۱] تبدیل می‌شود. برای به‌دست آوردن برآوردگر حوزه‌ای کافی است، ابتدا تعریف می‌کنیم

$$y_{(ij)} = \begin{cases} y_j & j \in U_i \\ ۰ & o.w \end{cases}$$

$$a_{(ij)} = \begin{cases} ۱ & j \in U_i \\ ۰ & o.w \end{cases}$$

بنابراین

$Y(y_i) = \sum_{j\in U} y_{ij} = \sum_{j\in U} y_i = Y_i$ و $Y(y_i) = \sum_{j\in U} a_{ij} = \sum_{j\in U} ۱ = N_i$ که در آن U_i حوزه‌ی مورد نظر و N_i اندازه‌ی آن است.

برآوردگرهای حوزه‌ای به‌راحتی به‌صورت $\hat{Y}_i = \hat{Y}(y_i) = \sum_{j\in s} w_j y_{ij} = \sum_{j\in s_i} w_j y_j$ به‌دست می‌آیند.

برای مطالعه‌ی بیشتر در مورد سایر برآوردگرها نظیر برآوردگر رگرسیون تعمیم‌یافته و برآوردگر نسبی به رائو (۲۰۰۳) مراجعه کنید.

انتقادی که به رویکرد طرح پایه (یا نمونه‌گیری احتمالاتی) وجود دارد، این است که هر چند استنباط‌های حاصل از آن آزاد فرض‌اند، به‌جای آن‌که تنها به نمونه‌ی خاص s استخراج شده اشاره کنند به نمونه‌گیری مکرر اشاره دارند. یافتن طرح «**بهینه**» نمونه‌گیری برای استفاده در برآوردگرهای مستقیم برای مجموع یا میانگین بزرگ ناحیه‌ها، طی سال‌های گذشته توجه زیادی را به خود معطوف کرده است. به‌ویژه، موضوع‌های مهم طراحی از قبیل تعداد طبقه‌ها، تشکیل طبقه‌ها، تخصیص نمونه، و احتمال‌های گزینش، مورد بررسی دقیق قرار گرفته است

[۳۱] San-yates-groendi variance esimator

(کوکران، ۱۹۷۷). هدف در اینجا پیدا کردن طرحی «**بهینه**» است که MSE برآوردگر مستقیم را با توجه به هزینه‌ای ثابت می‌نیمم سازد.

این هدف در عمل به واسطه‌ی محدودیت‌های عملیاتی و عامل‌های دیگر به‌ندرت حاصل می‌شود. در نتیجه، طرحی «**بینابینی**» که «**نزدیک**» به طرح بهینه باشد به‌کار می‌رود. در عمل، پیش‌بینی و برنامه‌ریزی برای همه‌ی ناحیه‌ها و کاربست‌ها، در داده‌های آمارگیری میسر نیست، زیرا «**مشتری همیشه بیش از آنچه در مرحله‌ی طراحی مشخص شده است، خواهد خواست**» (فولر، ۱۹۷۳). در نتیجه، با توجه به تقاضای رو به رشد آمارهای کوچک ناحیه‌ای قابل اعتماد، برآوردگرهای غیرمستقیم همیشه در عمل لازم خواهند بود. همچنین موضوع‌هایی از طراحی نمونه‌گیری که تاثیری بر روش برآورد کوچک ناحیه‌ای دارند، به‌ویژه در برنامه‌ریزی و طراحی آمارگیری‌های نمونه‌ای بزرگ مقیاس را نیز می‌توان بررسی کرد که عبارت‌اند از: می‌نیمم‌سازی خوشه‌بندی، طبقه‌بندی، تخصیص نمونه، تلفیق آمارگیری‌ها، آمارگیری با چارچوب دوگانه و آمارگیری مکرر.

هرچند این روش‌ها برای به‌دست آوردن برآوردگرهای قابل اعتماد کاربرد دارند، اما در عمل هنگامی که اندازه‌ی نمونه کم باشد کمکی در به‌دست آوردن برآوردگرهای با دقت کافی برای کوچک ناحیه‌ها نمی‌کنند. برای مطالعه‌ی این روش‌ها به سینگ و همکاران (۱۹۹۴)، مارکر (۲۰۰۱) و رائو (۲۰۰۳) مراجعه کنید.

۶.۱ روش‌های برآورد حوزه‌ای غیرمستقیم

زمانی که اندازه‌ی نمونه به‌قدر کافی بزرگ باشد یا به‌طورکلی نمونه بهینه شده باشد خطاهای استاندارد قابل قبول خواهند بود. اما در ساختار برآورد کوچک ناحیه‌ای، برآوردهای مستقیم به‌دلیل اینکه اندازه نمونه‌ها به‌طور غیر قابل قبولی کوچک است، منجر به خطاهای استانداردی می‌شوند که به‌طور غیرقابل قبولی بزرگ‌اند؛ حتی گاهی اوقات ممکن است برای برخی ناحیه‌ها هیچ نمونه‌ای انتخاب نشده باشد. همچنین لازم است از برآوردگرهای غیرمستقیم که با بهره‌گیری از مدل‌ها و اطلاعات کمکی ساخته می‌شوند نیز استفاده شود چرا که خطای استاندارد کاهش می‌یابد. برآوردگرهای حوزه‌ای غیرمستقیم مبتنی بر مدل‌های پیوندی ضمنی را برآوردگرهای نامستقیم سنتی گویند، که عبارت‌اند از: برآوردگرهای هم‌گذاشتی، برآوردگرهای ترکیبی و برآوردگرهای جیمز–استاین[۳۲]. با توجه به این‌که برآوردگرهای غیرمستقیم حاضر برخلاف برآوردگرهای غیرمستقیم جمعیت شناختی (رائو، ۲۰۰۳) تنها از داده‌های ثبتی و سرشماری استفاده می‌کنند و نمونه‌گیری در بین نیست، تا حد زیادی مبتنی بر داده‌های آمارگیری نمونه‌ای در ترکیب با داده‌های کمکی از جامعه هستند.

[۳۲] James-stein estimator

۷.۱ دیدگاه‌های استنباطی در برآورد کوچک ناحیه‌ای

در دیدگاه استنباطی معمولا دو نوع دیدگاه وجود دارد، یکی از این دیدگاه‌ها، دیدگاه کلاسیک و دیگری بیزی است. در دیدگاه کلاسیک پارامتر مقدار ثابت ولی نامعلوم می‌باشد، اما هدف در این روش ارزیابی رفتار روش‌های استنباطی بر اساس اصل تکرارپذیری است. اما در دیدگاه بیزی مقدار پارامتر، تحققی از یک توزیع احتمالی می‌باشد که معروف به توزیع پیشین است. از طرفی هدف در این رویکرد ارزیابی عدم قطعیت موجود در روش‌های استنباطی است که توسط توزیع پسین تحقق می‌یابد. از این رو با توجه به پیشرفت‌هایی که در حیطه رایانه بوجود آمده است برازش مدل در دو دیدگاه کار سختی نیست، اما در دیدگاه کلاسیک برازش مدل رگرسیونی با اثرات تصادفی، در برآورد نواحی کوچک، از روش بهترین پیشگوی نااریب خطی تجربی میسر می‌باشد (رابینسون، ۱۹۹۱؛ جیانگ و لاهیری، ۲۰۰۶). اما این رویکرد همیشه عملی نخواهد بود زیرا اگر متغیر هدف غیرنرمال باشد به‌دست آوردن آن در روش[۳۳] EBLUP عملی نخواهد بود (گوش و همکاران، ۱۹۹۸). با توجه به مشکلات این روش دیدگاه بیزی اهمیت پیدا می‌کند چرا که در این روش محاسبه مدل‌های نواحی کوچک به دلیل وجود الگوریتم MCMC قابل محاسبه خواهد بود.

۸.۱ روش‌های مبتنی بر نمونه‌گیری

روش‌های مبتنی بر نمونه‌گیری که به سه صورت زیر می‌باشد برای استنباط بیزی مورد استفاده قرار می‌گیرد.

۱.۸.۱ مونت کارلوی زنجیره مارکوفی MCMC

یکی از روش‌های معمول که به دلیل انتگرال‌های پیچیده موجود در تابع درستنمایی برای تعیین تحلیلی توزیع پسین به کار می‌رود MCMC است، چرا که شبیه‌سازی از توزیع پسین به کمک این روش آسان‌تر خواهد بود. این روش در سال ۱۹۴۹ با انتشار مقاله‌ای به نام روش مونت کارلو توسط دو ریاضیدان آمریکایی به نام‌های نیمن و یولام شناخته شد. پایه‌های نظری این روش از مدت ها قبل شناخته شده بود و در قرن نوزدهم و اوایل قرن بیستم گاهی اوقات مسائل آماری را با کمک شبیه‌سازی کمیت‌های تصادفی که همان روش مونت کارلو است، حل می‌کردند. اما تا پیش از اختراع رایانه‌های الکترونیکی از این روش به دلیل خسته کننده بودن شبیه‌سازی کمیت‌های تصادفی زیاد استفاده نمی‌شد. بنابراین می‌توان شروع استفاده از روش مونت کارلو را به عنوان یک تکنیک عددی عمومی، همزمان با اختراع رایانه‌های الکترونیکی دانست. نام مونت کارلو از شهر مونت کارلو واقع در موناکو گرفته شده است.

[۳۳] Estimator best linear unbiased Predictor

به‌طور کلی، روش‌های MCMC بر اساس یک الگوریتم مکرر[۳۴]، زنجیری تقلیل ناپذیر[۳۵]، و نادوره‌ای ساخته شده است و نیز از دنباله‌ای از پارامترهای مورد نظر تولید می‌شود، به گونه‌ای که توزیع مانای آن همان توزیع پسین است. در واقع با تکرار متوالی مراحل الگوریتم یک نمونه وابسته مارکوفی از توزیع پسین تولید می‌شود. دنباله آغازین تکرارهای زنجیر مارکوف قبل از زمان همگرایی، دوره داغیدن نامیده می‌شود. مشاهدات بعد از این دوره را می‌توان به عنوان مشاهداتی از توزیع پسین دانست. در روش‌های MCMC همگرایی زنجیر مساله‌ای اساسی به شمار می‌رود. از جمله مهم‌ترین این روش‌ها الگوریتم‌های گیبز [۳۶] و متروپولیس ـ هستینگز[۳۷] است. با به‌کارگیری از قضیه‌ی بیز، مدل (۱.۷)برای به‌دست آوردن توزیع پسین پارامترهای ناحیه‌ی کوچک استفاده می‌شود.

$$
\begin{aligned}
f(\boldsymbol{\mu},\boldsymbol{\lambda}|\mathbf{y}) &= f(\mathbf{y},\boldsymbol{\mu}|\boldsymbol{\lambda})f(\boldsymbol{\lambda})/f_1(\mathbf{y}),\\
f_1(\mathbf{y}) &= \int f(\mathbf{y},\boldsymbol{\mu}|\boldsymbol{\lambda})f(\boldsymbol{\lambda})d\boldsymbol{\mu}d\boldsymbol{\lambda} \qquad (۵.۱)\\
f(\boldsymbol{\mu}|\mathbf{y}) &= \int f(\boldsymbol{\mu}|\mathbf{y},\boldsymbol{\lambda})d\boldsymbol{\lambda} = \int f(\boldsymbol{\mu}|\mathbf{y},\boldsymbol{\lambda})f(\boldsymbol{\lambda}|\mathbf{y})d\mathbf{y}
\end{aligned}
$$

پس بنابر مدل بالا ارزیابی $f(\boldsymbol{\mu}|\mathbf{y})$ و کمیت‌های پسینی مربوط مانند $E[h(\boldsymbol{\mu}|\mathbf{y})]$ شامل انتگرال‌گیری‌های چند بعدی است. اما اغلب اجرای انتگرال‌گیری به‌طور تحلیلی تنها نسبت به برخی از مولفه‌های $\boldsymbol{\mu}$ و $\mathbf{y}$ میسر است. اگر مسئله‌ی کاهش یافته شده تنها شامل انتگرال‌گیری یک یا دو بعدی باشد، انتگرال‌گیری عددی مستقیم می‌تواند به‌کار گرفته شود تا کمیت‌های پسینی مطلوب به‌دست آورده شوند. اما برای مسئله‌های پیچیده، لازم می‌آید که انتگرال‌های با بعد زیاد ارزیابی شوند. به‌نظر می‌رسد روش‌های تازه ابداع شده‌ی زنجیر مارکوفی MCMC می‌تواند تا حد زیادی مشکلات محاسباتی را برطرف کند. روش‌های زنجیره‌ی مارکف مونت کارلویی MCMC شامل انواع الگوریتم‌های محاسباتی که برای شبیه‌سازی توزیع‌های پسین در مدل بیزی به‌کار می‌روند (گیلکس و همکاران، ۱۹۹۶). هنگام مطالعه مدل‌های پیچیده، تنها روش نمونه‌برداری از توزیع پسین، برآوردگرهای بیزی است. این روش نمونه‌هایی از توزیع پسینی تولید می‌کند، و سپس نمونه‌های شبیه‌سازی شده را برای تقریب کمیت‌های پسینی مطلوب به‌کار می‌برد. می‌توان برای تشخیص همگرایی و کارایی از بسته‌های نرم‌افزاری BUGS[۳۸] و CODA روش MCMC استفاده کرد (گیلکس و همکاران، ۱۹۹۶).

با توجه به مطالب بیان شده فرض کنید $\boldsymbol{\eta} = (\boldsymbol{\mu}^T, \boldsymbol{\lambda}^T)^T$ بردار پارامترهای کوچک ناحیه‌ای $\boldsymbol{\mu}$ و پارامترهای مدل $\boldsymbol{\lambda}$ باشد. به‌طور کلی، به‌دلیل مخرج غیرقابل پیگیری $f_1(\mathbf{y})$ در مدل (۱.۷)، استخراج نمونه‌های مستقل از پسین توام $f(\boldsymbol{\mu},\boldsymbol{\lambda}|\mathbf{y})$ میسر نیست. روش MCMC با

---

[۳۴] Iterative
[۳۵] Irreducibility
[۳۶] Gibbs
[۳۷] Metropolis-hastings
[۳۸] Bayesian inference using gibbs sampling

تشکیل یک زنجیر مارکوفی $\{\boldsymbol{\eta}^{(k)}, k = ۰,۱,۲,...\}$ به گونه‌ای که توزیع $\boldsymbol{\eta}^{(k)}$ به توزیع مانا (یا ناوردای [۳۹]) یکتا برابر با $f(\boldsymbol{\eta}|\mathbf{y})$ همگرا شود، از این مشکل پرهیز می‌کند؛ و این توزیع را با $\pi(\boldsymbol{\eta})$ نشان می‌دهند. بنابراین، بعد از یک دوره‌ی «تطبیق» به‌قدر کافی بزرگ مانند d می‌توانید $\boldsymbol{\eta}^{(d+D)}...\boldsymbol{\eta}^{(d+۱)}$ را، صرف‌نظر از نقطه‌ی شروع $\boldsymbol{\eta}^{(۰)}$ ، همانند D نمونه‌ی مستقل از توزیع هدف تلقی $f(\boldsymbol{\eta}|\mathbf{y})$ کنید.

برای ساختن یک زنجیر مارکوفی، نیاز به مشخص کردن احتمال انتقال یک گامی (یا هسته‌ی) $P(\boldsymbol{\eta}^{(k+۱)}|\boldsymbol{\eta}^{(k)})$ دارید که تنها به «وضعیت» جاری زنجیر $\boldsymbol{\eta}^{(k)}$ وابسته باشد. یعنی، توزیع شرطی $\boldsymbol{\eta}^{(k+۱)}$ به شرط $\boldsymbol{\eta}^{(k)}, ..., \boldsymbol{\eta}^{(۰)}$ به «سابقه‌ی» زنجیر $\{\boldsymbol{\eta}^{(۰)}, ..., \boldsymbol{\eta}^{(k+۱)}\}$ وابسته نیست. هسته‌ی انتقال باید در شرط مانایی صدق کند:

$$\int \pi(\boldsymbol{\eta}^{(k)}) P(\boldsymbol{\eta}^{(k+۱)}|\boldsymbol{\eta}^{(k)}) d\boldsymbol{\eta}^{(k)} = \pi(\boldsymbol{\eta}^{(k+۱)}) \qquad (۶.۱)$$

معادله‌ی (۶.۱) نشان می‌دهد که اگر $\boldsymbol{\eta}^{(k)}$ از توزیع $\pi(.)$ باشد، $\boldsymbol{\eta}^{(k+۱)}$ نیز از توزیع $\pi(.)$ خواهد بود. بنابراین اگر زنجیر «برگشت‌پذیر» باشد، مانایی برقرار خواهد بود:

$$\pi(\boldsymbol{\eta}^{(k)}) P(\boldsymbol{\eta}^{(k+۱)}|\boldsymbol{\eta}^{(k)}) d\boldsymbol{\eta}^{(k)} = \pi(\boldsymbol{\eta}^{(k+۱)}) \qquad (۷.۱)$$

درستی رابطه‌ی (۷.۱) این اطمینان را می‌دهد که توزیع مانای زنجیر ایجاد شده با $P(.|.)$ همان $\pi(.)$ می‌باشد. همچنین لازم است این اطمینان حاصل شود که توزیع $\boldsymbol{\eta}^{(k)}$ به شرط $\boldsymbol{\eta}^{(۰)}$ که با $P^{(k)}(\boldsymbol{\eta}^{(k)}|\boldsymbol{\eta}^{(۰)})$ نشان داده می شود، صرف نظر از $\boldsymbol{\eta}^{(۰)}$، به $\pi(\boldsymbol{\eta}^{(k)})$ همگرا می‌شود. برای برآورده شدن این شرط، لازم است زنجیر «نافروکاستنی[۴۰] » و «نادوره‌ای» باشد. نافروکاستنی بدین معنی می‌باشد که زنجیر از همه‌ی نقطه‌های شروع $\boldsymbol{\eta}^{(۰)}$ با احتمال مثبت سرانجام به هر مجموعه‌ی غیرتهی در فضای وضعیت خواهد رسید. نادوره‌ای بودن به این معنی است که زنجیر بین مجموعه‌ها مجاز نیست و به گونه‌ای دوره‌ای نوسان می‌کند. در مورد یک زنجیر فروکاستنی و نادوره‌ای، قضیه‌ی ارگودیک[۴۱] زیر صدق می‌کند: هنگامی که $D \longrightarrow \infty$

$$\bar{h}_D = \frac{۱}{D} \sum_{k=d+۱}^{d+D} h(\boldsymbol{\eta}^{(k)}) \xrightarrow{p} E[h(\boldsymbol{\eta}|\mathbf{y})] \qquad (۸.۱)$$

که در آن $\xrightarrow{p}$ همگرایی در احتمال را نشان می‌دهد. بدین صورت، به‌ازای اندازه‌ی بزرگ از D ممکن است، بتوانید برآوردگر $\bar{h}_D$ از $E[h(\boldsymbol{\eta}|\mathbf{y})]$ را با دقت کافی به‌دست آورید. اما یافتن خطای استاندارد مونت کارلویی $\bar{h}_D$ به‌دلیل وابستگی بین نمونه‌های شبیه‌سازی شده‌ی $\boldsymbol{\eta}^{(d+D)}, ..., \boldsymbol{\eta}^{(d+۱)}$ آسان نیست.

عملا، روش‌های MCMC را می‌توان با مثالی از زنجیره مارکف که توسط تانر و وانگ (۱۹۸۷) مطرح کرده‌اند، نشان داد. فرض کنید که در توزیع پسین از $\mathbf{X} = (\mathbf{z}, \mathbf{y})$ استفاده کنید.

---

[۳۹] Invariance

[۴۰] irreducible

[۴۱] Ergodic

چگالی‌های پسین حاشیه‌ای $\mathbf{z}$ و $\mathbf{y}$ را می‌توان همانند زیر مشتق‌گیری کرد:

$$p(\mathbf{z}) = \int p(\mathbf{z}|\mathbf{y})p(\mathbf{y})d\mathbf{y}$$

و

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

بنابراین $p(\mathbf{z}|\mathbf{y})$ و $p(\mathbf{y}|\mathbf{z})$ مشخص می‌باشد و برای $p(\mathbf{z})$ نیاز به حل کردن است. و از رویکرد مبتنی بر نمونه‌گیری استفاده می‌کنند، بنابراین می‌توان از $p_0(\mathbf{z})$ شروع کرد و $p(\mathbf{z})$ را برآورد کرد و در نهایت $\mathbf{Z}^{(0)} \sim p_0(\mathbf{z})$ را نوشت. و همین‌طور می‌توان نشان داد $\mathbf{Y}^{(1)} \sim p(\mathbf{y}|\mathbf{z}^{(0)})$.

توزیع حاشیه‌ای $\mathbf{Y}^{(1)}$ برابر است با

$$p_1(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{z})p_0(\mathbf{z})d\mathbf{z}$$

در مرحله‌ی بعد $\mathbf{Z}^{(1)} \sim p(\mathbf{z}|\mathbf{y}^{(1)})$ را با توزیع حاشیه‌ای زیر نشان می‌دهند:

$$\begin{aligned} p_1(\mathbf{z}) &= \int p(\mathbf{z}|\mathbf{y})p_1(\mathbf{y})d\mathbf{y} \\ &= \int p(\mathbf{z}|\mathbf{y}) \int p(\mathbf{y}|\acute{\mathbf{z}})p_0(\acute{\mathbf{z}})d\acute{\mathbf{z}}d\mathbf{y} \\ &= \int (\int p(\mathbf{z}|\mathbf{y})p(\mathbf{y}|\acute{\mathbf{z}})d\mathbf{y}p_0(\acute{\mathbf{z}})d\acute{\mathbf{z}} \\ &= \int h(\mathbf{z},\acute{\mathbf{z}})p_0(\acute{\mathbf{z}})d\acute{\mathbf{z}} = I_h(p_0(\mathbf{z})) \end{aligned}$$

تانر و وانگ (۱۹۸۷) نشان دادند که اگر این فرآیند را تکرار کنند، به تعداد کافی جفت $(\mathbf{Z}^{(i)}, \mathbf{Y}^{(i)})$، دنباله‌ی $\mathbf{Z}^{(i)}, \mathbf{Y}^{(i)}$ که همگرا به $p(\mathbf{z},\mathbf{y})$ و نیز با توزیع حاشیه‌ای $p(\mathbf{z})$ و $p(\mathbf{y})$ همراه هستند، به‌دست آورده می‌شود. به‌عبارت دیگر برای بسندگی بزرگ i، $\{\mathbf{Z}^{(i)}\}$ را می‌توان از یک نمونه در توزیع حاشیه‌ای $p(\mathbf{z})$ درنظر گرفت و $\{\mathbf{Y}^{(i)}\}$ را نیز می‌توان از یک نمونه با توزیع حاشیه‌ای $p(\mathbf{y})$ فرض کرد.

بنابراین، الگوریتم تکراری است و نیاز به یک دوره سوختن قبل از نمونه‌های ایجاد شده دارد و می‌توان فرض کرد از تراکم حاشیه‌ای شکل گرفته است. هنگامی که خروجی MCMC یک مدل تولید کند، می‌توان مقادیر پسین مورد نظر را محاسبه نمود و با توجه به نمونه‌هایی که برای هر جز بردار پارامترها به‌دست می‌آیند، آن‌ها را می‌توان به عنوان نمونه‌هایی از توزیع پسین حاشیه‌ای مربوط درنظر گرفت. ساده‌ترین روش برای به‌دست آوردن برآورد یک تابع از پارامتر خاص (برای مثال، $\phi = f(\boldsymbol{\theta}_i)$) از طریق میانگین ارگودیک امکان‌پذیر می‌باشد و به‌صورت زیر است:

$$\hat{\phi} = \hat{f}(\boldsymbol{\theta}_i) = \frac{1}{D}\sum_{k=d+1}^{d+D} \boldsymbol{\phi}^{(k)},$$

d طول دوره‌ی سوختن و برای به‌دست آوردن همگرایی استفاده می‌شود. D اندازه‌ی نمونه باقی‌مانده و $\phi^{(k)} = f(\boldsymbol{\theta}_{(i)}^{(k)})$ با $\boldsymbol{\theta}_{(i)}^{(k)}$ از زنجیره‌ی kام در نمونه‌ی MCMC برای $\boldsymbol{\theta}_i$ می‌باشد. با توجه به برآورد دقت برآوردگرها، واریانس پسین را می‌توان طبق رابطه‌ی زیر محاسبه نمود:

$$\hat{Var}(\phi|\mathbf{y}) = \frac{1}{D-1}\sum_{k=d+1}^{d+D}(\phi^{(k)} - \hat{\phi})^2.$$

اگر صورت ریاضی برای امیدریاضی شرطی ϕ به شرط داده‌ی $\boldsymbol{y}$ و تمامی پارامترها شناخته شده باشد، می‌توان از برآوردگر بهبود یافته رائو ـ بلک ولز[۴۲] استفاده کرد (گلفند و اسمیت، ۱۹۹۱):

$$\phi^{(RB)} = \frac{1}{D}\sum_{k=d+1}^{d+D} E(\phi|\mathbf{y}, \boldsymbol{\xi}^{(k)}),$$

که $\{\boldsymbol{\xi}^{(k)}\}$ توزیع پسین حاشیه‌ای از $\boldsymbol{\xi}|\mathbf{y}$ است. واریانس پسین همانند زیر محاسبه می‌شود:

$$\tilde{Var}(\phi|\mathbf{y}) = \frac{1}{D}\sum_{k=d+1}^{d+D} Var(\phi|\mathbf{y}, \boldsymbol{\xi}^{(k)}) + \frac{1}{D-1}(\sum_{k=d+1}^{d+D} Var(\phi|\mathbf{y}, \boldsymbol{\xi}^{(k)} - \phi^{(RB)})^2.$$

رابطه‌ی فوق معتبر خواهد بود، اگر مقادیر نمونه تولید شده تقریبا مستقل و یکسان (i.i.d) باشند. در برنامه‌های کاربردی، اغلب چندین زنجیره‌ی موازی با نقاط اولیه بیش‌پراکنش شده قابل اجرا خواهد بود.

فرض کنید زنجیره‌های موازی L قابل اجرا باشد، در این هنگام معادلات ارگودیک را می‌توان به‌صورت زیر تغییر داد:

$$\hat{\phi} = \frac{1}{D}\sum_{l=1}^{L}\sum_{k_l=d+1}^{d+D}\phi^{(k_l)}$$

d طول دوره‌ی سوختن و D اندازه‌ی نمونه‌ی باقی‌مانده (d و D فرض می‌شود برای همه زنجیرها یکسان باشد)؛ و $\phi^{(k_1)} = f(\boldsymbol{\theta}_i^{(k_1)})$ با داشتن $\boldsymbol{\theta}_i^{(k_1)}$ در زنجیره‌ی (k_1)ام از نمونه (MCMC) برای $\boldsymbol{\theta}_i$ در زنجیره‌ی lام می‌باشد. واریانس پسین ϕ می‌تواند به‌صورت زیر برآورده شود:

$$\hat{Var}(\phi|\mathbf{y}) = \frac{D-1}{D}W + \frac{1}{D}B$$

از این‌رو

- $\bar{\phi}_l = \frac{1}{D}\sum_{k_l=d+1}^{d+D}\phi^{(k_l)}$ ـ مقادیر میانگین در زنجیره‌ی l
- $\bar{\phi} = \frac{1}{L}\sum_{l=1}^{L}\bar{\phi}_l$ ـ میانگین هدف در میان تمام زنجیرها
- $W = \frac{1}{L(D-1)}\sum_{l=1}^{L}\sum_{k_l=d+1}^{d+D}(\phi^{(k_l)} - \bar{\phi}_l)^2$ ـ واریانس درون زنجیره‌ای
- $B = \frac{D}{L-1}\sum_{l=1}^{L}(\bar{\phi}_l - \bar{\phi})^2$ ـ واریانس بین زنجیره‌ای.

[۴۲] Rao black well

۲.۸.۱ نمونه‌گیر گیبز

نمونه‌برداری گیبز به افتخار نام فیزیکدان جوسی ویلارد گیبز، نامگذاری شده است و اشاره به یک مقایسه بین الگوریتم نمونه برداری و فیزیک آماری دارد. الگوریتم شرح داده شده، توسط برادران استوارت و دونالد گمن در سال ۱۹۸۴، یعنی هشت دهه پس از مرگ گیبز تشریح شد.

نمونه‌برداری گیبز در تجسم اولیه، یک حالت خاص از این الگوریتم متروپلیس ـ هستینگز است. نکته‌ای که در نمونه‌برداری گیبز وجود دارد این است که، برای یک توزیع چند متغیره، نمونه‌برداری از توزیع شرطی ساده‌تر است از محاسبه توزیع حاشیه‌ای، که با انتگرال‌گیری بر روی توزیع توام به‌دست می‌آید.

منظور از نمونه‌برداری گیبز یا نمونه‌بردار گیبز در مطالعات آماری، الگوریتمی است که بر مبنای تئوری زنجیره مارکوف مونت کارلو طراحی شده است. کاربرد این الگوریتم در تولید دنباله‌ای از مشاهدات از یک تابع توزیع احتمالاتی چند متغیره است که تولید نمونه از آن به صورت مستقیم دشوار است. این دنباله را می‌توان برای تخمین توزیع هم‌زمان (مثلاً برای تولید هیستوگرام توزیع)، تخمین توزیع حاشیه‌ای بر روی یک یا زیر مجموعه‌ای از متغیرهای توزیع (مانند پارامتر پنهان یا متغیرهای پنهان)، یا برای محاسبه یک انتگرال (مانند امید ریاضی متغیرها) استفاده نمود. اغلب برخی از متغیرها وابسته به مشاهدات هستند که مقدار آن‌ها مشخص است و بنابراین نیازی به نمونه‌برداری برای آن‌ها نیست.

نمونه‌برداری گیبز معمولاً به عنوان ابزاری برای استنباط آماری و به‌ویژه در استنباط بیزی استفاده می‌شود. این روش یک الگوریتم تصادفی (از آن جهت که با استفاده از اعداد تصادفی نمونه تولید می‌کند) که می‌تواند جایگزینی برای الگوریتم‌های قطعی در استنباط آماری باشد. نمونه‌برداری گیبز مانند دیگر الگوریتم‌های زنجیره مارکوف مونت کارلو، زنجیره مارکوفی را از نمونه‌ها تولید می‌کند، به‌طوری که هر نمونه وابسته به نمونه‌های نزدیک است؛ بنابراین اگر نمونه‌های مستقل مورد نظر است، باید نمونه‌برداری محتاطانه انجام پذیرد، این کار اغلب با نازک‌سازی زنجیرهٔ نمونه‌های حاصل شده انجام می‌شود، بدین شکل، تنها مقدار nام زنجیره مثلاً ۱۰۰ام انتخاب می‌شود. علاوه بر این، نمونه‌های ابتدای زنجیره (تکرارهای سوخته) احتمالاً نمایانگر خوبی برای توزیع مورد نظر نخواهند بود.

با توجه به مطالب بالا می‌توان گفت که نمونه‌گیری گبیز یک الگوریتم MCMC با هسته تغییر وضعیت است که به کمک توزیع‌های شرطی کامل ساخته شده است.

بنابراین برای تولید نمونه‌های $\boldsymbol{\eta}^{(k)}$، بردار $\boldsymbol{\eta}$ را به بلوک‌هایی مناسب $\eta_1, ..., \eta_r$ افراز می‌کنند. برخی از بلوک‌ها ممکن است تنها یک عنصر داشته باشند در صورتی که بلوک‌های دیگر ممکن است شامل بیش از یک عنصر باشند. برای مثال، مدل پایه‌ای در سطح ناحیه با $\boldsymbol{\mu} = (\theta_1, ..., \theta_m)^T = \boldsymbol{\theta}$ و $\boldsymbol{\lambda} = (\boldsymbol{\beta}^T, \sigma_v^2)^T$ را در نظر بگیرید. در این حالت $\boldsymbol{\eta}$ را می‌توان به‌صورت $\boldsymbol{\eta_1} = \boldsymbol{\beta}$، $\eta_2 = \theta_1, ..., \eta_{m+1} = \theta_m$، $\eta_{m+2} = \sigma_v^2$ (برای $r = m + 2$) افراز کرد. ما به مجموعه‌ی

توزیع‌های شرطی گیبزی زیر نیاز داریم:

$$f(\eta_1|\eta_2,...,\eta_r,\mathbf{y}), f(\eta_r|\eta_1,...,\eta_{r-1},\mathbf{y}),...,f(\eta_2|\eta_1,\eta_3,...,\eta_r,\mathbf{y}).$$

نمونه‌گیر گیبزی از این توزیع‌های شرطی برای ساختن یک هسته‌ی انتقالی $P(.|.)$، به‌طوری که توزیع مانای زنجیر مارکوفی حاصل $\pi(\eta) = f(\eta|\mathbf{y})$ باشد، استفاده می‌شود. این نتیجه از این واقعیت بر می‌آید که $f(\eta|\mathbf{y})$ به‌وسیله‌ی مجموعه‌ی توزیع‌های شرطی گیبزی به‌طور یکتا تعیین می‌شود.

اگر یک توزیع شرطی دارای صورتی استاندارد و بسته مثلا مانند نرمال یا گامای وارونه باشد، آن‌گاه نمونه‌ها می‌توانند مستقیما از آن توزیع شرطی تولید شوند. در غیر این‌صورت، الگوریتم‌های بدیل مانند نمونه‌گیری ردی متروپلیس‌ـ هستینگز می‌توانند به‌کار گرفته شوند تا نمونه‌هایی از توزیع شرطی تولید شوند. برخی مولفان به‌کارگیری M-H را برای حالت‌هایی به‌صورت بسته نیز پیشنهاد می‌کنند. اگر M-H تنها برای حالت‌های بدون صورت بسته به‌کار رود، آن‌گاه الگوریتم را M-H درون گیبزی گویند. الگوریتم نمونه‌گیری گیبزی شامل مراحل زیر است:

- مرحله‌ی اول. یک نقطه‌ی شروع $\eta^{(0)}$ با مولفه‌های $\eta_1^{(0)},...,\eta_r^{(0)}$ را انتخاب کنید؛ قرار دهید $k = 0$. مثلا، می‌توان از برآوردهای بیزتجربی (EB)ی $\boldsymbol{\mu}$ یا پارامترهای مدل $\boldsymbol{\lambda}$ به‌عنوان مقدارهای شروع استفاده کرد. اما نقطه‌های شروع می‌توانند دلخواه باشند.
- مرحله دوم. بردار $\eta^{(k+1)} = (\eta_1^{(k+1)},...,\eta_r^{(k+1)})$ را به شرح زیر تولید کنید:

مولفه‌ی $\eta_1^{(k+1)}$ را از $f(\eta_1|\eta_2^{(k)},...,\eta_r^{(k)},\mathbf{y})$، مولفه‌ی $\eta_2^{(k+1)}$ را از $f(\eta_2|\eta_1^{(k+1)},\eta_3^{(k)},...,\eta_r^{(k)},\mathbf{y}),...,$ و $\eta_r^{(k+1)}$ را از $f(\eta_r|\eta_1^{(k+1)},...,\eta_{r-1}^{(k+1)},\mathbf{y})$ استخراج کنید.

- مرحله‌ی سوم. قرار دهید $k = k+1$ و به مرحله‌ی اول بروید.

مراحل دوم و سوم یک چرخه برای هر k را تشکیل می‌دهند. دنباله‌ی $\{\eta^{(k)}\}$ که به‌وسیله‌ی نمونه‌گیر گیبز تولید شده یک زنجیر مارکوفی با توزیع مانای $\pi(\eta) = f(\eta|\mathbf{y})$ است؛ توجه کنید که هسته‌ی انتقالی تک گامی برابر حاصل‌ضرب r توزیع شرطی گیبز است (اسمیت و گلفند، ۱۹۹۰).

۳.۸.۱ الگوریتم متروپلیس‌ـ هستینگز درون گیبزی

الگوریتم متروپلیس‌ـ هستینگز توسط متروپلیس و همکاران در سال (۱۹۵۳) و هستینگز (۱۹۷۰) معرفی شد. الگوریتم M-H برای هر یک از متغیرهای مورد نظر، با استفاده از تابع چگالی توام کامل و توزیع‌های پیشنهاد شده (مستقل)، نمونه‌ها را از توزیع احتمال شبیه‌سازی می‌کند.

این روش خاصا در مواردی به کار می‌رود که نمونه‌گیری مستقیم دشوار است. ترتیب به دست آمده را می‌توان برای تقریب‌گیری توزیع (تولید یک هیستوگرام) یا برای محاسبه

یک انتگرال استفاده کرد. متروپلیس‌ـ هستینگز و سایر الگوریتم‌های زنجیره مونت کارلویی معمولاً برای نمونه‌گیری از توزیع‌های چند بعدی در ابعاد زیاد، استفاده می‌شوند. همگرایی این الگوریتم نسبت به نمونه‌گیر گیبز منظم‌تر و کندتر می‌باشد. اما الگوریتم M-H انعطاف‌پذیری لازم را از طریق انتخاب چگالی پیشنهادی به‌دست می‌آورد. و از مشکلات الگوریتم M-H می‌توان تشخیص همگرایی زنجیر و زمان طولانی محاسبات دانست.

حال اگر همه‌ی توزیع‌های شرطی گیبزی صورت‌هایی بسته متعلق به خانواده‌های استاندارد داشته باشند، آن‌گاه تولید نمونه‌ها از توزیع‌های شرطی سر راست است. اگر توزیع شرطی صورتی بسته نداشته باشد، در این صورت برای تولید نمونه‌ها از آن توزیع شرطی چند روش وجود دارند. به‌ویژه، نمونه‌گیری ردی را برای توزیع‌های شرطی از نوع لگ کاو تک متغیره (گیلکز و وایلد، ۱۹۹۲) و به‌طور کلی‌تر M-H درون گیبزی را خواهیم داشت (متروپلیس و همکاران، ۱۹۵۳؛ هستینگز، ۱۹۷۰).

فرض کنید $f(\boldsymbol{\eta}|\boldsymbol{\eta}^k_{-i},\mathbf{y})$ توزیع شرطی گیبزی پس از تکمیل $(i-1)$ استخراج اول از چرخه‌ی $(k+1)$ام نمونه‌گیری گیبزی را نشان دهد، که در آن $\boldsymbol{\eta}^k_{-i}=\{\boldsymbol{\eta}^{k+1}_1,...,\boldsymbol{\eta}^{k+1}_{i-1},\boldsymbol{\eta}^k_{i+1},...,\boldsymbol{\eta}^k_r\}$ الگوریتم M-H برای تولید یک نمونه‌ی $\boldsymbol{\eta}^{(k+1)}_i$ از $f(\boldsymbol{\eta}_i|\boldsymbol{\eta}^{(k)}_{-i},\mathbf{y})$ شامل مراحل زیر است.

- مرحله ی اول. توزیع $f(\boldsymbol{\eta}_i|\boldsymbol{\eta}^{(k)}_{-i},\mathbf{y})$ را با یک چگالی پیشنهادی مانند $q_i(\boldsymbol{\eta}_i|\boldsymbol{\eta}^{(k)}_i,\boldsymbol{\eta}^{(k)}_{-i})$ که نمونه‌گیری از آن آسان است از قبیل توزیع نرمال یا تی‌استودیت، تقریب بزنید. چگالی پیشنهادی ممکن است به مقدارهای جاری $\{\boldsymbol{\eta}^{(k)}_i,\boldsymbol{\eta}^{(k)}_{-i}\}$ وابسته باشد.
- مرحله‌ی دوم. برای $\boldsymbol{\eta}_i$ «نامزدی» مانند $\boldsymbol{\eta}^*_i$ از چگالی نامزد و $\mathbf{u}$ را از توزیع یکنواخت $(0,1)$ تولید کنید.
- مرحله‌ی سوم. اگر $\mathbf{u}\le \mathbf{a}(\boldsymbol{\eta}^{(k)}_{-i},\boldsymbol{\eta}^{(k)}_i,\boldsymbol{\eta}^*_i)$ قرار دهید $\boldsymbol{\eta}^{(k+1)}_i=\boldsymbol{\eta}^*_i$ وگرنه $\boldsymbol{\eta}^{(k+1)}_i=\boldsymbol{\eta}^{(k)}_i$، که در آن احتمال پذیرش $a(\boldsymbol{\eta}^{(k)}_{-i},\boldsymbol{\eta}^{(k)}_i,\boldsymbol{\eta}^*_i)$ با رابطه‌ی زیر تعریف می‌شود:

$$a(\boldsymbol{\eta}^{(k)}_{-i},\boldsymbol{\eta}^{(k)}_i,\boldsymbol{\eta}^*_i)=min\{\frac{f(\boldsymbol{\eta}^*_i|\boldsymbol{\eta}^{(k)}_i,\mathbf{y})q_i(\boldsymbol{\eta}^{(k)}_i|\boldsymbol{\eta}^*_i,\boldsymbol{\eta}^{(k)}_{-i})}{f(\boldsymbol{\eta}^{(k)}_i|\boldsymbol{\eta}^{(k)}_{-i},\mathbf{y})q_i(\boldsymbol{\eta}^*_i|\boldsymbol{\eta}^{(k)}_i,\boldsymbol{\eta}^{(k)}_{-i})},1\} \qquad (۹.۱)$$

دقت کنید که احتمال پذیرش (۹.۱) به نسبت $f(\boldsymbol{\eta}^*_i|.)/f(\boldsymbol{\eta}^{(k)}_i|.)$ بستگی دارد، پس لازم است که $f(\boldsymbol{\eta}_i|.)$ را تنها تا حد ثابت تناسب بدانید، یعنی ثابت نرمال‌گر در (۹.۱) از صورت و مخرج حذف می‌شود. همچنین اگر چگالی پیشنهادی متقارن باشد، یعنی داشته باشیم $q_i(\boldsymbol{\eta}^{(k)}_i|\boldsymbol{\eta}^*_i,.)=q_i(\boldsymbol{\eta}^*_i|\boldsymbol{\eta}^{(k)}_i,.)$، آن‌گاه احتمال پذیرش به مقدار زیر فرو می‌کاهد (متروپلیس و همکاران، ۱۹۵۳).

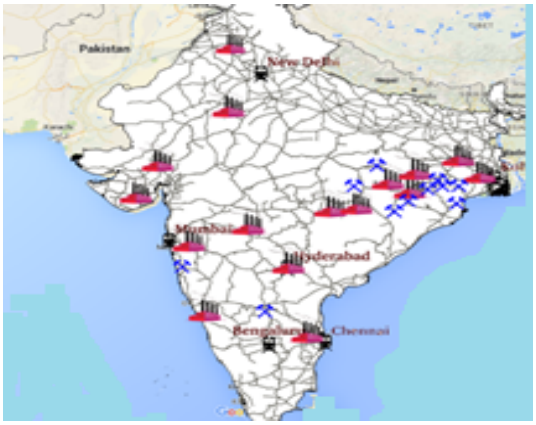
$$a(\boldsymbol{\eta}^{(k)}_{-i},\boldsymbol{\eta}^{(k)}_i,\boldsymbol{\eta}^*_i)=min\{\frac{f(\boldsymbol{\eta}^*_i|\boldsymbol{\eta}^{(k)}_{-i},\mathbf{y})}{f(\boldsymbol{\eta}^{(k)}_i|\boldsymbol{\eta}^{(k)}_{-i},\mathbf{y})},1\} \qquad (۱۰.۱)$$

نمونه‌گیر گیبزی حالتی خاص از M-H به‌ازای $q_i(\boldsymbol{\eta}_i|\boldsymbol{\eta}^{(k)}_i,\boldsymbol{\eta}^{(k)}_{-i})=f(\boldsymbol{\eta}_i|\boldsymbol{\eta}^{(k)}_{-i},\mathbf{y})$، و احتمال پذیرش متناظر برابر ۱ است به‌طوری که نامزد گیبزی $\boldsymbol{\eta}^*$ خودبه‌خود پذیرفته می‌شود. برای بهبود بخشیدن به همگرایی نمونه‌گیری MCMC، چند صورت مختلف از M-H پیشنهاد شده‌اند. برای مطالعه بیشتر می‌توانید به متروپلیس و همکاران، (۱۹۵۳) مراجعه کنید.

۹.۱ مقدمه‌ای بر آمار فضایی

گاهی اوقات در مطالعات محیطی با مشاهداتی سر و کار خواهیم داشت که نسبت به‌هم مستقل نمی‌باشند، و همبستگی بین آن‌ها ناشی از موقعیت و مکان قرار گرفتن مشاهدات در فضای مورد بررسی می‌باشد، بنابراین به این نوع از مشاهدات داده‌های فضایی[۴۳] گویند (کرسی، ۱۹۹۳). به دلیل وجود همبستگی فضایی[۴۴] بین آن‌ها روش‌های آماری برای تحلیل چنین داده‌هایی قابل به‌کارگیری نمی‌باشند و به‌همین منظور، لازم است ساختار همبستگی داده‌ها را در تحلیل‌شان به‌کار برد. داده‌های فضایی را می‌توان با توجه به انواع موقعیت‌ها، به سه گروه زیر تقسیم کرد:

داده‌های زمین آماری: این نوع از داده‌ها در موقعیت‌های ثابت و مشخص در ناحیه‌ای پیوسته مشاهده می شوند. حال ممکن است که متغیر مورد بررسی گسسته یا پیوسته باشد. به‌عنوان مثال، شکل ۲.۱ نشان‌دهنده‌ی میزان سنگ‌های فیروزه در یک معدن و یا نوعی از گیاه که در یک محدوده‌ی جنگلی است، پس، به آن‌ها داده‌های زمین آماری [۴۵] گویند. هدف از تحلیل این نوع داده‌ها، پیش‌گویی پاسخ در مکان‌های مشاهده نشده است.



شکل ۲.۱: نقشه محل سنگ‌های فیروزه در کشور هندوستان

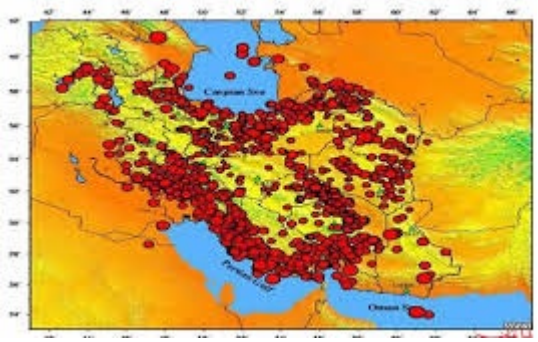
داده‌های مشبکه‌ای: این نوع از داده‌ها مرتبط به مکان‌های ناحیه‌ای هستند یعنی بین مکان‌های مشاهده شده مکان دیگری وجود ندارد. این مکان‌ها ممکن است به دو حالت منظم و نامنظم باشند. معمولا در عمل داده‌های مشبکه‌ای[۴۶] بر روی یک شبکه‌ی نامنظم جای دارند، مشابه‌ی مجموعه‌ای از استان‌ها یا دیگر مرزهای منطقه‌ای. مثلا، تعداد افراد بیمه شده در تمام مناطق استان گیلان می‌تواند نامنظم باشد و یا نرخ رشد اقتصادی هر استان

[۴۳] Spatial data
[۴۴] Spatial correlation
[۴۵] Geostatistical data
[۴۶] Lattice data

مثال‌هایی از داده‌های مشبکه‌ای می‌باشد. این نوع از داده‌ها به‌صورت گسسته و پیوسته تعریف شده‌اند. به‌طور معمول هدف از تحلیل داده‌های فضایی مشبکه‌ای، مدل‌بندی احتمالاتی مشاهدات است، که در این پایان‌نامه از این الگو استفاده می‌شود. همانند شکل ۳.۱ خطوط آبی رنگ در نقشه مراکز بیمه استان گیلان را نشان می‌دهد.

شکل ۳.۱: نمودار محل افراد بیمه شده در استان گیلان

داده‌های الگو نقطه‌ای: در این حالت مکان یا موقعیت مشاهده شده خود متغیری تصادفی است. الگوهای نقطه‌ای[۴۷] شامل تعدادی متناهی از مکان‌ها در یک ناحیه‌اند که در آن‌ها یک صفت خاص اندازه‌گیری می‌شود. به‌طور معمول الگوهای نقطه‌ای به سه دسته به‌طور کامل تصادفی فضایی CSR[۴۸] ، منظم[۴۹] و خوشه‌ای[۵۰] تقسیم می‌شوند. برای مثال، در شکل ۴.۱، موقعیت گونه‌ای از درختان را در یک ناحیه‌ی جنگلی یا موقعیت مراکز زلزله را نشان می‌دهد، که جز این نوع از داده‌های الگو نقطه‌ای به‌شمار می‌روند.



شکل ۴.۱: نمودار نقاط زلزله در سطح کشور

[۴۷] Point data
[۴۸] Complete spatial randomness
[۴۹] Regular
[۵۰] Cluster

۱.۹.۱ مدل بندی ساختار وابستگی فضایی

برای تحلیل داده‌های فضایی لازم است یک مدل آماری در نظر گرفته شود. در آمار فضایی به‌طور معمول یک میدان تصادفی به‌عنوان مدل آماری داده‌های فضایی در نظر گرفته می‌شود.

تعریف ۱.۹.۱. (میدان تصادفی)

میدان تصادفی مجموعه‌ای از متغیرهای تصادفی مانند $\{Z(s); s\epsilon D\}$ است، که در آن مجموعه اندیس‌گذار D یک زیر مجموعه از فضای اقلیدسی d بعدی، $d \geq ۱$، از R^d است. پس خواهیم داشت:

تعریف ۲.۹.۱. (توابع میانگین، واریانس و کوواریانس میدان تصادفی)

در مورد میدان تصادفی $Z(.)$ میانگین در موقعیت s و کوواریانس در موقعیت‌های $s_۱$ و $s_۲$ به‌ترتیب به‌صورت زیر :

$$\mu(s) = E[Z(s)], \quad s\epsilon D$$

$$C(s_۱, s_۲) = Cov[Z(s_۱), Z(s_۲)] =$$

$$E[(Z(s_۱) - \mu(s_۱))(Z(s_۲) - \mu(s_۲))], \quad s_۱, s_۲\epsilon D$$

تعریف می‌شوند. برای $s = s_۲, s = s_۱$، واریانس میدان تصادفی $Z(.)$ در مکان s به‌صورت

$$Var[Z(s)] = E[Z(s) - \mu(s)]^۲ = C(s, s)$$

حاصل می‌شود. هر میدان تصادفی $Z(.)$ را می‌توان به‌صورت

$$Z(s) = \mu(s) + \delta(s), \quad s\epsilon D$$

تجزیه کرد، که در آن $\mu(s)$ تغییرات بزرگ‌مقیاس یا روند و $\delta(.)$، فرایند خطا یا تغییرات کوچک‌مقیاس میدان تصادفی نامیده می‌شوند. تغییرات کوچک‌مقیاس ممکن است ناشی از خطای اندازه‌گیری یا تغییرپذیری در درون موقعیت مشاهده باشد و تغییرات بزرگ‌مقیاس ممکن است ناشی از تغییرپذیری بین موقعیت‌های مشاهده شده باشند. به‌طور معمول تحلیل داده‌های فضایی بر اساس مشاهدات نمونه دشوار است، اما برخی ویژگی‌های میدان تصادفی موجب ساده‌سازی مساله خواهند شد.

تعریف ۳.۹.۱. (مانای ذاتی). میدان تصادفی $Z(.)$ را **مانای ذاتی**[۵۱] گویند هرگاه:

۱. میانگین میدان تصادفی[۵۲] ثابت یا مستقل از s باشد، به‌طوری که :

[۵۱] Intrinsic stationary

[۵۲] Random field

$$E(Z(s)) = \mu, \quad s\epsilon D$$

۲.واریانس مربوط به عبارت $(Z(s_۱) - Z(s_۲))$ نیز تنها تابعی از فاصله بین دو موقعیت $s_۱$ و $s_۲$ باشد، بنابراین واریانس را به‌صورت زیر نشان داده می‌شود:

$$Var(Z(s_۱) - Z(s_۲)) = ۲\gamma(s_۱ - s_۲), \quad s_۱, s_۲\epsilon D \subset \mathbb{R}^d$$

تعریف ۴.۹.۱. (مانای مرتبه دوم یا ضعیف). میدان تصادفی $Z(.)$ را **مانای مرتبه دوم**[۵۳] **یا ضعیف** [۵۴] نامند هرگاه:

۱. میانگین میدان تصادفی $Z(.)$ مستقل از s و ثابت باشد. به همین دلیل :

$$E(Z(s)) = \mu, \quad s\epsilon D \subset \mathbb{R}^d$$

۲. کوواریانس $Z(s_۱)$ و $Z(s_۲)$ در صورتی که فاصله تابعی از موقعیت‌های $s_۱$ و $s_۲$ باشد بنابراین،

$$Cov(Z(s_۱), Z(s_۲)) = C(s_۱ - s_۲), \quad s_۱, s_۲\epsilon D \subset \mathbb{R}^d$$

که تحت این نوع از مانایی

$$Var(Z(s), Z(s)) = Cov(Z(s), Z(s)) = C(۰) = \sigma^۲$$

که بنابر مطالب فوق میدان تصادفی به موقعیت فضایی بستگی نخواهد داشت و تغییرپذیری میدان تصادفی در همه‌جا یکسان می‌باشد.

تعریف ۵.۹.۱. (مانای قوی). میدان تصادفی $Z(.)$ را **مانای قوی** گویند، زمانی که برای تمامی موقعیت‌های $s_۱, ..., s_n$ و تاخیر h، توزیع توأم $Z(s_۱), ..., Z(s_n)$ نیز مانند توزیع توأم

$$Z(s_۱ + h), ..., Z(s_n + h)$$

باشد. بنابراین

$$(Z(s_۱), ..., Z(s_n)) \overset{D}{=} (Z(s_۱ + h), ..., Z(s_n + h))$$

در صورت انتقال موقعیت‌های $s_۱, ..., s_n$ در راستای $\{h \epsilon \mathbb{R}^d\}$، توزیع توأم $Z(s_۱), ..., Z(s_n)$ تغییر نخواهد کرد.

تذکر ۱.۹.۱. برای جمع‌بندی مطالب فوق می‌توان گفت که، مانای قوی شرط کافی برای مانای ضعیف و نیز مانای ضعیف شرطی کافی برای مانای ذاتی خواهد بود که عکس مطالب فوق صحت نخواهد داشت.

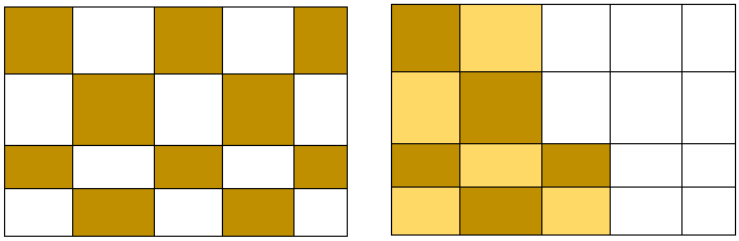
---

[۵۳] Second

[۵۴] stationary

۲.۹.۱ وارد کردن ساختار همبستگی فضایی در مدل

همان طوری که می‌دانید همبستگی فضایی یک خاصیت ذاتی است که این ویژگی در داده‌های ناحیه‌ای وجود دارد و باید این همبستگی در مطالعات لحاظ گردد. این نوع از همبستگی می‌تواند به دو صورت مثبت و منفی وجود داشته باشد. زمانی این همبستگی مثبت عمل می‌کند که نواحی نزدیک مشابه هم باشند و بالعکس زمانی این همبستگی منفی است که نواحی نزدیک به صورت غیرمشابه عمل کنند. برای مثال، در همبستگی فضایی مثبت، هنگامی که مقدار اندازه‌گیری شده‌ی متغیر مورد نظر در نواحی بزرگ باشد، نواحی مجاور، متمایل به مقدار بزرگ می‌شوند ولی برای همبستگی منفی این مقدار برای نواحی مجاور متمایل به مقدار کوچک خواهد بود. در شکل ۵.۱ این نوع از همبستگی‌ها را نشان می‌دهیم که نواحی رنگ زده شده یعنی همبستگی مثبت و نواحی به غیر این حالت یعنی همبستگی منفی. اولین گام



شکل ۵.۱: انواع همبستگی فضایی، (شکل سمت راست)همبستگی فضایی مثبت، (شکل سمت چپ) همبستگی فضایی منفی

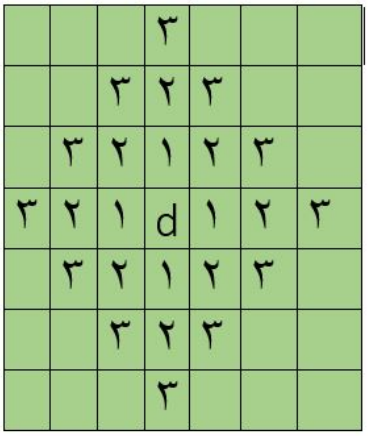
برای مدل‌بندی فضایی بین نواحی، مشخص کردن روابط بین نواحی است، برای این عمل از همسایگی بهره می‌گیرند.

تعریف ۶.۹.۱. مجموعه نواحی که بیشترین اطلاع در خصوص ناحیه‌ی دلخواه، مانند dام را ارائه دهد، مجموعه همسایه‌های ناحیه‌ی d می‌نامند و آن را با $\mathrm{N}(d)$ نشان می‌دهند.

۳.۹.۱ ماتریس همسایگی

به‌صورت‌های مختلف می‌توان همسایگی را داده‌های نواحی بیان کرد. یکی از این رویکردها به این صورت است که ناحیه‌هایی که در فاصله‌ی مشخصی از ناحیه‌ی مورد بررسی قرار دارند، به‌عنوان همسایه‌های آن ناحیه‌ی در نظر گرفت. در واقع با مشخص کردن یک مقدار در آستانه‌ی d همه‌ی ناحیه‌هایی که در فاصله‌ی γ یا کمتر از آن قرار دارند، به‌عنوان همسایه‌ی ناحیه‌ی مورد بررسی در نظر گرفته می‌شود. رویکرد دیگری از همسایگی به‌صورت راس مشترک جغرافیایی تعریف می‌شود. یعنی، همه‌ی ناحیه‌هایی که دارای راس مشترک با ناحیه‌ی مورد نظر هستند، به‌عنوان همسایه‌های آن ناحیه در نظر گرفته می‌شوند. همچنین همسایگی

می‌تواند به‌صورت تبادلات بیمه‌ای بین شهرها نیز تبیین شود، پس آن دسته از شهرهایی که با شهر d از نظر بیمه‌ای در ارتباط‌اند به‌عنوان همسایه‌های شهر d در نظر گرفته می‌شوند. همسایگی برای داده‌های ناحیه‌ای به‌صورت مرز مشترک جغرافیایی نیز است، به‌طوری که برای هر ناحیه‌ی خاص ناحیه‌هایی که دارای مرز مشترک با ناحیه‌ی مورد نظر هستند، به‌عنوان همسایه‌ی مرتبه‌ی اول آن ناحیه می‌باشد، البته به غیر از همسایگی مرتبه‌ی اول، برای هر ناحیه‌ی خاص همسایگی مرتبه دوم، سوم و مراتب بالاتر نیز تعریف می‌شود (همانند شکل۶.۱). حال بعد از اینکه همسایه‌های ناحیه‌ی مورد نظر را تعیین کردیم کافی است برای نشان دادن



شکل ۶.۱: همسایگی مرتبه ۱ تا ۳ برای ناحیه‌ی d

ارتباط بین آن‌ها از یک ماتریس همانند W با اعضای صفر و یک استفاده کنیم. برای مثال، اگر ناحیه‌ی مورد نظر d با ناحیه‌ی iام همسایه باشد، بنابراین $w_{dj} = w_{jd} = ۱$ ودر غیر این‌صورت صفر خواهد بود. به‌علاوه، برای هر ناحیه‌ی d ، $w_{dd} = ۰$ خواهد شد. برای نشان دادن مطالب گفته شده یک مثال را ارائه می‌دهیم (W را برای سه ناحیه در نظر می‌گیریم). ماتریس همسایگی طبق شکل ۷.۱ نشان داده شده است.

$$W = \begin{pmatrix} ۰ & w_{ij} & w_{ik} \\ w_{ji} & ۰ & w_{jk} \\ w_{ki} & w_{kj} & ۰ \end{pmatrix} = \begin{pmatrix} ۰ & ۱ & ۰ \\ ۱ & ۰ & ۱ \\ ۰ & ۱ & ۰ \end{pmatrix}$$

شکل ۷.۱: نواحی مورد بررسی

ماتریس فوق را ماتریس همسایگی[۵۵] می‌نامند. بنابراین نواحی که مجاور هم می‌باشند یعنی، مرز مشترک داشته باشند، به‌عنوان همسایه در نظر گرفته می‌شود. روش دیگری هم برای مشخص کردن ناحیه‌ی مورد نظر به‌کار گرفته‌اند که به‌صورت زیر خواهد بود:

$$w_{(}dj) = \begin{cases} ۱ & \gamma_{dj} < \delta \\ ۰ & o.w. \end{cases}$$

که γ_{dj} فاصله‌ی بین ناحیه‌ی d و همسایه‌ی jام است. توجه کنید که همیشه نیازی نیست که از صفر و یک برای انتساب وزن استفاده شود، پس می‌توان از تابع توانی منفی فاصله‌ی بین نواحی نیز استفاده کرد و وزن را اینگونه برای اعضای ماتریس همسایگی لحاظ کرد، در نتیجه α یک مقدار معلوم مثبت و نیز مقدار ۱ یا ۲ را می‌پذیرد.

$$w_{(}dj) = \begin{cases} \gamma_{dj}^{-\alpha} & \alpha > ۰ \\ ۰ & o.w. \end{cases}$$

از طرفی نیز می توان از تابع نمایی منفی[۵۶] به صورت زیر هم استفاده کرد.

$$w_{(}dj) = \begin{cases} exp\{-\alpha\gamma_{dj}\} & \alpha > ۰ \\ ۰ & o.w. \end{cases}$$

یکی از مشکلاتی که روش بالا وجود دارد، دارا بودن وزن برای همه‌ی همسایگی‌ها است که یکسان در نظر می‌گیرند و این منطقی نیست، زیرا، بعضی از همسایگی‌ها احتمال دارند که اطلاعات بیشتری در مورد ناحیه‌ی مورد نظر دهند و نشان‌دهنده‌ی این است که وزن بیشتری را در اختیار دارند. به عبارت دیگر، نیازی نیست ماتریس همسایگی متقارن باشد همین که وزن یکسانی به همسایگی ماتریس W داده شود کافی است. برای انتساب وزن به هر یک از این همسایگی‌ها، می‌توان از عواملی همچون مرز مشترک نواحی، جمعیت ناحیه، مساحت نواحی، طول مشترک نواحی و غیره استفاده کرد.

[۵۵] Neighborhood matrix
[۵۶] Negative exponential function

جمع‌بندی

در سال‌های اخیر، روش برآورد کوچک ناحیه‌ای به دلیل نیاز به فراهم آوردن برآوردهای قابل اعتماد آن هم برای وضعیت‌های کم‌نمونه یا بدون نمونه در زیر جوامع، مورد توجه دولت‌ها، بخش‌های اقتصادی و محققان قرار گرفته است. در این فصل مفاهیم اولیه برآورد کوچک ناحیه‌ای و الگوریتم‌های در این موضوع را بررسی کردیم. در فصل آتی، به معرفی مدل‌های کوچک ناحیه‌ای خواهیم پرداخت.

فصل ۲

معرفی مدل‌های کوچک ناحیه‌ای

در ابتدای این فصل، اهمیت برآوردگرهای غیرمستقیم و روش‌های مبتنی بر مدل‌هایی که در این روش وجود دارند را معرفی می‌کنیم. همچنین آن دسته از مدل‌هایی را که در برآورد نواحی کوچک معروف می‌باشند، شرح می‌دهیم. از جمله این مدل‌ها می‌توان مدل فی‌ـ‌هریوت را نام برد. و در انتهای فصل مدل‌های آمیخته خطی [۱] (LMM) و خطی تعمیم‌یافته‌ی فضایی[۲] (SGLMM) را با ذکر مثال‌هایی بیان می‌کنیم. مطالب ذکر شده در این فصل عمدتا برگرفته از کتاب برآورد کوچک ناحیه‌ای رائو (۲۰۰۳) و مقاله میلان کارگانیس (۲۰۰۹) است.

۱.۲ برآوردگرهای کوچک ناحیه‌ای

بنابر مطالب بیان شده در فصل اول، نمی‌توان از برآوردگرهای مستقیم برای برآورد پارامتر نواحی کوچک استفاده کرد. بنابراین، به برآوردگرهای غیرمستقیم توجه بیشتری می‌شود، زیرا این دسته از برآوردگرها دقت‌شان نسبت به برآوردگرهای مستقیم بیشتر است و برآوردهای قابل اطمینان و با کیفیت بهتری را ارائه می‌دهد. از طرفی برای برآورد پارامترهای نواحی کوچک می‌توان از طرح نمونه‌گیری استفاده کرد. اما اگر این طرح‌ها مبنی بر طرح برآوردگرهای مستقیم باشند، مستلزم هزینه‌های مالی و زمانی خواهند بود.

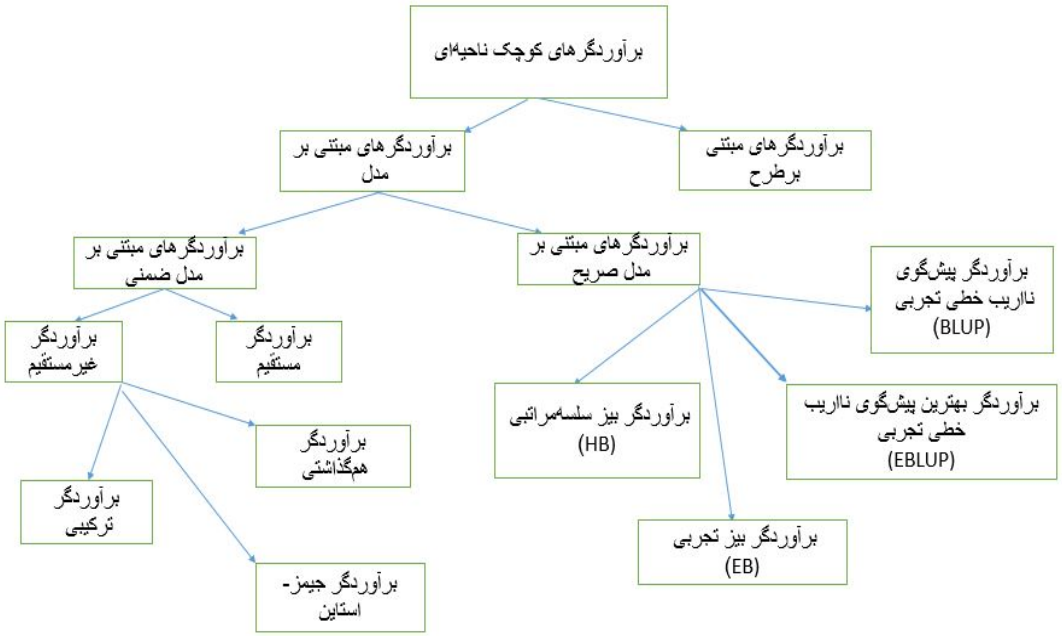
---

[۱] Linear mixed models

[۲] Spatial generalized Linear mixed models

به این منظور برای اینکه بتوان به برآوردگرهای نواحی کوچک دسترسی داشت برآوردگرهای مبتنی بر مدل را جایگزین برآوردگرهای مستقیم برطرح کرده‌اند. در روش مبتنی بر مدل لزوما از یک مدل آماری استفاده می‌کنند، بنابراین این مدل برای اینکه بتواند به برآوردگر یک ناحیه‌ی کوچک دسترسی داشته باشد، تنها باید از اطلاعات داده‌های نمونه‌گیری شده‌ی سایر نواحی کوچک و یا دوره‌های زمانی دیگر بهره بگیرد. بنابراین دو مورد از عواملی که کارایی برآوردگرهای ناحیه‌ی کوچک به آن بستگی خواهد داشت را می‌توان یکی متغیرهای پیشگو و دیگری انتخاب یک مدل مناسب برشمرد، زیرا در دسترس بودن داده‌های کمکی مناسب در سطح ناحیه یا در سطح واحد حائز اهمیت است. مدل‌های آماری بنابر ماهیت داده‌ها و داده‌های کمکی قابل دستیابی هستند.

با توجه به مطالب گفته شده، جدول زیر را برای آشنایی بیشتر نسبت به شاخص روش‌های برآورد کوچک ناحیه‌ای مطرح کرده‌ایم.



شکل ۱.۲: نمودار انواع برآوردگرهای کوچک ناحیه‌ای

سعی بر این است، مطالبی درباره‌ی برآوردگرهایی مبتنی برمدل را ارائه دهیم.

برآوردگرهای مبتنی بر مدل

بنابر مطالب گفته شده، برآوردگرهای مبتنی بر طرح به‌دلیل دارا بودن واریانس بزرگ برای برآورد پارامتر نواحی کوچک مناسب نمی‌باشند. در روش مبتنی بر مدل، یک توزیع احتمال را روی متغیر تصادفی Y قرار می‌دهند. عملکرد این روش بر مبنای نمونه‌گیری جامعه متناهی از یک جامعه فرضی، متناهی است که با یک مدل تصادفی مشخص شده است. در برآوردگر مبتنی بر مدل برای اینکه بتوان دقت برآوردگر را در برآورد کوچک ناحیه‌ای افزایش داد و همچنین میزان واریانس برآوردگر کوچک ناحیه‌ای را کاهش داد کافی است از اطلاعات سایر نواحی و یا از اطلاعات سایر دوره‌های زمانی استفاده کرد.

در برآوردگرهای غیرمستقیم مبتنی بر مدل، اطلاعات مورد نیاز مربوط به مقادیر متغیرهای مورد مطالعه و کمکی نواحی و یا دوره‌های زمانی است و تمامی این مقادیر با به‌کارگیری از یک مدل ربط (ضمنی یا صریح) در برآورد پارامتر وارد می‌شود. مدل‌های ربط شامل دو نوع می‌باشند، یکی مدل‌های صریح و دیگری مدل ضمنی، برآوردگرهای مبتنی بر مدل شامل دو دسته‌ی کلی برآوردگرهای مبتنی بر مدل ضمنی و مبتنی بر مدل صریح هستند. در مدل ضمنی، به‌طور غیرصریح از ساختار موجود در جامعه استفاده می‌شود و تغییرپذیری نواحی کوچک را شامل نمی‌شود. با توجه به این مدل‌ها، فرض بر این است، که تمامی نواحی نسبت به متغیر مورد بررسی به‌طور مشابه رفتار می‌کند و تغییرپذیری در میان نواحی کوچک وجود ندارد. ولی در مدل‌های صریح از روابط بین متغیرهای مورد بررسی و اطلاعات کمکی و ترکیب آن‌ها با اثرهای تصادفی ناحیه‌ی مشخص، به‌صورت صریح به‌کار گرفته می‌شود.

برآوردگرهای مبتنی بر مدل ضمنی

این نوع از برآوردگرها را برآوردگرهای غیرمستقیم سنتی نیز می‌نامند. همان‌طوری که قبلا اشاره شد، این نوع از برآوردگرها شامل سه دسته می‌باشند. برای جزئیات بیشتر می‌توانید به کتاب رائو (۲۰۰۳) مراجعه کنید.

۲.۲ معرفی مدل‌های کوچک ناحیه‌ای

به‌منظور اتخاذ اثرهای ناحیه‌ای و آگاهی از تغییرات بین ناحیه‌ای، مدل‌های ضمنی کمتر مورد استفاده قرار می‌گیرند و مدل‌های صریح جایگزین شده‌اند. معمولا مدل‌های صریح با به‌کارگیری مدل‌های آمیخته‌ی خطی، اثرهای تصادفی ناحیه‌ای را به حساب می‌آورند. در واقع مدل‌های صریح کاراتر از مدل‌های ضمنی هستند زیرا، علاوه بر اتخاذ داده‌های کمکی تغییرات بین ناحیه‌ای را فراتر از آنچه که توسط اطلاعات کمکی موجود در مدل تبیین می‌شوند به حساب می‌آورند. مدل‌های صریح کوچک ناحیه‌ای را می‌توان به دو دسته‌ی زیر تقسیم کرد: «مدل در سطح واحد آماری[۳]» و «مدل در سطح ناحیه[۴]»

[۳] Unit level model

[۴] Area level model

● مدل در سطح واحد آماری

این مدل‌ها مقادیر واحد متغیر تحت مطالعه را با متغیرهای کمکی خاص آن واحد معین ربط می‌دهند. در این مدل‌ها، واحدهای قابل مشاهده، یک واحد انفرادی در ناحیه هستند و نواحی اغلب اثر خوشه‌ای دارند یعنی مشاهدات واحدهای انفرادی درون یک ناحیه، نسبت به کل مشاهدات به یکدیگر نزدیک‌تراند. این مدل‌ها نیز انواع گوناگونی دارند. هنگامی که مقادیر متغیرهای کمکی واحد جامعه در دسترس باشند و بتوان روی آن‌ها مدل‌سازی کرد، این مدل‌ها به‌کار گرفته می‌شوند. این نوع مدل‌ها مقادیر واحد انفرادی از متغیر تحت مطالعه را با متغیرهای کمکی آن واحد معین پیوند می‌دهند.

● مدل در سطح ناحیه

مدل‌هایی که برآوردگرهای مستقیم ناحیه‌ی کوچک را با متغیرهای کمکی ناحیه‌ی معین ربط می‌دهند، مدل در سطح ناحیه می‌نامند. زمانی که داده‌های مربوط به سطح واحد در دسترس نباشند، استفاده از این مدل‌ها ضرورت پیدا می‌کند. این مدل‌ها برآوردگرهای مستقیم ناحیه‌ی کوچک را با متغیرهای کمکی سطح ناحیه‌ی مورد نظر پیوند می‌دهند. در این مدل‌ها واحدهای قابل مشاهده برای متغیرهای پاسخ و کمکی ناحیه‌ها می‌باشند و اثرهای تصادفی ناحیه نیز به‌عنوان مانده‌های مدل خطی معین شده برای متغیر پاسخ هستند. به‌طور کلی یک مدل در سطح ناحیه دو قسمت دارد، یک قسمت مربوط به مدل نمونه‌گیری برای برآوردگرهای مستقیم که مبتنی بر طرح است و قسمت دوم مدل ربطی است که پارامترهای ناحیه‌ی کوچک را به متغیرهای کمکی در سطح ناحیه ربط می‌دهد که معمولا از نوع رگرسیونی است. البته لازم به ذکر است در صورت وجود اطلاعات کمکی در سطح واحد نیز می‌توان مدل در سطح ناحیه را به‌کار برد.

مثال ۱.۲.۲. برای مثال، فرض کنید در برآورد نرخ بیکاری ناحیه‌ی کوچک اطلاعات مربوط به تعداد افراد باسواد خانوار به عنوان داده‌ی کمکی در سطح خانوارها (واحد آماری) در ناحیه‌ی مورد نظر در دسترس است. لذا با میانگین‌گیری می‌توان داده‌ی کمکی را در سطح ناحیه‌ی مورد نظر بوجود آورد.

روش‌های سنتی غیرمستقیم بر پایه‌ی مدل‌های ضمنی استوارند که از طریق داده‌های تکمیلی، پیوندی با ناحیه‌های کوچک مرتبط با ناحیه‌ی مربوطه برقرار می‌سازند. حال اگر به مدل‌های کوچک ناحیه‌ای صریح رجوع کنید می‌توانید به این نتیجه برسید که برای تغییرات بین ناحیه‌ای فرصتی خاصایجاد شده است. کاربست مدل‌های صریح چند مزیت خواهد داشت که عبارت‌اند از

۱. برای پی بردن مدل(هایی) مناسب که به داده‌ها خوب برازش دهد، می‌توان روش‌های تشخیصی مدل را به‌کار گرفت. این نوع از روش‌های تشخیصی شامل تحلیل مانده‌ها برای کشف انحراف‌ها از مدل مفروض، گزینش متغیرهای کمکی برای مدل، و مباحث تشخیصی حذف برخی مشاهده‌ها برای کشف مشاهده‌ها موثر است.

۲. می‌توان معیارهای دقت ناحیه ویژه را برای هر برآورد کوچک ناحیه‌ای به‌دست آورد. این مطلب برتری نسبتا خوبی را به معیارهای کلی که روی ناحیه‌های کوچک و متوسط قرار دارند، و اغلب اوقات با برآوردهای هم‌گذاشتی به‌کار می‌روند.

۳. مدل‌های آمیخته‌ی خطی و نیز مدل‌های غیرخطی، همچون مدل‌های رگرسیون لجستیک[۵] و مدل‌های خطی تعمیم‌یافته[۶] شامل اثرهای تصادفی ناحیه‌ای هستند که می‌توانند مورد بررسی قرار گیرند. با این مدل‌ها می‌توان ساختار داده‌های پیچیده را همانند وابستگی فضایی را که یکی از اهداف اصلی این پایان‌نامه است، وارد مدل‌بندی کرد.

۴. پیشرفت‌های روش‌شناختی اخیر برای مدل‌هایی که شامل اثرهای تصادفی هستند راهکاری برای استنباط‌های دقیق کوچک ناحیه‌ای می‌باشد.

دقت کنید که می‌توان مدل‌های گوناگونی برای برآورد کوچک ناحیه‌ای انتخاب کرد. اما مدل باید با مشورت متخصصان موضوع مورد مطالعه و کاربران نهایی این حیطه انتخاب شود. آن‌ها نقش موثری در انتخاب متغیرهای کمکی نیز دارند. لذا پرکاربرد بودن هر روش مدل پایه به در دسترس بودن داده‌های کمکی خوب بستگی دارد. بنابراین باید به جمع‌آوری آن دسته از متغیرهای کمکی که پیش‌گویی‌های خوبی از متغیرهای مورد بررسی دارند، توجه بیشتری شود. اکنون به شرح مختصری از مدل‌های کوچک ناحیه‌ای می‌پردازیم.

۳.۲ مدل پایه‌ای در سطح ناحیه

فرض کنید که $\theta_i = g(\bar{Y}_i)$ به‌ازای تابعی مشخص مانند $g(.)$ با داده‌های کمکی ناحیه ویژه‌ی $\mathbf{z}_i = (z_{1i}, ..., z_{pi})^T$ از طریق مدل خطی زیر مرتبط است:

$$\theta_i = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i, \qquad i = 1, ..., m \tag{۱.۲}$$

که در آن b_iها مقادیر ثابت مثبت و معلوم‌اند و $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ برداری با ابعاد $p \times 1$ از ضریب‌های رگرسیونی است. به‌علاوه، که، v_iها اثرتصادفی ناحیه ویژه هستند، که فرض می‌شوند مستقل و هم‌توزیع (i.i.d) با

$$E_m(v_i) = 0, \qquad V_m(v_i) = \sigma_v^2 (\geqslant 0) \tag{۲.۲}$$

می‌باشند، که در آن E_m امید ریاضی مدل و V_m واریانس مدل را نشان می‌دهند. این فرض به‌صورت $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ است. نرمال بودن اثرهای تصادفی v_i نیز اغلب اوقات در نظر گرفته می‌شود. پارامتر σ_v^2 معیاری از همگنی ناحیه‌ها پس از به‌حساب آوردن متغیرهای کمکی $\mathbf{z}_i$ است.

---

[۵] Logistic regression models

[۶] Generalized linear mixed models

در بعضی کاربردها، همه‌ی ناحیه‌ها به عنوان عضو نمونه انتخاب نمی‌شوند. فرض کنید M ناحیه در جامعه داریم و تنها m ناحیه در نمونه انتخاب می‌شوند، $m < M$ مدلی را به مدل (۱.۲) برای جامعه فرض می‌کند. به‌علاوه فرض کنید که ناحیه‌های نمونه از مدل جامعه پیروی می‌کنند، یعنی اریبی در گزینش ناحیه‌های نمونه‌ای وجود ندارد به‌طوری که در (۱.۲) برای ناحیه‌های نمونه‌گیری شده صادق است.

برای انجام استنباط‌هایی درباره‌ی میانگین‌های کوچک ناحیه‌ای[۷] $\bar{Y}_i$ تحت مدل (۱.۲)، فرض کنید که برآوردگرهای مستقیم $\hat{\bar{Y}}_i$ قابل دسترس باشد. حال فرض بر این است که

$$\hat{\theta}_i = g(\hat{\bar{Y}}_i) = \theta_i + e_i, \qquad i = ۱, ..., m \tag{۳.۲}$$

در آن e_iها خطاهای نمونه‌گیری مستقل هستند و

$$E_p(e_i|\theta_i) = ۰, \quad V_p(e_i|\theta_i) = \Psi_i \tag{۴.۲}$$

همچنین مرسوم می‌باشد. پس، فرض می‌شود که واریانس‌های نمونه‌گیری، Ψ_i، معلوم‌اند. فرض‌های بالا ممکن است در برخی کاربردها کاملا محدود کننده باشد. مثلا، اگر $g(.)$ تابع غیرخطی و اندازه‌ی نمونه‌ی ناحیه n_i کوچک باشد، برآوردگر مستقیم $\hat{\theta}_i$ ممکن است طرح اریب باشد. اگر ناحیه‌های کوچک طرح نمونه‌گیری را قطع کنند، خطاهای نمونه‌گیری ممکن است وابسته باشد. فرض واریانس معلوم Ψ_i را می‌توان با برآورد Ψ_i از داده‌های نمونه‌ای در سطح واحد آماری و سپس هموارسازی واریانس‌های برآورد شده‌ی $\hat{\Psi}_i$ برای به‌دست آوردن برآوردی پایدارتر از Ψ_i، تخفیف داد. نرمال بودن $\hat{\theta}_i$ نیز اغلب فرض می‌شود، اما به‌واسطه‌ی تاثیر قضیه‌ی حد مرکزی بر $\hat{\theta}_i$، این فرض ممکن است به اندازه‌ی فرض نرمال بودن اثرهای تصادفی محدود کننده نباشد. مدل‌های یقینی برای θ_i با قرار دادن $\sigma_v^۲ = ۰$ به‌دست می‌آیند، یعنی $\theta_i = \mathbf{z}_i^T\boldsymbol{\beta}$. چنین مدل‌هایی به برآوردگرهای هم‌گذاشتی می‌انجامند که تغییرات محلی را به غیر از تغییرات بازتاب‌یافته در متغیرهای کمکی $\mathbf{z}_i$، در نظر نمی‌گیرند.

با استفاده از ترکیب دو مدل (۱.۲) و (۳.۲) مدل زیر به‌دست می‌آید:

$$\hat{\theta}_i = \mathbf{z}_i^T\boldsymbol{\beta} + b_i v_i + e_i, \qquad i = ۱, ..., m \tag{۵.۲}$$

با توجه به این‌که مدل (۵.۲) شامل خطاهای ایجادشده از طرح e_i و نیز خطاهای مدل v_i است. فرض کنید e_i و v_i مستقل‌اند. مدل (۵.۲) حالت خاصی از مدل آمیخته‌ی خطی است.

اگر اندازه‌ی نمونه‌ای n_i در ناحیه‌ی iام کوچک باشد و θ_i تابعی غیرخطی از مجموع Y_i باشد، حتی اگر برآوردگر مستقیم $\hat{Y}_i$ طرح نااریب باشد، فرض $E_p(e_i|\theta_i) = ۰$ در مدل نمونه‌گیری (۳.۲) ممکن است معتبر نباشد. مدلی واقعی‌تر با فرمول

$$\hat{Y}_i = Y_i + e_i^*, \qquad i = ۱, ..., m \tag{۶.۲}$$

[۷] Small area means

و $E_P(e_i^*|Y_i) = 0$ مطرح می‌شود، یعنی $\hat{Y}_i$ برای مجموع Y_i طرح ناریب است.

در این حالت، مدل‌های نمونه‌گیری و پیونددهنده با هم جور نیستند. در نتیجه نمی‌توان مدل (۶.۲) را با مدل پیونددهنده‌ی (۱.۲) ترکیب کرد تا مدل آمیخته‌ی خطی به شکل مدل (۳.۲) به‌دست آید. ضمن این‌که نتایج استاندارد نظریه‌ی مدل‌های آمیخته‌ی خطی، کاربرد ندارد (رائو، ۲۰۰۳). روش پیشنهادی آماردانان‌ها برای حل این مشکل، با استفاده از رویکرد بیزی سلسه‌مراتبی[۸] HB برای اداره کردن مدل‌های ناجور نمونه‌گیری و پیونددهنده است.

مثال ۱.۳.۲. (شمار افراد فقیر). مدل پایه‌ای در سطح ناحیه (۵.۲) اخیرا برای تولید برآوردهای شهرستانی مدل پایه از کودکان فقیر در سن تحصیل در ایالت‌های متحده به‌کار رفته است (شورای پژوهش ملی، ۲۰۰۰). با استفاده از این برآوردها، وزارت آموزش و پرورش ایالت‌های متحده سالانه بیش از ۷ میلیارد دلار از بودجه‌ی عمومی را به شهرستان‌ها تخصیص می‌دهد، و سپس ایالت‌ها این بودجه را در بین منطقه‌های مدرسه‌ای توزیع می‌کنند. در گذشته، بودجه‌ها بر اساس شمارهای برآوردشده از سرشماری قبلی تخصیص می‌یافتند، اما شمار افراد فقیر در طی زمان به‌طور معنی داری تغییر کرده است.

در این کاربرد، $\theta_i = log(Y_i)$، که در آن Y_i شمار واقعی افراد فقیر شهرستان (کوچک ناحیه‌ی) iام است. برآوردهای مستقیم $\hat{Y}_i$ به‌صورت میانگین موزون سه‌ساله‌ی کودکان فقیر در سن تحصیل (زیر ۱۸ سال) از نشریه‌ی تکمیلی ماه مارس آمارگیری جمعیت جاری (CPS) به‌دست آمدند. متغیرهای پیشگو در سطح ناحیه، $\mathbf{z}_i$، از داده‌ی ثبتی به‌دست آمده است. شهرستان‌های با نمونه‌ی (CPS) ولی، بدون کودکان فقیر در سن تحصیل کنار گذاشته شده است، زیرا $\hat{Y}_i$ و $\hat{\theta}_i = log\hat{Y}_i = -\infty$ می‌باشد. برای شهرستان‌هایی که در نمونه‌ی (CPS) انتخاب نشدند، تنها متغیرهای کمکی $\mathbf{z}_i$ موجودند.

۱.۳.۲ گسترش‌های مدل در سطح ناحیه (A)

شکل (۱.۲) بعضی از مدل‌های معروف آماری را تا حدودی نشان می‌دهد. اما این مدل‌ها تنها بخشی از کل مطالبی است که معرفی نموده‌ایم، برای آشنایی بیشتر می‌توانید به کتاب رائو (۲۰۰۳) مراجعه کنید. گسترش مدل در سطح ناحیه شامل پنج دسته کلی که عبارت‌اند از؛ مدل فی‌ـ‌هریوت، مدل فی‌ـ‌هریوت چندمتغیره، مدل با خطاهای نمونه‌گیری همبسته، مدل‌های سری زمانی و مقطعی و مدل‌های فضایی است که در بخش‌های زیر به‌طور مفصل شرح خواهیم داد.

مدل فی‌ـ‌هریوت

مدل فی‌ـ‌هریوت یکی از مدل‌های موجود در سطح ناحیه است، این مدل نخستین بار توسط فی و هریوت در سال (۱۹۷۹) مطرح شد. ترکیب این مدل شامل مدل پیونددهنده و مدل

[۸] Hierarchical bayes

نمونه‌گیری می‌باشد که در آن مدل نمونه‌گیری نشان‌دهنده‌ی ارتباط بین پارامتر مورد نظر کوچک ناحیه‌ای و برآوردگر مستقیم مرتبط به آن است، همچنین مدل پیونددهنده به کمک یک مدل رگرسیونی در بین پارامتر کوچک ناحیه‌ای و متغیرهای کمکی در سطح ناحیه پیوند برقرار می‌کند. حال اگر نواحی بزرگ مورد نظر دارای طرح نمونه‌گیری متناسب باشد و شامل M نواحی کوچک باشد، مدل پیونددهنده درباره‌ی هریک از این M نواحی صدق می‌کند (پراتسی و سالواتی، ۲۰۰۶). بنابراین تنها m نواحی کوچک در میان سایر نواحی نمونه‌گیری قرار می‌گیرد، در این صورت فرض بر این است که اریبی انتخاب نمونه وجود ندارد و این m نواحی کوچک از یک مدل پیونددهنده‌ی موجود در جامعه پیروی می‌کند. و برای آشنایی با این مدل می‌توان آن را به‌صورت زیر نمایش داد:

$$\theta_i = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i; \qquad i = 1, ..., m \tag{۷.۲}$$

که θ_i پارامتر مورد نظر ناحیه‌ی iام و $\acute{\mathbf{z}}_i = (z_{i1}, z_{i2}, ..., z_{ip})$ بردار p بعدی از متغیرهای کمکی ناحیه‌ی iام است و b_i یک مقدار ثابت معلوم و $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ بردار $p \times 1$ از ضرایب رگرسیونی می‌باشد و عناصر آن نیز نامعلوم و ثابت هستند. اثر تصادفی ناحیه‌ی v_i است که مستقل و هم‌توزیع فرض می‌شود (i.i.d) بنابراین،

$$E(v_i) = 0, \qquad V(v_i) = \delta_v^2.$$

توزیع v_iها را نرمال در نظر می‌گیرند، پس $v_i \overset{iid}{\sim} N(0, \delta_v^2)$ و v_iها و e_iها نیز مستقل از هم هستند. مدل نمونه‌گیری هم به صورت زیر است:

$$\hat{\theta}_i = g(\hat{\bar{Y}}_i) = \theta_i + e_i; \qquad i = 1, ..., m. \tag{۸.۲}$$

در مورد مدل فوق می‌توان گفت که $\hat{\theta}_i$ برآوردگر مستقیم مرتبط به θ_i و e_i خطای نمونه‌گیری و از هم مستقل‌اند لذا بنابر $E_p(e_i|\theta_i) = 0$ و $V_p(e_i|\theta_i) = \Psi_i$ می‌توان گفت، برآوردگر مستقیم θ_i طرح ناریب است و به‌صورت تابعی از میانگین طرحی ناحیه‌ی کوچک iام یعنی $\hat{\bar{Y}}_i$ در نظر گرفته می‌شود. برای میانگین طرحی کوچک ناحیه‌ای در ناحیه‌ی iام معمولا از برآوردگر هورویتز– تامسون استفاده می‌کنند که این نوع از برآوردگر دارای ویژگی طرح ناریبی است.

Ψ_iها در واقع همان واریانس‌های نمونه‌گیری‌اند که معلوم فرض می‌شوند. بنابر شرایط مطرح شده، محدودیت در بعضی مسائل کاربردی وجود دارد. زمانی $\hat{\theta}_i$ها برای θ_iها می‌تواند طرح اریب باشد، که به عنوان مثال، اگر g یک تابع غیرخطی باشد، آن‌گاه n_i کوچک می‌شود. اگر خطاهای نمونه‌گیری (منظور همان e_iها) نسبت به هم‌وابسته باشند و نواحی کوچک، طرح نمونه‌گیری را قطع کنند، بایستی از سایر مدل‌ها استفاده کرد. اشاره به این نکته لازم است که با فرض معلوم بودن Ψ_iها و نرمال بودن برآوردگر مستقیم، $\hat{\theta}_i$ها محدودکننده نمی‌باشند، زیرا به‌جای e_iها می‌توان از برآورد هموارشده‌ی آن‌ها که از طریق مشاهدات نمونه‌ای حاصل می‌شوند، استفاده کرد و علاوه بر این برآوردگرهای $\hat{\theta}_i$ها یک ترکیب خطی از $\hat{\bar{Y}}_i$ها می‌باشند.

بنابر قضیه‌ی حد مرکزی می‌توان نتیجه گرفت که $\hat{\bar{Y}}_i$ها به دنبال $\hat{\theta}_i$ ها هستند و از توزیع نرمال پیروی می‌کنند که هر کدام از $\hat{\theta}_i$ها به صورت حاشیه‌ای دارای توزیع نرمال خواهند بود. در انتها با ترکیب مدل‌های (۷.۲) و (۸.۲) می‌توان مدل نهایی را به صورت زیر نمایش داد:

$$\hat{\theta}_i = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i + e_i, \qquad i = ۱, ..., m. \tag{۹.۲}$$

مدل گفته شده در بالا، یک صورت خاص از یک مدل آمیخته خطی می‌باشد. اگر $E_p(e_i|\theta_i) = ۰$ در مدل (۸.۲) برقرار نباشد، برآوردگر مستقیم $\hat{\theta}_i$ برای θ_i طرح اریب می‌شود، همانند زمانی که g یک تابع غیرخطی از یک مجموعه در ناحیه‌ی کوچک iام باشد و n_i کوچک باشد، در این صورت می‌توان گفت که اگر برآوردگر مستقیم $\hat{Y}_i$ برای Y_i طرح نااریب باشد آن‌گاه $\hat{\theta}_i$ برای θ_i طرح اریب خواهد بود. و به‌همین دلیل مدل نمونه‌گیری زیر را به جای مدل (۸.۲) به کار می‌برند. یعنی:

$$\hat{Y}_i = Y_i + e_i^* \tag{۱۰.۲}$$

اگر $E_p(e_i^*|\theta_i) = ۰$ برقرار باشد، برآوردگر مستقیم $\hat{Y}_i$ برای Y_i یک برآوردگر طرح نااریب خواهد شد. به همین دلیل مدل (۱۰.۲) یک مدل نمونه‌گیری است و با مدل (۷.۲) که یک مدل پیونددهنده است قادر به ترکیب شدن نمی‌باشند پس، از بوجود آوردن مدل ترکیبی (۹.۲) معذور خواهیم بود. در نتیجه یک مدل آمیخته‌ی خطی و نیز امکان دسترسی به پیشگوی EBLUP را نخواهیم داشت و این به منزله استفاده از مدل سلسله مراتبی HB خواهد بود.

مدل فی‌ـ هریوت چندمتغیره

در نظر بگیرید که بردار $r \times ۱$ از برآوردگرهای آمارگیری $\hat{\boldsymbol{\theta}}_i = (\hat{\theta}_{i۱}, ..., \hat{\theta}_{ir})^T$ باشد، پس خواهیم داشت:

$$\hat{\boldsymbol{\theta}}_i = \boldsymbol{\theta_i} + \mathbf{e_i}, \qquad i = ۱, ..., m \tag{۱۱.۲}$$

و $\boldsymbol{\theta_i} = (\theta_{i۱}, ..., \theta_{ir})^T$ است که $\theta_{ij} = g_i(\bar{Y}_{ij})$ به‌ازای $j = ۱, ..., r$ برقرار می‌باشد. خطاهای نمونه‌گیری $\mathbf{e_i} = (e_{i۱}, ..., e_{ir})^T$ برداری با متغیرهای مستقل نرمال r بعدی و دارای توزیع $N_r \sim (۰, \boldsymbol{\Psi_i})$ با میانگین ۰ و ماتریس‌های کوواریانس معلوم $\boldsymbol{\Psi_i}$ به شرط $\boldsymbol{\theta_i}$ است. در اینجا ۰ بردار r بعدی و $\hat{Y}_{ij}$ میانگین ناحیه‌ی کوچک iام برای مشخصه‌ی jام است. همچنین فرض کنید که $\boldsymbol{\theta_i}$ با داده‌های کمکی ناحیه ویژه‌ی $\mathbf{z_{ij}}$ از طریق مدل خطی

$$\boldsymbol{\theta_i} = \mathbf{Z_i}\boldsymbol{\beta} + \mathbf{v_i}, \qquad i = ۱, ..., m \tag{۱۲.۲}$$

ارتباط دارد که در آن اثرهای تصادفی ناحیه ویژه‌ی $\mathbf{v_i}$ متغیرهای $N_r(۰, \boldsymbol{\Sigma_v})$ و مستقل‌اند، ماتریس $\mathbf{Z_i}$ از مرتبه‌ی $r \times rp$ است که سطر iام آن با $(۰^T, ..., ۰^T, \mathbf{z_{ij}}^T, ۰^T, ..., ۰^T)$ و $\boldsymbol{\beta}$ بردار rp

بعدی ضریب‌های رگرسیونی است. لذا بردار صفر دارای ابعاد $p \times ۱$ و $\mathbf{Z_{ij}}$ در مکان jام از بردار سطری (سطر jام) واقع است.

با ترکیب مدل‌های (۱۱.۲) و(۱۲.۲)، مدل آمیخته‌ی خطی چند متغیره‌ی

$$\hat{\boldsymbol{\theta}}_i = \mathbf{Z_i}\boldsymbol{\beta} + \mathbf{v_i} + e_i \qquad (۱۳.۲)$$

به‌دست می‌آید. مدل (۱۳.۲) نوعی از مدل فی‌ـهریوت (۵.۲) است زمانی که $b_i = ۱$ باشد. فی (۱۹۸۷) و داتا و همکاران (۱۹۹۱) گسترش چند متغیره‌ی (۷.۲) را پیشنهاد کردند و نشان دادند این مدل می‌تواند به برآوردگرهای کاراتر از میانگین‌های کوچک ناحیه‌ای $\bar{Y}_{ij}$ منجر شود زیرا، برخلاف مدل تک متغیره‌ی (۵.۲)، از هم‌بستگی بین مولفه‌های $\hat{\boldsymbol{\theta}}_i$ بهره می‌گیرد. همچنین داتا و همکاران (۱۹۹۱) مدل چندمتغیره‌ی (۱۳.۲) را برای برآورد میانه‌ی درآمد جاری مربوط به خانوارهای چهارنفری در هریک از ایالت‌های متحده به‌کار بستند.

مدل با خطاهای نمونه‌گیری هم‌بسته

یک گسترش طبیعی از مدل فی‌ـهریوت (۵.۲) عبارت است از مدلی با خطاهای نمونه‌گیری همبسته e_i. فرض کنید $\hat{\boldsymbol{\theta}} = (\hat{\theta}_۱, ..., \hat{\theta}_m)^T$، $\boldsymbol{\theta} = (\theta_۱, ..., \theta_m)^T$ و $\mathbf{e} = (e_۱, ..., e_m)^T$ برقرار باشد، بنابراین مدل فی‌ـهریوت زیر

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \mathbf{e}, \qquad (۱۴.۲)$$

دارای توزیعی با $\mathbf{e}|\boldsymbol{\theta} \sim N_m(0, \boldsymbol{\Psi})$، که در آن ماتریس کوواریانس دارای خطاهای نمونه‌گیری $\boldsymbol{\Psi} = (\psi_{ij})$ است. بعد از ترکیب شدن مدل (۱۴.۲) با مدل (۱.۲) برای θ_iها تعمیمی از مدل فی‌ـهریوت (۵.۲) را به‌دست می‌آورند. لذا اگر به‌ازای تمامی iها در مدل (۵.۲) $b_i = ۱$، جای‌گذاری شود، بنابراین مدل ترکیبی به‌دست می‌آید که می‌توان آن را به صورت زیر نوشت:

$$\hat{\boldsymbol{\theta}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v} + \mathbf{e} \qquad (۱۵.۲)$$

با توجه به مدل فوق $\mathbf{v} = (v_i, ..., v_m)^T$ و $\mathbf{x}$ ماتریسی با بعد $m \times p$ با سطر iام در $\mathbf{x_i^T}$ است. اما در عمل، به‌جای $\boldsymbol{\Psi}$ یک برآوردگر آمارگیری $\hat{\boldsymbol{\Psi}}$ یا برآوردگری هموار شده جانشین می‌شود، اما در بعضی موارد تغییرپذیری مربوط به این برآوردگر اغلب نادیده گرفته می‌شود.

مثال ۲.۳.۲. (کم‌شماری سرشماری). در ساختار برآورد مقدار کم‌شماری در سرشماری ۱۹۹۰ ایالت‌های متحده، جامعه به ۳۵۷ پساطبقه مرکب از ۵۱ گروه پساطبقه‌ای تقسیم شد و هر یک از آن‌ها به ۷ دسته‌ی سنی‌ـجنسی تقسیم شدند. تعداد ۵۱ گروه پساطبقه‌ای بر اساس نژاد یا قومیت، تصرف مسکن (مالک، مستاجر)، نوع ناحیه و منطقه تعریف شدند. برآوردهای

دوگان سامانه $\hat{\theta}_i$ برای هر پساطبقه $i = ۱, ..., ۳۵۷$ با استفاده از داده‌های حاصل از آمارگیری پساشمارشی (PES) سال ۱۹۹۰ به‌دست آمدند. در اینجا $\hat{\theta}_i$ ضریب تعدیل برآورد شده برای پساطبقه‌ی iام است. مدلی به شکل (۱۵.۲) برای به‌دست آوردن برآوردهای هموارشده‌ی ضریب‌های تعدیل θ_i به کار گرفته شد (ایساکی و همکاران، ۲۰۰۰).

مدل‌های سری زمانی و مقطعی

سری زمانی به مجموعه‌ای از مشاهدات کمی گفته می‌شود که در فواصل زمانی و به صورت متوالی اندازه‌گیری می‌شود. به بیان دیگر، سری زمانی مجموعه‌ای از مشاهدات یک متغیر است که در نقاط گسسته‌ای از زمان، که معمولا فاصله‌های مساوی دارند، اندازه‌گیری و بر حسب زمان مرتب شده‌اند. بنابراین یک سری زمانی از مشاهده یک پدیده در طول زمان به دست می‌آید. بنابراین، بسیاری از آمارگیری‌های نمونه‌ای با جانشین کردن پاره‌ای از عنصرهای نمونه‌ای، در طی زمان تکرار می‌شوند. مثلا، در آمارگیری جمعیت (CPS) ماهانه‌ی ایالت‌های متحده یک خانوار انفرادی به مدت چهار ماه پیاپی در نمونه باقی می‌ماند. سپس برای هشت ماه بعدی از نمونه خارج می‌شود، و برای چهار ماه بعدی دیگر به نمونه باز می‌گردد. در آمارگیری نیروی کار (LFS) ماهانه‌ی کانادا، هر خانوار انفرادی به مدت شش ماه پیاپی در نمونه می‌ماند و سپس از نمونه بیرون می‌رود. در مورد چنین آمارگیری‌های مکرر با وام گرفتن قرضی هم از ناحیه‌های کوچک و هم از دوره‌های زمانی، می‌توان افزایش قابل ملاحظه‌ای در کارایی برآوردگرها به‌دست آورد.

رائو و یو (۱۹۹۲، ۱۹۹۴) گسترشی از مدل پایه‌ای فی‌ـ‌هریوت (۵.۲) را برای اداره‌ی داده‌های سری زمانی و مقطعی پیشنهاد کردند. مدل آن‌ها شامل یک مدل خطای نمونه‌گیری

$$\hat{\theta}_{it} = \theta_{it} + e_{it}, \qquad t = ۱, ..., T; i = ۱, ..., m \tag{۱۶.۲}$$

و مدل پیوند دهنده‌ی

$$\hat{\boldsymbol{\theta}}_{it} = \mathbf{z}_{it}^T \boldsymbol{\beta} + v_i + u_{it} \tag{۱۷.۲}$$

بود. در اینجا $\hat{\theta}_{it}$ برآوردگر آمارگیری مستقیم برای کوچک ناحیه‌ای i در زمان t است، پس، $\theta_{it} = g(\bar{Y}_{it})$ تابعی از میانگین کوچک ناحیه‌ای $\bar{Y}_{it}$ است، e_{it}ها عبارت‌اند از، خطاهای نمونه‌گیری با توزیع شرطی نرمال به شرط θ_{it}ها که دارای میانگین‌های صفر و یک ماتریس کوواریانس قطری بلوکی معلوم $\boldsymbol{\Psi}$ با بلوک‌های $\boldsymbol{\Psi}_i$اند، و $\mathbf{z}_{it}$ برداری از متغیرهای کمکی ناحیه ویژه است که برخی از آن‌ها، مثلا داده‌های ثبتی، ممکن است با زمان تغییر کنند. به‌علاوه، $v_i \overset{iid}{\sim} N(۰, \sigma_v^۲)$ و فرض می‌شود که u_{it} از یک فرآیند اتورگرسیو مرتبه‌ی اول مشترک برای هر i پیروی می‌کند. یعنی،

$$u_{it} = \rho u_{i,t-۱} + \varepsilon_{it}, \quad |\rho| < ۱ \tag{۱۸.۲}$$

که در آن $\varepsilon_{it} \overset{iid}{\sim} N(۰, \sigma^۲)$. فرض می‌شود که خطاهای $\{e_{it}\}$، $\{v_i\}$ و $\{\varepsilon_{it}\}$ نیز مستقل از یکدیگر هستند. مدل‌هایی که به شکل (۱۷.۲)و (۱۸.۲) با نادیده گرفتن خطای نمونه‌گیری، به‌طور گسترده مقالات اقتصادسنجی نیز به‌کار رفته‌اند (اندرسون و هسیااو، ۱۹۸۱).

مدل (۱۷.۲) برای $\boldsymbol{\theta}_{it}$ها هم به اثرهای ناحیه ویژه‌ی v_i و هم به اثرهای (ناحیه×زمان) ویژه‌ی u_{it} که به ازای هر i در طول زمان همبسته‌اند، و وابستگی دارند. همچنین می‌توانید (۱۷.۲) را به‌صورت مدلی با تاخیر توزیع‌شده بیان کنید:

$$\theta_{it} = \rho\theta_{i,t-۱} + (\mathbf{z}_{it} - \rho\mathbf{z}_{i,t-۱})^T\boldsymbol{\beta} + (۱ - \rho)v_i + \varepsilon_{it} \qquad (۱۹.۲)$$

صورت بدیل (۱۹.۲) متغیر θ_{it} را به میانگین دوره‌ی قبلی $\theta_{i,t-۱}$، مقدارهای متغیرهای کمکی در نقطه‌های زمانی t و $(t-۱)$، اثرهای تصادفی کوچک ناحیه‌ای v_i، و اثرهای (ناحیه × زمان) ε_{it} ارتباط می‌دهد. مدل‌های پیچیده‌تر از (۱۸.۲) برای u_{it} را می‌توان با فرض یک فرآیند میانگین متحرک اتورگرسیو (ARIMA) فرمول‌بندی کرد، اما مقدار افزایش حاصل در کارایی نسبت به (۱۸.۲) نامحتمل است که معنی‌دار باشد.

گوش و نانجیا (۱۹۹۳) و گوش، نانجیا و کیم (۱۹۹۶) نیز یک مدل سری زمانی مقطعی برای برآورد کوچک ناحیه‌ای پیشنهاد کردند. مدل آن‌ها به‌صورت زیر است:

$$\hat{\theta}_{it}|\theta_{it} \overset{ind}{\sim} N(\theta_{it}, \psi_{it}) \qquad (۲۰.۲)$$

$$\theta_{it}|\boldsymbol{\alpha}_t \overset{ind}{\sim} N(F_{it}\boldsymbol{\beta} + \mathbf{w}_{it}^T\boldsymbol{\alpha}_t, \sigma_t^۲) \qquad (۲۱.۲)$$

و

$$\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-۱} \overset{ind}{\sim} N_r(\mathbf{H}_t\boldsymbol{\alpha}_{t-۱}, \Delta) \qquad (۲۲.۲)$$

در اینجا $\mathbf{z}_{it}$ و $\mathbf{w}_{it}$ بردارهای متغیرهای کمکی ناحیه ویژه‌اند، ψ_{it}ها واریانس نمونه‌گیری‌اند که معلوم فرض می‌شوند، $\boldsymbol{\alpha}_t$ برداری $r \times ۱$ از اثرهای تصادفی زمان ویژه، و $\mathbf{H}_t$ یک ماتریس $r \times r$ی معلوم است. فضای وضعیت (۲۲.۲) در حالت تک متغیره $(r = ۱)$ با $H_t = ۱$ به مدل مشهور گام‌زدن تصادفی فرو می‌کاهد. مدل بالا مبتلا به دو محدودیت است: (یک) فرض می‌شود که برآوردهای مستقیم $\hat{\theta}_{it}$ برای هر i در طول زمان مستقل باشند. این فرض در ساختار آمارگیری مکرر با نمونه‌های هم‌پوش، از قبیل CPS و LFS، واقع گرایانه نیست. (دو) اثرهای تصادفی ناحیه ویژه در مدل گنجانده نشده‌اند که منجر به ترنجش زیاد برآوردگرهای کوچک ناحیه‌ای همانند برآوردگرهای هم‌گذاشتی می‌شود.

داتا و همکاران (۲۰۰۲) و یو (۱۹۹۹) مدل‌های نمونه‌گیری و پیونددهنده‌ی رائو–یو (۱۶.۲) و (۱۷.۲) را به‌کار بستند، اما به‌جای AR(1) مدل (۱۸.۲)، برای u_{it} یک مدل گام‌زدن تصادفی $u_{it} = u_{i,t-۱} + \varepsilon_{it}$ را گذاشتند که با رابطه‌ی (۱۸.۲) به ازای $\rho = ۱$ داده می‌شود. داتا و همکاران (۱۹۹۹) مدلی مشابه را در نظر گرفتند اما جمله‌هایی دیگر به مدل پیونددهنده افزودند تا تغییرات فصلی را در کاربرد آن‌ها بازتاب دهد.

پففرمن و بورک (۱۹۹۰) مدلی کلی شامل اثرهای تصادفی (ناحیه×زمان) ویژه را پیشنهاد کردند. مدل آن‌ها به‌صورت زیر است:

$$\hat{\theta}_{it} = \theta_{it} + e_{it} \tag{۲۳.۲}$$

$$\theta_{it} = \mathbf{z}_{it}^T \boldsymbol{\beta_{it}} \tag{۲۴.۲}$$

که در آن ضریب‌های $\boldsymbol{\beta}_{it} = (\beta_{it}, ..., \beta_{itp})^T$ به‌طور مقطعی و در طی زمان مجاز به تغییر هستند. فرض می‌شود که خطاهای e_{it} برای هر ناحیه‌ی i به‌طور پیاپی ناهمبسته با میانگین صفر و واریانس ψ_{it} باشند. تغییرات β_{it} در طی زمان با مدل فضای وضعیت زیر مشخص می‌شود:

$$\begin{bmatrix} \beta_{itj} \\ \beta_{it} \end{bmatrix} = \mathbf{T}_j \begin{bmatrix} \beta_{i,t-۱,j} \\ \beta_{ij} \end{bmatrix} + \begin{bmatrix} ۱ \\ ۰ \end{bmatrix} v_{itj}, \quad j = ۰, ۱, ..., p \tag{۲۵.۲}$$

در اینجا β_{ij}ها ضریب‌های ثابت‌اند، $\mathbf{T}_j$ یک ماتریس معلوم ۲×۲ با سطر دوم $(۰, ۱)$ است، و خطاهای مدل $\{v_{itj}\}$ به ازای هر i در طی زمان ناهمبسته‌اند و دارای میانگین صفر و واریانس‌های $E_m(v_{itj}, v_{itl}) = \sigma_{vjl}$، به ازای $j, l = ۰, ۱, ..., p$ هستند.

فرمول‌بندی (۲۵.۲) چند مدل سودمند را در بردارد. اول، گزینه‌ی $\mathbf{T}_j = \begin{bmatrix} ۰ & ۱ \\ ۰ & ۱ \end{bmatrix}$ مدل مشهور رگرسیونی با ضریب تصادفی $\beta_{itj} = \beta_{ij} + v_{itj}$ را می‌دهد. مدل آشنای گام‌زدن تصادفی $\beta_{itj} = \beta_{i,t-۱,j} + v_{itj}$ از انتخاب $\mathbf{T}_j = \begin{bmatrix} ۱ & ۰ \\ ۰ & ۱ \end{bmatrix}$ حاصل می‌شود. در این حالت ضریب β_{ij} اضافی است و باید حذف شود به‌طوری که $\mathbf{T}_j = ۱$. گزینه‌ی $\mathbf{T}_j = \begin{bmatrix} \rho & ۱-\rho \\ ۰ & ۱ \end{bmatrix}$ یک مدل AR(1) را می‌دهد: $\beta_{itj} - \beta_{ij} = \rho(\beta_{i,t-۱,j} - \beta_{ij} + v_{itj})$. مدل فضای وضعیت (۲۵.۲) کاملا کلی است، اما فرض خطاهای نمونه‌گیری به‌طور پیاپی ناهمبسته‌ی e_{it} در (۲۳.۲) در ساختار آمارگیری‌های مکرر با نمونه‌های هم‌پوش، محدود کننده است.

مثال ۳.۳.۲. (نرخ‌های بیکاری ایالت‌های متحده). داتا و همکاران (۱۹۹۹) مدل خود را برای برآورد نرخ‌های بیکاری ماهانه‌ی مربوط به ۴۹ ایالت آمریکا (با کنار گذاشتن ایالت نیویورک) و منطقه کلمبیا (شهر واشنگتن) $(m = ۵۰)$ که توسط انواع سازمان‌های فدرال به منظور تخصیص بودجه‌ها و تنظیم سیاست‌ها مورد استفاده قرار می‌گیرند، به‌کار بستند. آنان دوره‌ی زمانی ژانویه‌ی ۱۹۸۵ تا دسامبر ۱۹۸۸ را در نظر گرفتند و برآوردهای CPS را به‌عنوان $\hat{\theta}_{it}$ و نرخ ادعای بیمه‌ی بیکاری UI (درصد کارکنان بیکاری که از بین شاغلان غیر کشاورزی ادعای مزایای بیکاری می‌کنند) را به‌عنوان داده‌های کمکی، $\mathbf{z}_{it}$ به‌کار گرفتند. تغییرات فصلی در نرخ‌های بیکاری ماهانه از راه وارد کردن اثرهای تصادفی ماه و سال در مدل به حساب آورده شدند.

مدل فضایی

مدل پایه‌ای فی‌ـ‌هریوت (۵.۲) اثرهای کوچک ناحیه‌ای v_i را به صورت (i.i.d) فرض می‌کند، اما در برخی کاربردها شاید امتحان مدل‌هایی که همبستگی بین v_iها را مجاز می‌شمارند واقع گرایانه‌تر باشد. مدل‌های فضایی برای v_iها وقتی به‌کار می‌روند که بتوان ناحیه‌های «**همسایه**»ی هر ناحیه را تعریف کرد. چنین مدل‌هایی همبستگی‌های بین v_iها، مثلا، همبستگی‌هایی را که در ساختار برآورد نرخ‌های محلی بیماری و مرگ‌ومیر، به هم‌جواری جغرافیایی وابسته‌اند، القا می‌کنند. کرسی (۱۹۹۱) از یک مدل فضایی برای برآورد کوچک ناحیه‌ای در ساختار سرشماری استفاده کرد.

۴.۲ مدل پایه‌ای در سطح واحد آماری

فرض کنید که داده‌های کمکی واحد ویژه $\mathbf{x}_{it} = \{x_{ij1}, ..., x_{ijp}\}$ برای هر عنصر jی در جامعه در هر کوچک ناحیه‌ی i در دسترس‌اند. اغلب کافی است فرض کنید که تنها میانگین‌های جامعه‌ای $\bar{\mathbf{X}}_i$ معلوم‌اند. این اطلاعات از سرشماری و اطلاعات ثبتی به‌دست می‌آیند. به‌علاوه فرض می‌شود متغیر مورد نظر، y_{ij} از طریق یک مدل رگرسیون خطی با خطاهای آشیانی یک لایه‌ای e_{ij} به x_{ij} به‌صورت زیر مربوط می‌شود:

$$y_{ij} = \mathbf{x}_i^T \boldsymbol{\beta} + v_i + e_{ij}, \qquad j = ۱, ..., N_i, i = ۱, ..., m \qquad (۲۶.۲)$$

که در اینجا فرض بر این است که اثرهای ناحیه‌ی ویژه v_i متغیرهای تصادفی (i.i.d) هستند که در شرایط مدل زیر صدق می‌کنند،

$$E_m(v_i) = ۰, \qquad V_m(v_i) = \sigma_v^۲ (\geqslant ۰)$$

آن‌گاه $e_{ij} = k_{ij}\tilde{e}_{ij}$ با ثابت‌های معلوم k_{ij} و $\tilde{e}_{ij}$ها متغیرهای تصادفی (i.i.d) مستقل از v_iها هستند و

$$E_m(\tilde{e}_{ij}) = ۰, \qquad V_m(\tilde{e}_{ij}) = \sigma_e^۲. \qquad (۲۷.۲)$$

افزون بر آن، نرمال بودن v_iها و e_{ij}ها اغلب فرض می‌شود. پارامترهای مورد نظر میانگین‌های $\bar{Y}_i$ یا مجموع‌های Y_i کوچک ناحیه‌ای اند. مدل‌های رگرسیونی استاندارد با قرار دادن $\sigma_v^۲ = ۰$ یا به‌طور هم‌ارز $v_i = ۰$ در (۲۶.۲) به‌دست می‌آیند. این نوع مدل‌ها به برآوردگرهایی از نوع هم‌گذاشتی می‌انجامند.

فرض کنید که نمونه‌ی s_i با اندازه‌ی n_i از N_i واحد درون ناحیه iام $(i = ۱, ..., m)$ گرفته می‌شود و مقدارهای نمونه‌ای نیز از مدل (۲۶.۲) پیروی می‌کنند. فرض اخیر تحت نمونه‌گیری تصادفی ساده از هر ناحیه یا به‌طور کلی‌تر برای طرح‌های نمونه‌گیری‌یی که اطلاعات کمکی x_{ij}

را در گزینش نمونه‌های s_i به کار می‌گیرند، صادق است. برای درک بهتر این نکته، (۲۶.۲) را به‌صورت ماتریسی می‌نویسید.

$$\mathbf{y}_i^P = \mathbf{X}_i^P \boldsymbol{\beta} + v_i \mathbf{1}_i^P + \mathbf{e}_i^P, \qquad i = ۱, ..., m \tag{۲۸.۲}$$

که در آن $\mathbf{X}_i^P$ از مرتبه‌ی $N_i \times p$، بردارهای 1_i^P و $\mathbf{e}_i^P$ از مرتبه‌ی $N_i \times ۱$اند، و $1_i^P = (۱, .., ۱)^T$. سپس (۲۸.۲) را به دو بخش نمونه‌ای و غیرنمونه‌ای افراز می‌کنیم:

$$\mathbf{y}_i^P = \begin{bmatrix} \mathbf{y}_i \\ \mathbf{y}_i^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_i \\ \mathbf{X}_i^* \end{bmatrix} \boldsymbol{\beta} + v_i \begin{bmatrix} 1_i \\ 1_i^* \end{bmatrix} + \begin{bmatrix} \mathbf{e}_i \\ \mathbf{e}_i^* \end{bmatrix} \tag{۲۹.۲}$$

که در آن بالا نویس $*$ واحدهای نمونه‌گیری نشده را نشان می‌دهد. اگر مدل برای نمونه صادق باشد، یعنی، اگر اریبی گزینش وجود نداشته باشد، آن‌گاه استنباط‌ها درباره‌ی پارامترهای زیر

$$\boldsymbol{\psi} = (\beta^T, \sigma_v^۲, \sigma_v^۲)^T$$

بر پایه‌ی توزیع احتمال زیر استوارند:

$$f(\mathbf{y}_i | \mathbf{X}_i^P, \boldsymbol{\psi}) = \int f(\mathbf{y}_i, \mathbf{y}_i^* | \mathbf{X}_i^P, \boldsymbol{\psi}) d\mathbf{y}_i^*, \qquad i = ۱, ..., m \tag{۳۰.۲}$$

که در آن $f(\mathbf{y}_i, \mathbf{y}_i^* | \mathbf{X}_i^P, \boldsymbol{\psi})$ توزیع توام مفروض $\mathbf{y}_i^*$ و $\mathbf{y}_i$ است. از طرفی با در نظر گرفتن فرض $\mathbf{a}_i = (a_{i۱}, ..., a_{iNi})$ که در آن $a_{ij} = ۱$ اگر $\{j \epsilon s_i\}$ و در غیر این صورت $a_{ij} = ۰$، توزیع داده‌های نمونه‌ای $(\mathbf{y}_i, \mathbf{a}_i)$ با فرمول زیر داده می‌شود.

$$f(\mathbf{y}_i, \mathbf{a}_i | \mathbf{X}_i^P, \boldsymbol{\psi}) =$$

$$\int f(\mathbf{y}_i, \mathbf{y}_i^* | \mathbf{X}_i^P, \boldsymbol{\psi}) f(\mathbf{a}_i | \mathbf{y}_i, \mathbf{y}_i^*, \mathbf{X}_i^P) d\mathbf{y}_i^* =$$

$$[\int f(\mathbf{y}_i, \mathbf{y}_i^* | \mathbf{X}_i^P, \boldsymbol{\psi}) d\mathbf{y}_i^*] f(\mathbf{a}_i | \mathbf{X}_i^P)$$

به شرط آنکه رابطه‌ی زیر برقرار باشد:

$$f(\mathbf{a}_i | \mathbf{y}_i, \mathbf{y}_i^*, \mathbf{X}_i^P) = f(\mathbf{a}_i | \mathbf{X}_i^P)$$

یعنی احتمال‌های گزینش نمونه به $\mathbf{y}_i^p$ وابسته نیستند اما ممکن است به $\mathbf{X}_i^p$ وابسته باشند. در این حالت، اریبی گزینش وجود ندارد و می‌توانید فرض کنید که مقدارهای نمونه‌ای نیز از مدل مفروض پیروی می‌کنند، یعنی، از $f(\mathbf{y}_i | \mathbf{X}_i^p, \boldsymbol{\psi})$ برای استنباط‌هایی درباره $\boldsymbol{\psi}$ می‌توان استفاده کرد (اسمیت، ۱۹۸۳).

اگر احتمال‌های گزینش نمونه‌ای به متغیری کمکی، مثلا $\mathbf{z}_i^p$، وابسته باشد که در $\mathbf{X}_i^P$ گنجانده نشده است، آن‌گاه توزیع داده‌های نمونه‌ای $(\mathbf{y}_i, \mathbf{a}_i)$ عبارت‌اند از

$$f(\mathbf{y}_i, \mathbf{a}_i | \mathbf{X}_i^P, \mathbf{z}_i^P, \boldsymbol{\psi}) = [\int f(\mathbf{y}_i, \mathbf{y}_i^* | \mathbf{X}_i^P, \mathbf{z}_i^P, \boldsymbol{\psi}) d\mathbf{y}_i^*] f(\mathbf{a}_i | \mathbf{z}_i^P, \mathbf{X}_i^P)$$

در این حالت، استنباط درباره‌ی ψ مبتنی بر $f(\mathbf{y}_i|\mathbf{X}_i^p, \mathbf{z}_i^p, \psi)$ است که با (۳۰.۲) تفاوت دارد، مگر آنکه، $\mathbf{z}_i^p$ به شرط $\mathbf{X}_i^p$ با $\mathbf{y}_i^p$ ارتباط نداشته باشد. در این حالت، اریبی گزینش خواهید داشت و بنابراین نمی‌توانید فرض کنید که مدل (۲۸.۲) برای مقدارهای نمونه‌ای برقرار است. شاید بتوانید مدل (۲۸.۲) را با گنجاندن $\mathbf{z}_i^p$ و سپس آزمون فرض معنی‌دار بودن ضریب رگرسیونی مربوط با استفاده از داده‌های نمونه‌ای، گسترش دهید. اگر فرض صفر رد نشود، آن‌گاه می‌توانید فرض کنید که مدل اصلی (۲۸.۲) برای مقدارهای نمونه‌ای نیز برقرار است (اسکینر، ۱۹۹۴).

مدل (۲۸.۲) تحت نمونه‌گیری خوشه‌ای دومرحله‌ای درون ناحیه‌های کوچک نیز مناسب نیست، زیرا اثرهای تصادفی خوشه‌ها در آن وارد نشده‌اند. اما می‌توانید مدل را برای در نظر گرفتن چنین ویژگی‌هایی گسترش دهید.

هدف یافتن برآورد پارامترهای میانگین‌های کوچک ناحیه‌ای $\theta_i = \bar{Y}_i = \mathbf{x}_i^T\boldsymbol{\beta} + v_i$ یا مجموع Y_i کوچک ناحیه‌ای است. در مدل (۲۸.۲) اگر قرار دهیم $\sigma_v^2 = 0$ یا به‌طور هم‌ارز $v_i = 0$، آن‌گاه مدل‌های رگرسیونی استاندارد به‌دست می‌آید که نوعی از برآوردگرهای هم‌گذاشتی است. میانگین کوچک ناحیه‌ای $\bar{Y}_i$ را به‌صورت زیر می‌نویسند:

$$\bar{Y}_i = f_i\bar{y}_i + (1 - f_i)\bar{Y}_i^* \qquad (۳۱.۲)$$

که $f_i = n_i/N_i$ و $\bar{y}_i$ و $\bar{Y}_i^*$ به ترتیب میانگین‌های واحدهای نمونه گرفته شده و نمونه گرفته نشده را نشان می‌دهند. از مدل (۳۱.۲) داریم برآورد میانگین کوچک ناحیه‌ای $\bar{Y}_i$ معادل با برآورد مقدار تحقق‌یافته متغیر تصادفی $\bar{Y}_i^*$ به‌شرط داده‌های نمونه‌ای $\{\mathbf{y}_i\}$ و داده‌های کمکی $\{\mathbf{X}_i^p\}$ است. اگر اندازه‌ی جامعه N_i بزرگ باشد، با توجه به اینکه $\bar{Y}_i = \bar{\mathbf{x}}_i^T\boldsymbol{\beta} + v_i + \bar{E}_i$ و $\bar{E}_i \approx 0$ که در آن $\bar{E}_i$ میانگین N_i خطای e_{ij} و $\bar{\mathbf{X}}_\mathbf{i}$ میانگین معلوم $\mathbf{X}_i^p$ است، می‌توانید میانگین‌های کوچک ناحیه‌ای را به‌صورت

$$\bar{Y}_i = \bar{\mathbf{X}}_i^T\boldsymbol{\beta} + v_i \qquad (۳۲.۲)$$

بنویسید. از مدل (۳۲.۲) چنین بر می‌آید که برآورد کردن $\bar{Y}_i$ هم‌ارز با برآورد کردن ترکیب خطی از $\boldsymbol{\beta}$ و مقدار تحقق‌یافته متغیر تصادفی v_i است.

مثال ۱.۴.۲. (دستمزدها و حقوق‌ها). رائو و چاودری (۱۹۹۵) جامعه مالیات پردازان غیرشرکتی از استان نووا اسکوشیای کانادا را بررسی کردند. آن‌ها مدل

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + v_i + x_{ij}^{1/2}\tilde{e}_{ij}$$

را که حالت خاصی از (۲۶.۲) با $k_{ij} = x_{ij}^{1/2}$ است، پیشنهاد کردند. در اینجا y_{ij} و x_{ij} کل دستمزد و حقوق‌ها و درآمد ناخالص را برای بنگاه jام در ناحیه‌ی iام نشان می‌دهند. برای برآورد مجموع‌های کوچک ناحیه‌ای Y_i یا میانگین‌های $\bar{Y}_i$ از نمونه‌گیری تصادفی ساده از جامعه‌ی سراسری استفاده کردند.

۱.۴.۲ گسترش‌های مدل در سطح واحد آماری (B)

گسترش مدل در سطح واحد آماری انواع گوناگونی دارند که عبارت‌اند از: مدل رگرسیونی با خطای آشیانی چندمتغیره، مدل خطی با واریانس خطای تصادفی، مدل رگرسیونی با خطای آشیانی دولایه، مدل دوسطحی و مدل آمیخته‌ی خطی کلی.

تذکر ۱.۴.۲. سازگاری ضرایب برآوردگرهای متغیرهای کمکی بین ناحیه و مدل‌های سطح واحد همیشه کامل نخواهد بود (گرینلند و همکاران، ۱۹۹۴). انتظار می‌رود که داشتن مقادیر مشابه در هر دو مورد مناسب باشد (برازش داده شده) اما هنگامی که از داده جمع‌آوری شده استفاده شود، ممکن است تعداد اریبی در برآوردگرها مشاهده شود که به اریبی بوم شناختی معرفی شده‌است. این حالت ممکن است اتفاق بیافتد، برای مثال هنگامی که اختلالی در متغیرهای کمکی سطح فردی و متغیرهای کمکی جمع‌آوری شده وجود داشته باشد نمی‌توان آن را شرح داد. از این رو، دقت کنید که به‌طور معمول ضرایب را همانند اندازه‌گیری اثرات سطح فردی تفسیر کنید. انتخاب مدل نیز وابسته به نوع داده خواهد بود. داده جمع‌آوری شده آسان‌تر به‌دست می‌آید، همانند موسسات و دفاتر مختلف آماری که داده‌ی آماری سطح ناحیه را با حجم‌های زیادی ارائه می‌دهند.

۵.۲ مدل رائو_یو

برای نشان دادن کارایی این مدل مثالی را مطرح کرده‌ایم. در نظرسنجی نیروی کار در کانادا، خانوارهای مورد نظر را طی ۶ ماه متوالی و در هر ماه مصاحبه می‌کنند. همچنین ارزیابی کودکان و نوجوانان سنین ۵ تا ۱۷ سال برای بررسی این‌که چه تعداد از این افراد در زیر خط فقر جای دارند یا خیر. از طرفی زمان سرشماری برای بررسی این موضوع هر دو سال و یا هر سال انجام می‌شود. زیرا دولت برای تخصیص دهی بودجه نیازمند داشتن چنین اطلاعاتی است که کدام مناطق نیازمند بودجه‌ی بیشتری است. اما همیشه سرشماری کردن در بازه‌های زمانی طولانی کارا نیست لذا در هر نظرسنجی ترکیب کردن داده‌ها در طول زمان متوالی منجر به افزایش اندازه نمونه و ... می‌شود، پس این روش دقت در برآوردهای نواحی کوچک را افزایش می‌دهد. لذا رائو و یو[۹] (۱۹۹۲، ۱۹۹۴) مدل فی_هریوت را به دو قسمت تقسیم کردند :

۱. مدل نمونه‌گیری
۲.مدل ربط (پیوند)

با این عمل توانستند خیلی از نواقص مدل فی_هریوت را برطرف کنند. (مانند نادیده گرفتن ویژگی مدل) بنابراین با توجه به مطالب گفته شده مدل رائو_یو را شرح خواهیم داد.

[۹] Rao-yo model

- $\mathbf{X}_i$ برداری از کاراکترهای ناحیه
- θ_i کاراکترهای ناحیه غیرشناخته شده از عواملی همچون میانگین، واریانس، مد و ...
- $\tilde{\theta}_i$ یک برآوردگر مستقیم از θ_i (میانگین نمونه، برآوردگرهای مبتنی بر طرح و ...)
- m تعدادی از نواحی با داده‌های نمونه
- n_i اندازه نمونه در ناحیه‌ی iام و N_i اندازه ناحیه

تذکر ۱.۵.۲. رابطه‌ی زیر صورت کلی مدل را در حالت نمونه‌گیری (یعنی در این مدل از خود مدل استفاده نمی‌شود، به‌طور حتم اگر یک برآوردگر مستقیم داشته باشیم این برآوردگر با واقعیت خود می‌تواند فرقی داشته باشد که این فرق ناشی از خطای نمونه‌گیری می‌باشد، یعنی e_i) نشان می‌دهد.

$$\tilde{\theta}_i = \theta_i + e_i$$

تذکر ۲.۵.۲. همچنین رابطه‌ی زیر هم صورت کلی مدل را در حالت ربط (در واقع θ شما خود از یک مدل تولید می‌شود و در این مدل از ویژگی‌های حالت نمونه‌گیری برای پیدا کردن برآوردگر استفاده نخواهیم کرد و تنها از ویژگی‌های خود مدل بهره‌مند می‌شویم تا برآوردگر را بهتر شناسایی نماییم) نشان می‌دهد.

$$\theta_i = \mathbf{X}_i^T \boldsymbol{\beta} + u_i$$

$\mathbf{X}_i$ بردارمقادیر متغیر کمکی و با ابعاد $(p \times ۱)$ و با استفاده از اطلاعاتی که در سطح ناحیه قرار دارد به‌دست آورده می‌شود و $\boldsymbol{\beta}$ نیز برداری با ابعاد $(p \times ۱)$ و نشان‌دهنده‌ی ضرایب رگرسیونی است. لذا e_i خطای نمونه‌گیری و مستقل همچنین u_i نشان‌دهنده‌ی اثرات تصادفی و دارای توزیع زیر می‌باشند:

$$e_i \overset{ind}{\sim} N(۰, \sigma^۲_{Di}) \text{ و } u_i \overset{iid}{\sim} N(۰, \sigma^۲_u)$$

۶.۲ مدل‌های آمیخته‌ی خطی کلی

۱.۶.۲ مقدمه

در برآورد کوچک ناحیه‌ای استفاده کردن از مدل‌های آمیخته خطی مناسب است، دلیل استفاده کردن از این نوع مدل‌ها در ترکیب کردن اطلاعات از منابع مختلف و در نظر گرفتن منابع مختلف تولید کننده‌ی خطا، که به‌طور ویژه موثر می‌باشند. مدل‌های آمیخته‌ی خطی نوعی از مدل‌های آماری هستند که از اثرهای ثابت و تصادفی تشکیل شده‌اند. بنابراین این نوع از مدل‌ها تفاوتی که بین نواحی وجود دارد را نمی‌تواند به کمک اثرهای ثابت به‌دست آورد، پس به کمک اثرهای تصادفی محاسبه می‌شوند، از طرفی مدل‌های خطی رگرسیونی وجود دارد که

تفاوت بین نواحی را به کمک متغیرهای کمکی نشان می‌دهد. به‌طور حتم در مفهوم کوچک ناحیه‌ای، این دسته از مدل‌ها نوعا به دو گروه کلی تقسیم می‌شوند یکی در سطح ناحیه و دیگری در سطح واحد. به منظور آشنایی بیشتر این موضوع مطالب مختصری از مدل‌های آمیخته‌ی خطی و تعمیم‌یافته و تعمیم‌یافته فضایی را شرح می‌دهیم.

۲.۶.۲ مدل آمیخته‌ی خطی LMM

همان‌طوری که در مقدمه اشاره کردیم، مدل آمیخته‌ی خطی یک مدل آماری است که شامل اثرهای ثابت و تصادفی می‌باشد. کاربرد این مدل را می‌توان در شاخه‌های علوم زیستی و اجتماعی دید. از طرفی مدل آمیخته‌ی خطی در برآورد کوچک ناحیه‌ای پرکاربرد است، زیرا استفاده از این مدل می‌تواند تفاوت بین نواحی را که توسط اثرهای ثابت مدل محاسبه نمی‌شوند، با به‌کارگیری از اثرهای تصادفی انجام داد و همین طور مدل‌های رگرسیونی خطی وجود دارد که تفاوت بین نواحی را فقط از طریق اطلاعات کمکی نشان می‌دهد. مدل‌های آمیخته‌ی خطی را می‌توان بر سه نوع برشمرد. مدل‌هایی با اثرهای ثابت، مدل‌هایی با اثرهای تصادفی و مدل‌هایی با اثرهای آمیخته.

بهترین راه حلی که به ما کمک می‌کند تا بتوان مدل آمیخته‌ی خطی را راحت‌تر درک کرد، ارائه یک مدل رگرسیونی است، این مدل برای اینکه ارتباط بین متغیر پاسخ و متغیرهای کمکی را ایجاد کند، استفاده می‌شود و آن را به‌صورت زیر نشان می‌دهیم:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \qquad (۳۳.۲)$$

باتوجه به معادله‌ی بالا $\mathbf{y}$ بردار مشاهدات، $\mathbf{X}$ ماتریس طرح معلوم، $\boldsymbol{\beta}$ بردار نامعلوم از ضرایب رگرسیونی است، که اغلب اثرهای ثابت نامیده می‌شوند و $\mathbf{e}$ یک بردار خطاهای تصادفی غیرقابل مشاهده است. به طور کل فرض نرمال بودن خطاها را در نظر بگیرید. خطاها دارای توزیع نرمال با میانگین صفر و واریانس σ^2 می‌باشند، یعنی $\mathbf{e} \overset{iid}{\sim} N(0, \sigma^2\mathbf{I})$. بنابراین توزیع مقدارهای مشاهده شده نیز نرمال خواهد شد، و در نتیجه داریم $\mathbf{y} \overset{iid}{\sim} N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ و صورت ماتریسی این مدل به شکل زیر است:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \qquad (۳۴.۲)$$

که $\mathbf{y}_{n\times 1}$ بردار مشاهدات، $\boldsymbol{\beta}_{p\times 1}$ بردار اثرهای ثابت، $\mathbf{v}_{q\times 1}$ بردار اثرهای تصادفی، $\mathbf{e}_{n\times 1}$ بردار خطاهای تصادفی، $\mathbf{X}_{n\times p}$ ماتریس طرح برای اثرهای ثابت و $\mathbf{Z}_{q\times 1}$ ماتریس طرح برای اثرهای تصادفی که هردوی آن‌ها معلوم هستند. از طرفی توجه به اینکه $\mathbf{e}$ و $\mathbf{v}$ متغیرهای تصادفی غیرهمبسته با میانگین‌های صفر و دارای ماتریس کوواریانس می‌باشد و می‌توان به‌صورت زیر آن را نشان داد:

$$Var(\mathbf{v}) = \mathbf{G}, Var(\mathbf{e}) = \mathbf{R}, Cov(\mathbf{v}, \mathbf{e}) = 0$$

لذا ماتریس‌های کوواریانس شامل بعضی از پارامترهای پراکندگی نامعلوم می‌باشند. اگر فرض نرمال بودن اثرهای تصادفی و خطاها آورده شود، بنابراین خواهیم داشت $\mathbf{v} \sim N(0, \mathbf{G})$ و $\mathbf{e} \sim N(0, \mathbf{R})$، که این مدل را مدل آمیخته‌ی خطی گاوسی [۱۰] می‌نامند. همچنین فرض می‌شود که توزیع حاشیه‌ای $\mathbf{y}$ نرمال چند متغیره‌ای است که مشمول میانگین $\mathbf{X}\boldsymbol{\beta}$ و ماتریس کوواریانس $\mathbf{V} = \mathbf{ZGZ}^T + \mathbf{R}$ که تابعی از بردار ν می‌باشد پس $\mathbf{V} = \mathbf{V}(\nu)$ و در نتیجه $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ را خواهیم داشت (جیانگ و لاهیری، ۲۰۰۶). همان‌طوری که در مدل آمیخته‌ی خطی (۲.۳۴) مشخص است، این مدل‌ها سه مولفه دارند (رائو،۲۰۰۳):

- مولفه تصادفی: این مولفه متغیر پاسخ $\mathbf{y}$ را با یک توزیع احتمال مناسب تعریف می‌کند. توزیع احتمال متغیر پاسخ می‌تواند توزیع‌های مختلفی باشد که معمولا برای داده‌های گسسته توزیع دوجمله‌ای و یا پواسون و برای پیوسته توزیع نرمال مورد استفاده قرار می‌گیرند.
- مولفه سیستماتیک: این مولفه متغیرهای کمکی استفاده شده یعنی، متغیرهایی که نقش x_j و z_j را در مدل دارند، به عنوان پیشگوهای مدل تعیین شده و به‌صورت ترکیب خطی از متغیرهای پیشگو یعنی $\boldsymbol{\eta} = \mathbf{x}\boldsymbol{\beta} + \mathbf{zv}$ تحت عنوان پیشگوی خطی در نظر می‌گیرند.
- مولفه ربط: تابع یا پیوندی بین مولفه سیستماتیک و امیدریاضی (میانگین) مولفه تصادفی است که در آن نحوه ارتباط $\boldsymbol{\mu} = E(\mathbf{y}|\mathbf{x}, \mathbf{z})$ با متغیرهای کمکی در پیشگوی خطی نشان داده می‌شود و می‌توان میانگین $\boldsymbol{\mu}$ را به‌طور مستقیم و یا به‌صورت تابع $\boldsymbol{\eta} = g(\boldsymbol{\mu})$ و به‌صورت $\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \mathbf{x}\boldsymbol{\beta} + \mathbf{zv}$ بیان کرد، بنابراین، $g(.)$ را تابع ربط می‌نامند. در تابع ربط، میانگین متغیر پاسخ با پیشگوی خطی ربط داده می‌شود و این مساله از تبدیل بر روی متغیر پاسخ متفاوت است. همچنین نکته مهم در انتخاب تابع ربط، این است که تابع انتخاب شده مقادیر پیشگویی قابل قبول از متغیر پاسخ را نتیجه می‌دهد (جیانگ و لاهیری، ۲۰۰۶؛ رائو، ۲۰۰۳).

باتوجه به مطالب گفته شده، برای وضوع بهتر مباحثی را در پیرامون مدل‌های آمیخته‌ی خطی تعمیم‌یافته GLMM شرح می‌دهیم.

۳.۶.۲ مدل آمیخته‌ی خطی تعمیم‌یافته GLMM

بنابر مطالبی که در بخش قبلی گفته شد، مدل‌های رگرسیونی را در مواقعی می‌توان استفاده کرد که متغیر پاسخ، متغیری پیوسته باشد اما با این حال شروطی نیز لازم است:

۱. مشاهدات از هم مستقل درنظر گرفته شود.

۲. می‌توان روابط بین متغیر پاسخ و اثرهای ثابت را به کمک یک تابع خطی مدل‌سازی کرد.

۳. از طرفی واریانس متغیر پاسخ یک مقدار ثابت باشد.

[۱۰] Gaussian linear mixed model

۴. خطاها معمولا از توزیع نرمال پیروی می‌کنند.

بنابر شروط گفته شده، در بعضی کاربردها می‌توان آن‌ها را نقض کرد، بنابراین استفاده از مدل‌های LMM و GLMM راه حلی برای برطرف کردن این نوع نقص‌ها می‌باشد. در مدل آمیخته‌ی خطی فرض بر این است که، متغیرهای پاسخ از هم مستقل‌اند و از یک توزیع خاص که جزء خانواده‌های نمایی است پیروی می‌کند که در بیشتر مواقع این توزیع‌ها شامل توزیع‌های نرمال، پواسون و برنولی می‌باشد. لذا میانگین متغیر پاسخ $\boldsymbol{\mu}$ به ترکیب خطی از متغیرهای کمکی $\mathbf{X}\boldsymbol{\beta} = \xi_1$ به کمک معکوس تابع پیوند مرتبط می‌شوند، پس خواهیم داشت:

$$E(\mathbf{y}) = \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta}) = g^{-1}(\xi_1)$$

β برداری از ضرایب رگرسیونی نامعلوم و همچنین g یک تابع پیوند است. اما در حالت کلی مدل رگرسیونی یک حالت خاص از مدل GLM می‌باشد که تابع پیوند آن‌ها تابع همانی است. لذا اگر مشاهدات نسبت به‌هم مستقل نباشند، مدل مطرح شده همان GLM است در اغلب موارد مناسب نیستند به همین دلیل، لازم است از مدل دیگری استفاده کرد و آن هم مدل‌های آمیخته‌ی خطی تعمیم‌یافته می‌باشد.

در مدل‌های آمیخته‌ی خطی تعمیم‌یافته پیشگوی خطی همانند مدل‌های آمیخته‌ی خطی شامل اثرهای ثابت و تصادفی می‌باشد، یعنی:

$$\xi_2 = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

همچنین در این دسته از مدل‌ها،

$$E(\mathbf{y}|\mathbf{u}) = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) = g^{-1}(\xi_2)$$

$\mathbf{X}_{n\times p}$ ماتریس طرح برای اثرهای ثابت $\boldsymbol{\beta}_{p\times 1}$، $\mathbf{Z}_{n\times q}$ یک ماتریس طرح برای اثرهای تصادفی $\mathbf{u}_{q\times 1}$ و g یک تابع پیوند است. حال فرض کنید که مدل آمیخته‌ی خطی را بسط دهیم آن‌گاه مشاهدات $y_1, y_2, ..., y_n$ به‌صورت شرطی مستقل هستند، یعنی $y_i|\mathbf{u} \sim N(\mathbf{x_i'}\boldsymbol{\beta} + \mathbf{z_i'}\mathbf{u}, \tau^2)$. با توجه به این می‌توان گفت که x_i' و z_i' سطرهای iام ماتریس‌های $\mathbf{X}$ و $\mathbf{Z}$ می‌باشند و پارامتر τ^2 یک واریانس نامعلوم است و با داشتن فرض نرمال و $\mathbf{R} = \tau^2\mathbf{I}$ منجر به مدل $\boldsymbol{\theta}_i = \mathbf{X}_i\boldsymbol{\beta} + v_i$ می‌شود. اکنون اگر فرض بر این باشد که، بردارهای تصادفی $\mathbf{u}$ برقرار باشند و $y_1, y_2, ..., y_n$ به‌طور شرطی مستقل باشند، آن‌گاه توزیع شرطی $y_i|\mathbf{u}$ یکی از اعضای خانواده توزیع‌های نمایی با تابع چگالی احتمال زیر می‌شود:

$$f_i(y_i|\mathbf{u}) = exp[\frac{y_i\xi_i - b(\xi_i)}{a_i(\phi)} + c_i(y_i, \phi)],$$

که ϕ پارامتر پراکندگی و $a_i(\circ), b_i(\circ), c_i(\circ,\circ)$ توابعی معلوم هستند و اگرچه در برخی موارد اینکه نامعلوم هستند نیز امکان دارد. جیانگ و لاهیری (۲۰۰۶) ژانگ و لی (۲۰۱۱)، بعد از انتخاب مدل آماری مناسب با توجه به شرایط مساله روش‌های مرسومی که برای برآورد کوچک ناحیه‌ای شدنی می‌باشد را تبیین کردند.

۴.۶.۲ مدل‌های آمیخته خطی تعمیم‌یافته فضایی SGLMM

فرض کنید G ناحیه فضایی مشبکه‌ای و $Y(s_i)$، $i = ۱, \ldots, n$ متغیر پاسخ شمارشی برای ناحیه s_i باشد. همچنین فرض کنید در هر ناحیه مقادیر بردار p بعدی متغیرهای تبیینی $\mathbf{x}(s_i)$ مشاهده شده باشند. یک SGLM به صورت زیر تعریف می‌شود:

$$\mathrm{E}\left(Y(s_i)|v(s_i)\right) = g^{-۱}(\mathbf{x}^T(s_i)\boldsymbol{\beta} + v(s_i) + \epsilon(s_i)), \quad i = ۱, \ldots, n \tag{۳۵.۲}$$

که در آن $g(\cdot)$ یک تابع پیوند مشتق‌پذیر و معکوس‌پذیر است، $\boldsymbol{\beta}$ بردار p بعدی پارامترهای ثابت مدل است و $\epsilon(s_i)$ جمله خطای تصادفی غیرساختارمند با توزیع نرمال $\mathrm{N}(۰, \sigma^۲_\epsilon)$ است که اغلب برای خطای اندازه‌گیری در نظر گرفته می‌شود. اثر تصادفی فضایی $\mathbf{v} = (v(s_۱), \ldots, v(s_n))'$ از یک مدل اتورگرسیو شرطی[۱۱] (CAR) همانند مدل زیر پیروی می‌کند.

$$v_i|\mathbf{v}_{-i}, \sigma^۲_v \sim \mathrm{N}(\sum_{j\epsilon\delta_i} \frac{d_{ij}}{|\delta_i|} v_j, \frac{\sigma^۲_v}{|\delta_i|})$$

و توزیع متغیر پاسخ $Y(s_i)$ به شرط $v(s_i)$، از خانواده توزیع‌های نمایی با تابع چگالی

$$f(y(s_i)|v(s_i); \boldsymbol{\beta}, \phi) = \exp[\frac{۱}{\phi}\left(a(\mu_i)y(s_i) - b(\mu_i)\right) + c(y(s_i), \phi)]$$

پیروی می‌کند، که در آن $\mu_i = \mu(s_i) = \mathrm{E}(Y(s_i)|v(s_i))$، ϕ پارامتر پراکندگی معلوم و توابع $a(\cdot)$، $b(\cdot)$ و $c(\cdot, \cdot)$ نیز معلوم هستند.

تابع درستنمایی کناری مدل(۳۵.۲)، برای بردار مشاهدات $\boldsymbol{y}$، به صورت

$$\mathrm{L}(\boldsymbol{\beta}, \sigma^۲_v, \sigma^۲_\epsilon|\mathbf{y}) \propto \int \prod_{i=۱}^{n} f(y(s_i)|v(s_i); \boldsymbol{\beta})\phi_n(\mathbf{v}; ۰, \sigma^۲_v A^{-۱})\mathrm{d}\mathbf{v} \tag{۳۶.۲}$$

است، که در آن $\phi_n(\mathbf{v}; ۰, \sigma^۲_v A^{-۱})$ تابع چگالی نرمال n متغیره با بردار میانگین صفر و ماتریس کوواریانس $\sigma^۲_v A^{-۱}$ است. محاسبه تابع درستنمایی (۳۶.۲) مستلزم حل انتگرالی با بعد برابر تعداد مشاهدات است و برای داده‌های (تعداد نواحی) حجیم، محاسبه آن کاری طاقت‌فرسا و مشکل‌ساز است. هر چند روش‌های تقریبی مختلفی در دیدگاه مبتنی بر درستنمایی برای محاسبه این درستنمایی پیشنهاد شده‌اند (بریسلو و کلیتون، ۱۹۹۳). اما استفاده از چارچوب بیزی برای رده SGLMMs ، به دلیل وجود الگوریتم‌های MCMC، انتخاب اول است.

کاربرد مدل GLM در برآورد کوچک ناحیه‌ای

۱. مدل‌های اثرات آمیخته در سطح ناحیه

مدل‌های اثرات آمیخته نوع **A** (یعنی همان مدل در سطح ناحیه) را می توانید با استفاده از مدلی که به فرم زیر، نمایش دهید:

$$\hat{\bar{\mathbf{Y}}} = \bar{\mathbf{X}}\boldsymbol{\beta} + \mathbf{Z}u + \varepsilon$$

[۱۱] Conditional autoregressive

که در آن $\hat{\tilde{\mathbf{Y}}}_D$ برداری از برآوردگرهای غیرمستقیم، $\mathbf{Z}$ نیز به توصیفی از ساختار اثرات تصادفی u می‌پردازد و ε جمله خطا می‌باشد. در اصل می‌خواهیم فرض کنیم که $\mathbf{Z}$ یک ماتریس همانی است، اما یک ساختار پیچیده‌تر، زمانی می‌تواند مورد استفاده قرار بگیرد که مقدار آن در یک ناحیه به چندین اثر تصادفی بستگی داشته باشد. فرض می‌کنیم که اثرات تصادفی u_i دارای توزیع نرمال با میانگین صفر و واریانس $\sigma_u^۲$ است. خطای نمونه‌گیری یا تغییرات فردی ε، مستقل با میانگین صفر و واریانس قطری $(\hat{\sigma}_i^۲)$ در نظر گرفته می‌شود. علاوه بر این $\mathbf{D}$ و $\mathbf{V}$ ماتریس واریانس u و ε اند. ناشرطی بودن روی اثرات تصادفی، برای واریانس $\hat{\tilde{\mathbf{Y}}}_D$ می‌تواند به این صورت $\mathbf{G} = \acute{\mathbf{Z}}\mathbf{D}\mathbf{Z}+\mathbf{V}$ باشد. با توجه اینکه واریانس شناخته شده است، پارامتر $\boldsymbol{\beta}$ می‌تواند توسط میانگین‌هایی از کمترین توان‌های دوم تعمیم یافته شده برآورد شود، هر چند که در این مورد با ماکسیم درستنمایی معادل است. اثرات تصادفی را می‌توان به چندین روش به شرح زیر برآورد کرد. از این رو، برآوردگر

$$\hat{\tilde{\mathbf{Y}}} = \bar{\mathbf{X}}_{\mathbf{i}}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{u}$$

به طوری که برآوردگرهای غیرمستقیم برای برخی نواحی از دست رفته است اما با این حال متغیرهای کمکی در سطح ناحیه وجود دارد، و مقدار $\hat{u}$ باید به صفر برسد و آن‌گاه برآوردگر ترکیبی محاسبه می‌شود.

$$\hat{\tilde{\mathbf{Y}}} = \bar{\mathbf{X}}_{\mathbf{i}}\hat{\boldsymbol{\beta}}$$

اگر اثرات تصادفی یک ساختار داده شود برآوردگر قبلی می‌تواند برای برآورد اثرات تصادفی حساب شود.

۲. مدل‌های اثرات آمیخته در سطح واحد

برای مدل‌های نوع $\mathbf{B}$ (یعنی همان مدل در سطح واحد) را می‌توانید مشابه فرمولی زیر داشته باشید:

$$\mathbf{y}_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \sum_l \mathbf{z}_{li}\mathbf{u}_i + e_{ij}$$

که در این صورت e_{ij} تغییرات تصادفی فردی یا واحد jام در ناحیه‌ی iام است. این مدل‌ها همچنین می‌توانند با تقسیم داده‌ها به واحدهای نمونه‌ی $(y_۱)$ و غیرنمونه‌ی $(y_۲)$ بیان شوند:

$$\begin{bmatrix} y_۱ \\ y_۲ \end{bmatrix} = \begin{bmatrix} x_۱ \\ x_۲ \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} Z_۱ \\ Z_۲ \end{bmatrix} u + \begin{bmatrix} e_۱ \\ e_۲ \end{bmatrix}$$

توجه داشته باشید که طول از $y_۱$ و $y_۲$ را به صورت $\sum_{i=۱}^n$ و $\sum_{i=۱}^n (N_i - n_i)$ است و تمامی ماتریس‌های دیگر نیز بر اساس اندازه‌ها می‌باشند. برآورد از مدلی که مبنی بر معادله داده مشاهده شده

$$y_۱ = x_۱\beta + Z_۱u + e_۱$$

و روش‌های متفاوتی که می توان برای ارایه برآوردهای پارامتر از مدل استفاده کرد امکان پذیر خواهد بود. برآورد از پارامتر β توسط ماکسیم درستنمایی[۱۲] (ML)، یا شبه احتمال تاوانیده[۱۳] نیز انجام می‌شود، ضمن اینکه ماکسیم درستنمایی مقید[۱۴] (REML) می‌تواند برای به‌دست آوردن غیراریبی برآورد واریانس استفاده شود. مک‌کولا و سیرل (۲۰۰۱)، استراتژی‌های دیگری را برای به‌دست آوردن برآوردگرهایی از پارامترهای مختلف مطرح کرده‌اند. علاوه بر این ایراریبا کانسرسیوم (۲۰۰۴) الگوریتم‌های دقیق و ترفندهای محاسباتی را برای برآورد پارامترهایی در نوع‌های متفاوت آن هم از مدل‌های آمیخته ارایه داده است. سپس برآورد کوچک ناحیه‌ای با جمع‌آوری مشاهدات و مقادیر برآورد شده در یک روش به‌دست می‌آید:

$$\hat{\bar{Y}}_i = \frac{n_i}{N_i}\sum_{j=1}^{n_i} y_{ij} + \frac{N_i - n_i}{N_i}\sum_{j=1}^{N_i-n_i} \hat{y}_{ij} = \frac{n_i}{N_i}\sum_{j=1}^{n_i} y_{ij} + \frac{N_i - n_i}{N_i}\sum_{j=1}^{N_i-n_i} x_{ij}\hat{\beta} + \hat{u}_i \qquad (۳۷.۲)$$

ضمن اینکه $\hat{y}_{ij}$ برآوردی برای واحد jام در ناحیه‌ی iام می‌باشد. هنگامی که اندازه نمونه‌ی n_i نسبت به N_i خیلی کوچک باشد برآوردگر مورد استفاده به شرح زیر است:

$$\hat{\bar{\mathbf{Y}}}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbf{x}_{ij}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_i = \bar{\mathbf{X}}_i\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_i$$

۷.۲ برآوردگرهای غیرمستقیم و روش های مبتنی بر مدل

۱.۷.۲ برآوردگر ترکیبی

هنگامی که متغیری مدام اندازه‌گیری شود، مدل خطی می‌تواند با ایجاد یک ارتباط بین متغیر هدف و متغیرهای کمکی استفاده شود. فی ــ هریوت مدل زیر را که ترکیبی از برآوردگر مستقیم و رگرسیونی می‌باشد پیشنهاد داده است:

$$\bar{\mathbf{Y}}_\mathbf{i} = \boldsymbol{\beta}\bar{\mathbf{X}}_\mathbf{i} + \epsilon_i; \qquad \epsilon_i \sim N(0, \mathbf{V}_\mathbf{i}^2) \qquad (۳۸.۲)$$

بنابراین $\mathbf{V}_\mathbf{i}^2$ واریانس طرحی است. برآوردگر ترکیبی می‌تواند با استفاده از برآورد β در مدل فوق به‌دست آورده شود، پس خواهیم داشت:

$$\hat{\bar{\mathbf{Y}}}_{\mathbf{SA},\mathbf{i}} = \hat{\boldsymbol{\beta}}\bar{\mathbf{X}}_\mathbf{i}$$

$\hat{\bar{\mathbf{Y}}}_{\mathbf{SA},\mathbf{i}}$ یک برآوردگر ترکیبی در ناحیه‌ی iام است. همانند روش‌های قبلی متغیرهای کمکی برای همه نواحی کوچک در دسترس‌اند. این رویکرد می‌تواند پاسخ‌های غیرنرمال را از طریق

[۱۲] Maximum likelihood

[۱۳] quasi-probability penalized

[۱۴] Restricted maximum likelihood

میانگین‌های مدل آمیخته خطی تعمیم‌یافته در پاسخ و پیش‌بینی خطی با یک تابع متناسب که مرتبط می‌باشد، گسترش دهد. برآورد واریانس را می‌توان به‌شیوه‌ی زیر انجام داد:

$$var[\hat{\bar{\mathbf{Y}}}_{\mathbf{SA},\mathbf{i}}] = E[(\hat{\bar{\mathbf{Y}}}_{\mathbf{SA},\mathbf{i}} - \hat{\boldsymbol{\beta}}\bar{\mathbf{X}}_{\mathbf{i}} - \epsilon_i)^2] = \bar{\mathbf{X}}_{\mathbf{i}}^{\mathbf{T}} var[\hat{\boldsymbol{\beta}}]\bar{\mathbf{X}}_{\mathbf{i}} + \mathbf{V}_{\mathbf{i}}^{\mathbf{2}}$$

معادله مدل در سطح واحد در این برآوردگر مشمول نمونه‌های در دسترس برای هر ناحیه و تحت مدل‌سازی زیر پیشنهاد شده‌است:

$$\mathbf{y}_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{e}_{ij}$$

$\mathbf{e}_{ij}$ خطای فردی و دارای توزیع نرمال با میانگین ۰ و واریانس σ_i^2 است. فرض کنیم تغییرات در داخل ناحیه داخلی می‌تواند از ناحیه‌ای به ناحیه‌ی دیگر متفاوت باشد. اگر نسبت n_{i/N_i} خیلی کوچک باشد همبستگی جمعیت کل را می‌توان چشم‌پوشی کرد و برآوردگر ترکیبی مدل را در سطح واحد به‌صورت زیر نشان داد:

$$\hat{\bar{\mathbf{Y}}}_{\mathbf{SB},\mathbf{i}} = \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbf{x_{ij}}\hat{\boldsymbol{\beta}} = \bar{\mathbf{X}}_{\mathbf{i}}\hat{\boldsymbol{\beta}}$$

هرچند این نسبت برای مدل‌های غیرخطی یا غیرنرمال صورت نمی‌پذیرد. این نوع از برآوردگر ترکیبی به‌طور گسترده توسط اداره آمار ملی در انگلیس برای ایجاد گزارش‌های متفاوت ناحیه‌ای مورد استفاده قرار می‌گیرد. یار و همکاران (۲۰۰۲) و هدی و همکاران (۲۰۰۳) متکی به نوعی از مدل در سطح واحد می‌باشند، اما با توجه به اینکه هیچ متغیر کمکی در دسترس در سطح فردی نیست آن را به‌صورت $\mathbf{x_{ij}} = \mathbf{X_i}$ درنظر می‌گیرند. واریانس این برآوردگر به‌صورت زیر خواهد بود:

$$var[\hat{\bar{\mathbf{Y}}}_{\mathbf{i}}] = \bar{\mathbf{X}}_{\mathbf{i}}^{\mathbf{T}} var\hat{\boldsymbol{\beta}}\bar{\mathbf{X}}_{\mathbf{i}} + \sigma_i^2/N_i$$

همچنین توجه کنید، برآورد ترکیبی مدل در سطح ناحیه برآوردگری برای ناحیه را فراهم می‌کند، برآوردگر مدل در سطح واحد مقادیر متغیر مورد نظر را برای افرادی که نمونه‌برداری نشده‌اند ($\hat{y}_{ij} = \hat{\beta}x_{ij}, \quad i$ در ناحیه $\forall j$) و برآوردگرهایی که باید بعد از آن به‌طور متوسط بر روی تمام افراد در ناحیه محاسبه شوند را در نظر می‌گیرد.

۸.۲ برآوردگرهای مبتنی بر مدل صریح

بعد از انتخاب مدل مناسب و گزینش متغیرهای کمکی خوب که با متغیر تحت بررسی وابستگی نسبتا خوبی را داشته باشد، نوبت به انتخاب برآوردگر مناسب برای نواحی کوچک است. به‌طور حتم دو رویکرد متفاوت در یافتن برآورد وجود دارد که یکی، رویکرد فراوانی‌گرای کلاسیک و دیگری روش بیزی می‌باشد. از طرفی می‌دانید که برآوردگر غیرمستقیم مبتنی بر مدل صریح را، برآوردگر غیرمستقیم مدرن گویند. این نوع از برآوردگرها، برآوردگرهای مبتنی بر مدل صریح

هستند که به ۴ دسته تقسیم می‌شوند و همین طور این ۴ دسته زیر گروه همان روش‌های فراوان‌گرای کلاسیک و روش بیزی‌اند پس، عبارت‌اند از:

۱. برآوردگرهای پیشگوی نااریب خطی تجربی BLUP

۲. برآوردگرهای بهترین پیشگوی نااریب خطی تجربی EBLUP

۳. برآوردگرهای بیز تجربی EB

۴. برآوردگرهای بیزی سلسله مراتبی HB

در ادامه درباره‌ی این برآوردگرها مطالبی را ارائه خواهیم داد (گوش و رائو، ۱۹۹۴).

۱.۸.۲ برآوردگرهای پیشگوی نااریب خطی تجربی BLUP

میانگین و مجموع نواحی کوچک را می‌توان به‌صورت ترکیب‌های خطی اثرهای ثابت و تصادفی بیان کرد. بنابراین روش برآوردگرهای پیشگوی نااریب خطی برای این دسته از پارامترها را می‌توان در چارچوب بسامدگرای کلاسیک، با مرتبط قرار دادن به نتایج کلی در مورد روش برآورد BLUP به‌دست آورد. این بهترین در این جمله یعنی کمترین میانگین توان دوم خطا را داشته باشد. این نوع از برآوردگرها در رده‌ی برآوردگرهای نااریب خطی MSE را مینیمم می‌سازد و همچنین به نرمال بودن اثرهای تصادفی وابسته نیست (مک کولا و سیرل، ۲۰۰۱). این دسته از برآوردگرها، واریانس‌ها و کوواریانس‌های اثرهای تصادفی را به کمک روش برازش ثابت‌ها یا روش گشتاوری برآورد می‌کند. یکی دیگر از روش‌های که در دسترس است و به‌وسیله‌ی آن می‌توان برآورد مولفه‌های واریانس و کوواریانس را به‌دست آورد و نیز با لحاظ فرض نرمال بودن اثرها، روش‌های ماکسیمم درستنمایی و یا ماکسیمم درستنمایی مقید می‌باشد. با به‌کارگیری این مولفه‌های برآوردشده در برآوردگر BLUP، برآوردگری دو مرحله‌ای به‌دست می‌آید که در مقایسه با برآوردگر بیز تجربی بهتر عمل می‌کند و از آن به‌عنوان برآوردگر BLUP تجربی یا EBLUP یاد می‌شود (هارویل، ۱۹۹۱).

بنابراین با توجه به مدلی که قبلا ارائه شده بود، برآوردگر را می‌توان به شرح زیر نشان داد:

$$E[\mathbf{u}|\mathbf{y}] = \hat{\mathbf{u}} = \mathbf{D}\mathbf{Z}^T{}_1\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta})$$

هنگامی که اثرات تصادفی در این روش برآورده شوند؛ معادله‌ی (۳۷.۲) به یک برآوردگر BLUP تبدیل می‌شود و مجموع برآوردگر مستقیم مبتنی بر نمونه در ناحیه‌ی iام می‌شود و قسمت ثابت مبتنی بر متغیرهای کمکی و برآوردگری از اثرتصادفی است.

توجه کنید اگر در ناحیه‌ی iام هیچ نمونه‌ای دردسترس نباشد برآوردگر BLUP به برآوردگر ترکیبی کاهش می‌یابد، زیرا، برآوردگر مستقیم وجود ندارد و $\hat{\mathbf{u}} = 0$ است. هنگامی که واریانس اثرات تصادفی ناشناخته باشد و از داده برآورد شود، برآوردگرهای مشابه می‌توانند توسعه یابند.

بنابراین موارد بیان شده هم، بهترین برآوردگر نااریب خطی تجربی BLUP نامیده می‌شود. رائو (۲۰۰۳) در مورد استفاده از برآوردگرهای BLUP و EBLUP در برآورد کوچک ناحیه‌ای پیشنهادهایی را مطرح نمود.

برای مدل در سطح ناحیه برآورد میانگین توان دوم خطا MSEاغلب با تقریب ۳ مولفه محاسبه می‌شود که به ترتیب عبارت‌اند از: $(\mathbf{G}_1)$ نشان‌دهنده‌ی عدم اطمینان دربرآوردگر ، $(\mathbf{G}_2)$ برآورد β و $(\mathbf{G}_3)$ برآورد واریانس σ^2 می‌باشند. که اغلب نیز به‌صورت زیر نشان داده می‌شود:

$$MSE[\hat{\bar{\mathbf{Y}}}_\mathbf{i}] \approx \mathbf{G}_1 + \mathbf{G}_2 + \mathbf{G}_3$$

از طرفی MSE برای مدل در سطح واحد نیز مشابه مدل در سطح ناحیه انجام می‌گیرد. برای توضیح بیشتر به روش‌های EBLUP می‌پردازیم.

۲.۸.۲ برآوردگرهای بهترین پیشگوی نااریب خطی تجربی EBLUP

برآوردگر مطرح شده در دو مرحله به‌دست می‌آید. همان‌طوری که از نام این برآوردگرمشخص است در مرحله‌ی اول، بهترین پیشگوی نااریب خطی در رده‌ی برآوردگرهای نااریب خطی، زمانی به‌دست می‌آید که مرتبط به کمیت مورد بررسی باشد و میانگین توان دوم خطا را مینیمم سازد. این دسته از برآوردگرها به واریانس و کوواریانس اثر تصادفی در مدل بستگی دارند. در مرحله‌ی بعدی، برآوردگر بهترین پیشگوی نااریب خطی، زمانی به‌دست می‌آید که برآوردگرهای مناسب را برای پارامترهای واریانس و کوواریانس، در بهترین پیشگوی نااریب خطی در مرحله‌ی اول جایگزین نماییم. همچنین برای به‌دست آوردن برآورد واریانس‌ها معمولا از روش گشتاوری یا روش لاکرانژ استفاده می‌شود. و گاهی نیز با فرض نرمال بودن، می‌توان از روش دیگری برای برآورد مولفه‌های واریانس استفاده کرد که برای این عمل روش‌های ماکسیمم درستنمایی یا ماکسیمم مقید معرفی شده است.

با توجه به مطالب بیان شده اکنون به توضیح روش برآوردگر بهترین پیشگوی نااریب خطی تجربی پرداخته می‌شود. فرض کنید داده‌های نمونه‌ای از یک مدل آمیخته‌ی خطی به صورت زیر پیروی می‌کند، لذا خواهیم داشت:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}. \qquad (۳۹.۲)$$

- $\mathbf{y}$ برداری از مشاهدات $\mathbf{y} = (y_1^T, ..., y_n^T)^T$ و $\mathbf{y}_i$ برداری با ابعاد $n \times 1$

- $\mathbf{X} = col_{1 \le i \le m}(\mathbf{X}_i)$ و $\mathbf{X}_i$ ماتریسی با ابعاد $(n_i \times p)$،

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ . \\ . \\ . \\ \mathbf{X}_m \end{pmatrix}$$

- $\mathbf{Z} = diag_{1 \le i \le m}(\mathbf{Z}_i)$ و $\mathbf{Z}_i$ ماتریسی با ابعاد $(n_i \times h_i)$

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & 0 & ... & 0 \\ 0 & \mathbf{Z}_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \mathbf{Z}_m \end{pmatrix}$$

همان‌طوری که همانند بالا اشاره داشته‌ایم، $\mathbf{y}$ بردار مشاهدات نمونه‌ای با بعد $n \times 1$، $\mathbf{X}$ و $\mathbf{Z}$ ماتریس‌های معلومی هستند که به ترتیب با ابعاد $n_i \times p_i$ و $n_i \times h$ نمایش داده می‌شوند. $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ برداری از ضرایب رگرسیونی و همچنین $\mathbf{v}$ بردار اثرات تصادفی که با ابعاد $(h_i \times 1)$ می‌باشند. از طرفی $\mathbf{v}$ و $\mathbf{e}$ نیز دارای توزیعی‌هایی با بردار میانگین 0 و ماتریس کوواریانس $\sum_v$ و $\sum_e$اند، در نتیجه مستقل نیز می‌باشند. ماتریس کوواریانس‌های اشاره شده، دارای مولفه‌های واریانسی می‌باشند که این مولفه‌ها در بردار $\boldsymbol{\delta} = (\delta_1, ..., \delta_q)^T$ آورده می‌شوند. ماتریس کوواریانس برای مدل (۳۹.۲) به‌صورت زیر به‌دست می‌آید:

$$Var(\mathbf{y}) = V = V(\boldsymbol{\delta}) = \boldsymbol{\Sigma_e} + \mathbf{Z}\boldsymbol{\Sigma_v}\mathbf{Z}^T,$$

که برای تمامی δها، نامنفرد است.

ترکیب خطی زیر که از ضرایب رگرسیونی $\boldsymbol{\beta}$ و مقادیر تحقق‌یافته‌ی $\mathbf{v}$ است را برای بردارهای ثابت و معین $\boldsymbol{\eta}$ و $\mathbf{m}$ در نظر بگیرید.

$$\boldsymbol{\mu} = \boldsymbol{\eta}^T\boldsymbol{\beta} + \mathbf{m}^T\tilde{\boldsymbol{v}}$$

در روش بهترین پیشگوی نااریب خطی تجربی در پی‌دستیابی برآوردگر $\boldsymbol{\mu}$ یعنی همان $\hat{\boldsymbol{\mu}}$ هستیم، که این نوع از برآوردگر برای $\boldsymbol{\mu}$ نااریب می‌باشد. و اگر فرض بر این باشد که، پارامترهای δ معلوم باشند، بنابراین پیشگوی نااریب خطی به صورت زیر به‌دست می‌آید:

$$\tilde{\boldsymbol{\mu}}^{\boldsymbol{H}} = t(\boldsymbol{\delta}, \boldsymbol{y}) = \boldsymbol{\eta}^T\tilde{\boldsymbol{\beta}} + \boldsymbol{m}^T\tilde{\boldsymbol{v}}, \qquad (۴۰.۲)$$

بالانویس H در $\tilde{\boldsymbol{\mu}}^{\boldsymbol{H}}$ نشان‌دهنده‌ی نام هندرسون[۱۵] است. و $\tilde{\boldsymbol{\beta}}$ بهترین برآوردگر نااریب خطی $\boldsymbol{\beta}$ است و آن را به صورت زیر نشان می‌دهیم:

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\delta}) = (\mathbf{X^T V^{-1} X})^{-1}\mathbf{X^T V^{-1} Y}$$

[۱۵] Henderson

و $\tilde{v}$ برابر است با

$$\tilde{\mathbf{v}} = \tilde{v}(\boldsymbol{\delta}) = \boldsymbol{\Sigma}_v \mathbf{Z}^T \mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}).$$

بنابراین $\tilde{\boldsymbol{\mu}}^H \tilde{\boldsymbol{\mu}}$ برآوردگر پیشگوی ناریب خطی است که در رده‌ی برآوردگرهای خطی ناریب قرار دارد که بهترین برآوردگر می‌باشد. برآوردگر $\boldsymbol{\delta}$ که مرتبط با عناصر $\boldsymbol{\Sigma}_v$ و $\boldsymbol{\Sigma}_e$ می‌باشد، فرض می‌شود که معلوم باشد در حالی که در عمل این‌گونه عمل نمی‌باشد و عبارت $\boldsymbol{\delta}$ را باید به روش‌های مختلفی همانند روش گشتاوری، حداقل مربعات تعمیم‌یافته، ماکسیمم درستنمایی و ماکسیمم درستنمایی مقید برآورده شود و در صورت برآورده شدن جایگزین می‌شود که بهترین برآوردگر پیشگوی ناریب خطی تجربی به‌دست آید.

۳.۸.۲ برآوردگر بیز تجربی EB

برای دستیابی به این نوع از برآوردگر، کافی است از سه مرحله‌ی زیر استفاده کنید:

مرحله‌ی اول: برای به‌دست آوردن توزیع پسین پارامترهای مورد نظر $\boldsymbol{\mu}$ بنابر شرط $\mathbf{y}$ یعنی $f(\boldsymbol{\mu}|\mathbf{y}, \boldsymbol{\lambda})$ نیاز است، لذا برای به‌دست آوردن آن از دو توزیع شرطی $f(\boldsymbol{\mu}|\mathbf{y}, \boldsymbol{\lambda}_2)$ و $f(\boldsymbol{\mu}|\mathbf{y}, \boldsymbol{\lambda}_1)$ که در آن بردار پارامتر مدل، مربوط به عبارت $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ می‌باشد، استفاده می‌کنند.

مرحله‌ی دوم: بنابر پارامترهای مدل گفته شده یعنی همان $\boldsymbol{\lambda}$ ها، از تابع چگالی حاشیه‌ای $f(\boldsymbol{y}|\boldsymbol{\lambda})$ و به کمک روش‌های ماکسیم درستنمایی، گشتاوری و غیره برآورد می‌شوند.

مرحله‌ی سوم: برای به‌دست آوردن پارامتر $\boldsymbol{\mu}$ از تابع چگالی پسین برآورد شده‌ی $f(\boldsymbol{\mu}|\mathbf{y}, \hat{\boldsymbol{\lambda}})$ استفاده می‌کنند و لذا برآوردگر $\hat{\boldsymbol{\lambda}}$ مرتبط به برآوردگر $\boldsymbol{\lambda}$ می‌باشد.

به‌طور کلی بهترین پیشگو برای یک کمیت، زمانی به‌دست می‌آید که در روش بیزی با در نظر گرفتن تابع زیان دوم خطاها، امید شرطی کمیت تحت مطالعه برای ناحیه‌ی کوچک بنابر اینکه داده‌ها و پارامترهای مدل داده شده باشند، به‌دست آورده می‌شود و همچنین برای این محاسبات بعضی فرض‌های توزیعی مورد نیاز است.

برآوردگر بیز تجربی با استفاده از برآوردگرهای مناسب پارامترهای مدل و با به‌کارگیری از روش‌های ماکسیمم درستنمایی و روش گشتاوری به‌دست می‌آید. بنابراین همانند برآوردگر بهترین پیشگوی خطی ناریب تجربی، مدل (۳۹.۲) را در نظر بگیرید، که دارای توزیع‌های زیر می‌باشند:

$$\mathbf{y}_i|\mathbf{v}_i \overset{iid}{\sim} N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{v}_i, \boldsymbol{\Sigma}_{e_i}),$$

$$\mathbf{v}_i \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}_{v_i}), \quad i = 1, ..., m.$$

ازاین رو، $\boldsymbol{\Sigma}_{e_i}$ و $\boldsymbol{\Sigma}_{v_i}$ پارامترهای واریانسی می‌باشند که با $\boldsymbol{\delta}$ نشان داده می‌شوند، و در معادله بالا $i = 1, ..., m$ نیز نشان دهنده‌ی نواحی کوچک هستند. اما برای بردارهای ثابت و معین $\boldsymbol{\eta}$

و $\boldsymbol{m}$ بنابر فرض اگر رابطه‌ی زیر برقرار باشد،

$$\mu_i = \eta_i^T \beta + m_i^T v_i$$

آن‌گاه برآورد پیشگوی بیزی به‌صورت امید شرطی μ_i مشروط بر y_i، $\boldsymbol{\beta}$ و $\boldsymbol{\delta}$ است که به‌صورت زیر نشان داده می‌شود:

$$\mu_i^B = \hat{\mu_i}^B(\boldsymbol{\beta}, \boldsymbol{\delta}) = E(\mu_i|y_i, \boldsymbol{\beta}, \boldsymbol{\delta}) = \eta_i^T \boldsymbol{\beta} + m_i^T \hat{v_i}^B; \qquad (۴۱.۲)$$

و در این برآوردگر

$$v_i^B = E(v_i|y_i, \boldsymbol{\beta}, \boldsymbol{\delta}) = \Sigma_{v_i} \mathbf{Z}_i^T \boldsymbol{V}_i^{-1}(y_i - X_i\boldsymbol{\beta}),$$

$$V_i = \Sigma_{e_i} + \boldsymbol{Z}_i \boldsymbol{\Sigma_{v_i}} \mathbf{Z}_i^T,$$

بنابراین $\hat{\mu_i}^B$ از توزیع پسین زیر به‌دست می‌آید:

$$\mu_i|y_i, \boldsymbol{\beta}, \boldsymbol{\delta} \overset{iid}{\sim} N(\hat{\mu_i}^B, g_i(\boldsymbol{\delta}))$$

پس

$$g_i(\boldsymbol{\delta})) = m_i^T(\Sigma_{v_i} - \Sigma_{v_i} Z_i^T V_i^{-1} \Sigma_{v_i}) m_i.$$

در رابطه‌ی بالا یعنی مدل (۴۱.۲)، دیده می‌شود که این برآوردگر به پارامترهای $\boldsymbol{\beta}$ و $\boldsymbol{\delta}$ وابسته است، به‌همین منظور لازم می‌شماریم که این پارامترها برآورد شوند. در مرحله‌ی بعدی، این پارامترها برای $i = ۱, ..., m$ از توزیع حاشیه‌ای $y_i \overset{iid}{\sim} N(\mathbf{X_i}\boldsymbol{\beta}, \mathbf{V_i})$ با به‌کارگیری از روش‌های ماکسیمم درستنمایی یا ماکسیمم درستنمایی مقید برآورده می‌شوند و به‌صورت $\hat{\boldsymbol{\beta}}$ و $\hat{\boldsymbol{\delta}}$ نشان داده می‌شوند و بعد در پارامتر $\hat{\mu_i}^B$ قرار می‌گیرند که بدین صورت برآوردگر بیز تجربی به صورت زیر خواهد بود:

$$\hat{\mu_i}^{EB} = \hat{\mu_i}^{EB}(\hat{\beta}, \hat{\delta}) = \eta_i^T \hat{\beta} + m_i^T \hat{v_i}^B(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}).$$

پس، $\hat{\mu_i}^{EB}$ میانگین برآورد شده‌ی توزیع پسین عبارت $f(\mu_i|y_i, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}})$ به شکل $N(\hat{\mu_i}^{EB}, g_i(\hat{\boldsymbol{\delta}}))$ می‌باشد.

۴.۸.۲ برآوردگر بیز سلسه‌مراتبی HB

اغلب اوقات روش بیز سلسله‌مراتبی را به‌عنوان روش کاملا بیزی می‌نامند، چرا که توزیع پیشین را برای پارامترهای مدل فرض می‌کنند و سپس نتیجه‌گیری را براساس توزیع پسین انجام می‌دهند. به‌طورکلی، تحت میانگین مربع خطا، پارامتر نواحی کوچک توسط میانگین‌های پیشین برآورد می‌شوند و عدم حتمیت توسط واریانس پسین اندازه‌گیری می‌شود (گلفند و اسمیت، ۱۹۹۰). روش بیز سلسه‌مراتبی، روش کاملا ساده و قابل درک می‌باشد. هر چند

نیازمند محاسبات پیچیده همراه با انتگرال‌گیری‌های چند بعدی است. پیشرفت‌های اخیر در ابعاد محاسباتی، به‌ویژه در نمونه‌گیری گیبز در رویکرد بیزی قضیه بیز را بسیار پرکاربرد و ساده کرده است. در روش سلسله‌مراتبی برای اینکه بتوان برآوردگری را به‌دست آورد، در ابتدا کافی است توزیع پسین $f(\boldsymbol{\mu}|\mathbf{y})$ را محاسبه کرد. لذا برآوردگر پارامتر مورد نظر $\phi = h(\boldsymbol{\mu})$ از طریق برآورد میانگین پسین $\hat{\phi}^{HB} = E(h(\boldsymbol{\mu})|\boldsymbol{y})$ به‌دست می‌آید و برای دقت نیز از عبارت $V(h(\boldsymbol{\mu})|\boldsymbol{y})$ استفاده می‌شود.

همان‌طوری که می‌دانید برای پیدا کردن توزیع پسین $f(\boldsymbol{\mu}|\mathbf{y})$ به جای برآورد $\boldsymbol{\lambda}$، از توزیع پیشین $f(\boldsymbol{\lambda})$ استفاده می‌کنند که روی مدل تعریف شده است، چرا که با این عمل توزیع پسین مورد نظر $f(\boldsymbol{\mu}|\mathbf{y})$ برای پارامترهای کوچک ناحیه‌ای برآورده می‌شوند. برای به‌دست آوردن توزیع پسین با استفاده از قضیه‌ی بیز، کافی است مدل را در دو مرحله برای داده‌ها و پارامترها با استفاده از توزیع‌های $f(\boldsymbol{\mu}|\lambda_۲)f(\boldsymbol{y}|\boldsymbol{\mu},\lambda_۱)$ که $\boldsymbol{\lambda} = (\lambda_۱^T, \lambda_۲^T)^T$ بردار پارامترهای مدل می‌باشد، صورت گیرد. با بهره‌گیری از قضیه‌ی بیز، برای به‌دست آوردن توزیع پسین پارامترهای ناحیه‌ی کوچک همانند رابطه‌ی (۱.۵) به‌عمل می‌آید. برآوردگر بیز سلسله‌مراتبی را می‌توان با استفاده از امیدریاضی بر روی توزیع پسین، برای پارامترهای کوچک ناحیه‌ای به‌دست آورد. همان‌طوری که در به‌دست آوردن این امیدریاضی، اغلب اوقات با انتگرال‌های پیچیده و با ابعاد بالا مواجه می‌شویم پس، محاسبات آن‌ها سختی نیز به همراه خواهد داشت. برای اینکه بتوان از سختی محاسبات کم کرد، بهتر است از روش‌های دیگر استفاده کرد و به همین دلیل به‌کاربردن از انواع روش‌های مونت کارلوی زنجیره‌ی مارکوفی که شامل نمونه‌گیری گیبز است، ملزوم می‌باشد. اساس عملکرد این روش‌ها، ساختن نمونه‌های شبیه‌سازی برای تمامی پارامترهای مدل $\boldsymbol{\eta} = (\lambda_۱, \lambda_۲, \boldsymbol{\mu})$ است، به‌طوری که این نمونه‌ها یک زنجیره‌ی مارکوفی $\boldsymbol{\eta}^{(k)}, K = ۰, ۱, ۲, ...$ که در آن توزیع $\boldsymbol{\eta}^{(k)}$ به توزیع پسین یکتای $\pi(\boldsymbol{\eta}) = f(\boldsymbol{\eta}|\mathbf{y})$ همگرا می‌شود، را تشکیل می‌دهند. در الگوریتم به‌کار گرفته شده، برای ایجاد نمونه‌ها ابتدا از نقطه‌ی اولیه $\boldsymbol{\eta}^{(۰)}$ بعد از انتخاب f نمونه‌ی اولیه، F نمونه‌ی $\boldsymbol{\eta}^{(f+۱)}, ..., \boldsymbol{\eta}^{(f+F)}$ را از توزیع هدف $f(\boldsymbol{\eta}|\mathbf{y})$ به‌دست می‌آورند. و در انتها میانگین نمونه‌های به‌دست آمده، برآورد پارامتر را نتیجه می‌دهند. چند تن از افرادی که در این زمینه کار کرده‌اند به شرح زیر می‌باشد:

بل (۱۹۹۹) با به‌کارگیری از روش بیز سلسله‌مراتبی در مدل سطح ناحیه، برآورد درآمد متوسط و نیز تعداد کودکان زیر خط فقر را در همه‌ی ایالت‌های متحده آمریکا به‌دست آورد. داتا و همکاران (۱۹۹۶) درآمد متوسط خانوارهای ۴نفره ایالت متحده را با استفاده از مدل‌های HB انجام دادند. و همچنین داتا و همکاران در سال (۱۹۹۹) با این روش بیان شده توانستند نرخ بیکاری را برای همه ایالت های متحده به‌دست آورند.

جمع‌بندی

در این فصل سعی بر آن بوده است که مدل‌های برآورد کوچک ناحیه‌ای را معرفی کنیم و همچنین اهمیت هر کدام از این مدل‌ها را با ذکر مثال‌هایی تشریح نمودیم. لذا تمامی مدل‌های بیان شده در این فصل اهمیت دارد ولی، آنچه که برایمان در این پایان‌نامه مهم می‌باشد، به‌کارگیری از مدل‌های فضایی است، چرا که این مدل‌ها موجب افزایش دقت برآوردگر مورد نظر می‌شوند. به همین منظور، در فصل بعدی مدل بیزی برآورد کوچک ناحیه‌ای را مطرح می‌نماییم.

فصل ۳

مدل‌بندی بیزی برآورد کوچک ناحیه‌ای

برای بالا بردن دقت برآوردگرها و کاهش واریانس از برآوردگرهای غیرمستقیم در ناحیه‌ی کوچک استفاده می‌شود و به‌کارگیری از روش‌های بیزی نیز راه حل مناسبی برای موارد فوق خواهد بود. در این فصل به بیان اهمیت روش‌های بیزی در برآورد کوچک ناحیه‌ای خواهیم پرداخت. لذا استفاده کردن از رهیافت بیزی، توزیع پیشین مناسبی را برای پارامترهای مدل ارائه خواهد داد. از طرفی با اعمال محدودیت روی پارامتر فضایی و تعریف مناسب ماتریس همسایگی، مدل CAR برای مدل پیوند مطرح می‌شود. بنابراین برای تعیین برآوردهای بیز، با در نظر گرفتن پیشین مناسب و استفاده از مدل‌های مطرح شده، توزیع پسین پارامترها به‌دست می‌آید و پارامتر مورد علاقه به روش بیز برآورد می‌شود. مطالب مطرح شده در این فصل، از مقاله استنباط بیزی برآورد کوچک ناحیه‌ای ریچاردسون و لی در سال (۲۰۱۰) گرفته شده است.

۱.۳ روش‌های بیزی

در دیدگاه کلاسیک، مدل‌بندی بر اساس تابع درستنمایی است. ولی در دیدگاه بیزی، مقدار واقعی پارامتر تحققی از یک توزیع تصادفی می‌باشد. در استنباط بیزی، علاوه بر مشخص کردن مدل برای داده‌های مشاهده شده‌ی بردار $\mathbf{y}$ به شرط یک بردار از پارامترهای مجهول $\boldsymbol{\theta}$ به‌صورت توزیع $f(\mathbf{y}|\boldsymbol{\theta})$، فرض می‌شود که $\boldsymbol{\theta}$ یک کمیت تصادفی با توزیع پیشین $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})$ و $\boldsymbol{\eta}$

نیز یک بردار ابر پارامتر می‌باشد. استنباط بیزی بر اساس توزیع پسینی به‌صورت

$$\pi(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\eta}) = \frac{\pi(\boldsymbol{\theta},\mathbf{y}|\boldsymbol{\eta})}{f(\mathbf{y}|\boldsymbol{\eta})} = \frac{\pi(\boldsymbol{\theta},\mathbf{y}|\boldsymbol{\eta})}{\int f(\mathbf{y},\mathbf{u}|\boldsymbol{\eta})d\mathbf{u}} = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\eta})}{\int f(\mathbf{y}|\mathbf{u})\pi(\mathbf{u}|\boldsymbol{\eta})d\mathbf{u}},$$

محاسبه می‌شود که این قضیه به قضیه‌ی بیز بر می‌گردد. اگر $\boldsymbol{\eta}$ معلوم باشد آن‌گاه در نوشتن $\pi(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\eta})$، $\boldsymbol{\eta}$ نوشته نمی‌شود. در دیدگاه بیزی، همه‌ی استنباط‌ها مبنی بر توزیع پسین به‌دست می‌آید.

انواع توزیع‌های پیشینی

به‌طور کلی توزیع‌های پیشینی دارای نوع‌های مختلفی می‌باشند، اما تنها ۴ نوع از آن‌ها را شرح می‌دهیم:

۱. توزیع پیشین سره[۱]: به توزیعی گویند که در آن

$$\int_{\Theta} \pi(\theta)d\theta = ۱$$

باشد.

۲. توزیع پیشین ناسره[۲]: به توزیعی گویند که در آن

$$\int_{\Theta} \pi(\theta)d\theta = +\infty$$

باشد.

۳. توزیع پیشین ناآگاهی‌بخش[۳]: یکی از مهمترین پیشین‌ها، پیشین ناآگاهی‌بخش می‌باشد و در مواردی که اطلاع چندانی در مورد پارامتر موجود نباشد، مورد استفاده قرار می‌گیرد. در این حالت محقق احتمال یکسانی را به همه پیشامدهای ممکن تخصیص می‌دهد؛ یعنی $\pi(\theta) = k$.

۴. توزیع پیشین مبهم[۴]: توزیع پیشین سره‌ای که واریانس خیلی بزرگ دارد، به‌طوری که به یک پیشین ناآگاهی بخش نزدیک می‌شود.

پیشین‌های متفاوتی را می‌توان در تحلیل بیزی در نظر گرفت. انتخاب توزیع پیشین یکی از مسائل مهم در آمار بیزی می‌باشد، زیرا این توزیع تاثیر نسبتا زیاد و قابل توجهی در نتیجه‌ی استنباط دارد. پس، این انتخاب باید به بهترین نحو انجام شود. بنابر مطالب فوق، تحلیل بیزی، مبنی بر معرفی پیشین‌ها، توزیع پسینی و توزیع‌های شرطی کامل است؛ که مهم‌ترین روش در این پایان‌نامه مدل‌های سلسله‌مراتبی بیزی می‌باشد.

---

[۱] Proper priori

[۲] ImProper priori

[۳] informative priori

[۴] vague priori

۲.۳ ساختار سلسله‌مراتبی مدل‌بندی بیزی برای نواحی کوچک

فرض کنید $\mathbf{y}_{(n)} = (y_1, ..., y_n)$ بردار نمونه‌گیری شده با اندازه‌ی n است. ساختار مدل‌بندی مدل سلسله‌مراتبی کلی زیر را در نظر می‌گیریم:

مرتبه‌ی اول: $\mathbf{y}_{(n)}|\mathbf{X} = \mathbf{x} \sim h(\mathbf{y}_{(n)}|\mathbf{X} = \mathbf{x}, \boldsymbol{\theta}_1)$

مرتبه‌ی دوم: $\boldsymbol{X} \sim g(\mathbf{x}|\boldsymbol{\theta}_2)$

در اینجا $\boldsymbol{y}_{(n)}$ قابل مشاهده اما $\mathbf{x}$ (پنهان) غیرقابل مشاهده است. پارامترهای مورد علاقه برای استنباط، پارامترهای $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ هستند. هدف تحلیل برآورد پارامترهای $\boldsymbol{\theta}$ و پیشگویی وضعیت پنهان $\mathbf{x}$ است. تابع درستنمایی ساختار مدل‌بندی سلسه‌مراتبی به صورت زیر است:

$$L(\boldsymbol{\theta}, \mathbf{y}_{(n)}) = \int h(\mathbf{y}_{(n)}|\mathbf{x}, \boldsymbol{\theta}_1) g(\mathbf{x}, \boldsymbol{\theta}_2) d\mathbf{x}$$

مشکلات مرتبط در استفاده از این تابع برای استنباط بیشتر، محاسباتی هستند، به این معنا که:

۱. به‌طور کلی محاسبه‌ی تابع درستنمایی شامل انتگرال‌هایی با بعد بسیار زیاد است.

۲. به‌دست آوردن مقداری از $\boldsymbol{\theta}$ که $L(\boldsymbol{\theta}; \mathbf{y}_{(n)})$ را ماکسیمم سازد، با به‌کار بردن روش جستجوی عددی به دلیل طبیعت تصادفی درستنمایی برآورد شده بسیار مشکل‌ساز است؛ دقت کنید که طبیعت تصادفی تابع درستنمایی در حقیقت ناشی از متفاوت بودن مقادیر نمونه در هر نمونه‌گیری مجدد است.

۳. و در آخر محاسبه‌ی خطاهای معیار به‌دست آمده از پارامترهای برآورد شده در روش محاسبه‌ی عددی مشتقات مرتبه‌ی دوم تابع لگ درستنمایی، شامل مشکلات و سختی‌های بیشتری می‌باشد.

۳.۳ بررسی رویکرد بیزی

فرض کنید مشاهدات $\mathbf{y}_{(n)} = (y_1, ..., y_n)$ یک نمونه‌ی nتایی از مدل آماری سلسه‌مراتبی زیر باشد:

$$\mathbf{y}_{(n)} \sim f(\mathbf{y}_{(n)}|\mathbf{X}, \boldsymbol{\theta}_1)$$
$$\mathbf{X} \sim g(\mathbf{x}|\boldsymbol{\theta}_2)$$

که در آن f و g توابع چگالی توأم، $\mathbf{X}$ بردار کمیت‌های تصادفی یا فرآیندهای موثر بر مشاهدات، و $\boldsymbol{\theta}_1 = (\boldsymbol{\theta}_{11}, ..., \boldsymbol{\theta}_{1q})$ بردار پارامترهای ثابت و نامعلوم موثر بر مشاهدات، و $\boldsymbol{\theta}_2 = (\boldsymbol{\theta}_{21}, ..., \boldsymbol{\theta}_{2p})$

بردار پارامترهای ثابت و نامعلوم مرتبط با فرآیند $\mathbf{X}$ است. دقت کنید که در ادبیات مدل‌های سلسه‌مراتبی، کمیت‌های تصادفی درون $\mathbf{X}$ را اثرات تصادفی، متغیر پنهان یا وضعیت سیستم گویند.

تابع درستنمایی برای مدل کلی سلسه‌مراتبی معرفی شده در بالا به‌صورت زیر است:

$$L(\boldsymbol{\theta};\mathbf{y}_{(n)}) = \int h(\mathbf{y}_{(n)}|\mathbf{x},\boldsymbol{\theta}_1)g(\mathbf{x},\boldsymbol{\theta}_2)d\mathbf{x} \qquad (۱.۳)$$

که در آن $\mathbf{y}$ بردار مشاهدات است (داده‌های مشاهده شده). برآورد ماکسیمم درستنمایی پارامترها که با $(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2) = (\hat{\theta}_{11},...,\hat{\theta}_{1q};\hat{\theta}_{21},...,\hat{\theta}_{2p})$ نشان داده می‌شود، مقادیر پارامترهای $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = (\theta_{11},...,\theta_{1q};\theta_{21},...,\theta_{2p})$ هستند که توأما تابع درستنمایی را ماکسیمم می‌کنند. واضح است که محاسبه‌ی تابع درستنمایی و برآوردهای ماکسیمم درستنمایی مشمول کار محاسبه‌ای بسیار سخت و پیچیده با انتگرال‌گیری‌هایی با بعد بالا است. به‌طور کامل رویکرد بیزی از عهده‌ی انتگرال‌هایی با بعد زیاد بر می‌آید. رویکرد بیزی با این فرض که پارامترهای مدل، متغیرهای تصادفی هستند، شروع می‌شود. توزیع توأم آماری این پارامترها را توزیع پیشینی می‌گویند. توزیع پیشینی باور شخص نسبت به مقادیر پارامترها قبل از جمع‌آوری داده‌ها را نشان می‌دهد. سپس توزیع پیشینی را با تابع درستنمایی و با استفاده از قضیه‌ی بیز توأم می‌کنند که در نتیجه‌ی آن را توزیع پسینی می‌نامند. این توزیع باور شخص می‌باشد که پس از جمع‌آوری داده‌ها نسبت به پارامترهای جامعه است.

توجه کنید که در مدل‌بندی مدل سلسه‌مراتبی، متغیرهای $\mathbf{X}$ نامعلوم و تصادفی هستند. توزیع پیشین را اغلب با نماد $\pi(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ نشان می‌دهند. بنابر قضیه‌ی بیز، توزیع پسین توأم کمیت‌های نامعلوم $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\mathbf{X})$ به شرط مقادیر مشاهده شده‌ی $\mathbf{y}_{(n)} = (y_1,...,y_n)$ به‌صورت زیر ارائه می‌دهد:

$$h(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\mathbf{X}|\mathbf{y}_{(n)}) = \frac{f(\mathbf{y}_{(n)}|\mathbf{x},\boldsymbol{\theta}_1)g(\mathbf{x}|\boldsymbol{\theta}_2)\pi(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)}{\int f(\mathbf{y}_{(n)}|\mathbf{x},\boldsymbol{\theta}_1)g(\mathbf{x}|\boldsymbol{\theta}_2)\pi(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)d\mathbf{X}d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2} \qquad (۲.۳)$$

توزیع پسینی حاشیه‌ای پارامترها که با $\pi(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\mathbf{X}|\mathbf{y}_{(n)})$ نشان داده می‌شود به سادگی با انتگرال‌گیری از تابع (۲.۳) نسبت به $\mathbf{X}$ به‌دست می‌آید. در ابتدا به نظر می‌رسد که مدل (۲.۳) مشکل انتگرال‌گیری‌های با بعد زیاد را برای محاسبات درستنمایی به یک شکل بزرگتر انتگرال‌گیری‌های با بعد بیشتر در مخرج معادله‌ی (۲.۳) و سپس انتگرال‌گیری از تابع توزیع پسینی حاشیه‌ای را جایگزین می‌کند. اما همان‌گونه که در زیر توضیح می‌دهیم این‌گونه نیست. الگوریتم زنجیر مارکوف مونت کارلویی ابزار محاسباتی است که اعداد تصادفی را از توزیع‌پسینی (۲.۳) تنها با استفاده از صورت کسر بدون این‌که نیازی به انتگرال‌گیری مخرج وجود داشته باشد را تولید می‌کند. برای اطلاعات بیشتر در این‌باره به کسلا و جرج (۱۹۹۲)، چیب و گرینبرگ (۱۹۹۵)، گیلکس و همکاران (۱۹۹۵) و رابرت و کسلا (۲۰۱۰) مراجعه کنید. توجه کنید که صورت کسر نیاز به انتگرال‌گیری ندارد. فرض کنید که اعداد تصادفی تولید شده به‌وسیله‌ی الگوریتم مونت کارلویی زنجیره‌ی مارکوف به‌صورت $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\mathbf{X})_j$ به‌ازای $j = 1,...,B$

نشان داده شود. در اینجا B تعداد مشاهدات استخراج شده از تابع توزیع (۲.۳) است. این مقدار به اندازه‌ی کافی بزرگ است تا بتواند برآورد مناسبی از (۲.۳) را به‌دست آورد. همان طوری که می‌دانید با به‌دست آوردن توزیع (۲.۳) محاسبه‌ی توزیع پسینی $\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}_{(n)})$ نیاز به انتگرال‌گیری نسبت به $\mathbf{X}$ در فرمول (۲.۳) دارد. بنابراین چنین انتگرال‌گیری‌هایی لازم نمی‌باشد. توزیع حاشیه‌ای $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ به‌راحتی با حذف مولفه‌ی $\mathbf{X}$ از اعداد تصادفی $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{X})_j$ به‌صورت $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)_j$ به‌ازای $j = 1, ..., B$ به‌دست می‌آید. به‌طور مشابه، مقادیر میانگین و واریانس اعداد تصادفی ایجاد شده‌ی $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)_j$ به‌ازای $j = 1, ..., B$ است. در نتیجه، فرآیند شبیه‌سازی توزیع پسینی حاشیه‌ای نیازی به هیچ انتگرال‌گیری‌ای ندارد.

۴.۳ مدل سلسله‌مراتبی بیزی فضایی

فرض کنید یک مورد از داده‌ی سطح فردی بر روی متغیر هدف و متغیرهای کمکی نمونه بررسی شده در همه‌ی نواحی وجود داشته باشد. توجه کنید، مدل‌ها با داده جمع‌آوری شده، در نظر گرفته می‌شود. برای متغیرهای گسسته، پاسخ می‌تواند با استفاده از توزیع پواسون مدل‌سازی شود (رائو، ۲۰۰۳):

$$\mathbf{y}_{ij}|\boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_e^2 \sim P(\boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_e^2) \qquad (۳.۳)$$

بنابراین $\boldsymbol{\mu}_{ij}$ مقدار واقعی (درست) متغیر هدف از فرد jام در نمونه‌ای از ناحیه‌ی i ام است، $\boldsymbol{\sigma}_e^2$ نشان‌دهنده‌ی تغییرات نمونه‌گیری فردی می‌باشد و فرض می‌شود در همه نواحی شدنی است $(\boldsymbol{\sigma}_i^2 = \boldsymbol{\sigma}_e^2)$. اگرچه به‌طور معمول متغیرهای کمکی، به‌صورت وابستگی مدل بین میانگین و تعداد عوامل توضیحی، تعریف شده است. و این احتمال هم وجود دارد، برخی واریانس باقی‌مانده توسط متغیرهای کمکی غیرقابل توضیح باقی بماند. به‌طور حتم می‌توان با توجه به اثرات تصادفی در مدل مورد توجه قرار بگیرد. این اثرات تصادفی، الگوهای ناشناخته‌ای را همانند وابستگی فضایی و تغییرات بین نواحی در نظر می‌گیرند. بنابراین، مدل رگرسیون اثرات تصادفی زیر را برای میانگین‌های سطح واحد در نظر می‌گیرد:

$$\boldsymbol{\mu}_{ij} = \boldsymbol{\alpha} + \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{v}_i \qquad (۴.۳)$$

که منجر به مدل میانگین سطح ناحیه می‌شود:

$$\boldsymbol{\mu}_i = \sum_j \frac{\boldsymbol{\mu}_{ij}}{N_i} = \boldsymbol{\alpha} + \bar{\mathbf{X}}_i\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{v}_i \qquad (۵.۳)$$

$\boldsymbol{\alpha}$ عرض از مبدا مدل و $\boldsymbol{\beta}$ برداری از ضرایب متغیرهای کمکی $\mathbf{X}_{ij}$، $\mathbf{u_i}$ اثرات تصادفی که برای تغییرات سطح ناحیه و توزیع مستقل شده زیر به‌شمار می‌رود:

$$\mathbf{u}_i|\boldsymbol{\sigma}_u^2 \sim P(0, \boldsymbol{\sigma}_u^2)$$

لذا $\mathbf{u}_i$ نشان‌دهنده‌ی همبستگی اثرات تصادفی فضایی است.

۱.۴.۳ داده‌های مشبکه‌ای

یکی از معمول‌ترین مدلی که در برآورد نواحی کوچک استفاده می‌شود، مدل اتورگرسیوشرطی CAR است. این مدل در سال (۱۹۹۱) توسط بی‌سگ و همکارانش معرفی شده است. این مدل به‌دلیل دارا بودن ساختار ماتریس مجاورت بسیار پرکاربرد خواهد بود. از طرفی این مدل‌ها توسط توزیع‌های احتمال شرطی ساخته می‌شوند. همچنین بسته به موقعیت قرار گرفتن داده‌ها در مدل‌بندی توزیع احتمال نقش مهمی را ایفا می‌کنند. پس ملزوم می‌باشد که نحوه‌ی تاثیر قرار گرفتن این نوع از موقعیت‌ها در مدل‌بندی مورد بررسی قرار گیرد. بیشتر داده‌های نواحی کوچک فضایی مشبکه‌ای هستند. از طرفی یکی از پرکاربردترین مدل‌های فضایی برای داده‌های مشبکه‌ای و وارد کردن وابستگی فضایی از طریق ساختار همسایگی نواحی، مدل اتورگرسیوشرطی است.

به‌دلیل به‌کارگیری این مدل در زمینه‌های مختلف از جمله روش MCMC، که برای برازش رده مشخصی از مدل‌های سلسله‌مراتبی به‌کار می‌رود، اهمیت می‌یابد. از این‌رو، در مدل‌های فضایی، فرضیه‌ی مستقل بودن برداشته می‌شود و اثرات تصادفی ناحیه $\mathbf{v}_i$ به‌صورت وابسته و تحت تاثیر موقعیت جغرافیایی و همسایگی نواحی قرار می‌گیرد. توزیع‌ها و مدل‌های مختلف فضایی اغلب، از مدل اتورگرسیوشرطی استفاده می‌کنند. تحت این مشخص‌سازی، توزیع شرطی از $\mathbf{u}_i$ به شرط مقادیر $\mathbf{v}_i$ در تمام نواحی باقی مانده تنها مستلزم همسایگی نواحی می‌باشد.

$$\mathbf{v}_i|\mathbf{v}_{-i}, \boldsymbol{\sigma}_v^2 \sim \mathrm{N}(\sum_{j\epsilon\delta_i} \frac{d_{ij}}{|\boldsymbol{\delta}_i|}\mathbf{v}_j, \frac{\boldsymbol{\sigma}_v^2}{|\boldsymbol{\delta}_i|}) \qquad (۶.۳)$$

$\boldsymbol{\delta}_i$ مجموعه‌ای از همسایگی‌های ناحیه‌ی iام و $|\boldsymbol{\delta}_i|$ تعداد همسایگی‌ها درنظر گرفته می‌شود. علاوه براین، قیدی را اضافه می‌کنند تا مجموع مقادیر همه‌ی اثرات تصادفی $\mathbf{v}_i$ برابر صفر شود تا عرض از مبدا و اثرات تصادفی را شناسایی کند (بانرجی و همکاران، ۲۰۰۴).

۲.۴.۳ داده‌های زمین آماری

در داده‌های زمین آماری، فرض می‌شود که $\mathbf{w}_i = \mathbf{w}(s_i)$ تحققی از یک میدان تصادفی گاوسی[۵] (GRF) به‌صورت $\{\mathbf{w}(s), s\varepsilon S\}$ است، و $\mathbf{w} = (w_1, ..., w_n)$ برداری از یک توزیع نرمال چندمتغیره[۶] $\mathbf{w} \sim MVN(0, \Sigma)$ است. بنابراین، به‌عنوان جایگزین برای مشخصات شرطی می‌توانید میانگین $\boldsymbol{\mu}_{ij}$ که شامل اثرات تصادفی فضایی $\mathbf{w}_i$ و همچنین همبسته است را انتخاب کنید، و از طرفی، مطابق فاصله‌ی d_{kl} که بین دو ناحیه‌ی k و l قرار می‌گیرد، مطابق زیر عمل کنید (دیاگ و همکاران، ۱۹۹۸):

$$\boldsymbol{\mu_{ij}} = \boldsymbol{\alpha} + \mathbf{x_{ij}}\boldsymbol{\beta} + \mathbf{w}_i \qquad (۷.۳)$$

[۵] Gaussion random field

[۶] Multivariate normality

$\mathbf{w}$ به‌عنوان نرمال چند متغیره توزیع شده است.

$$\mathbf{w}_i|\boldsymbol{\Sigma} \sim MVN(0, \boldsymbol{\Sigma}); \boldsymbol{\Sigma}_{kl} = \boldsymbol{\sigma}_w^2 exp\{-(\boldsymbol{\phi} d_{kl})\} \qquad (۸.۳)$$

$\boldsymbol{\sigma}_w^2$ واریانس در هر نقطه و ϕ پارامتر هموارسازی که مقیاس بین نواحی را کنترل می‌کند. برخلاف مدل (۵.۳) اثرات تصادفی مستقل $\mathbf{u}_i$ را در مدل (۷.۳) وارد نمی‌کنند. هدف برای انجام مدل (۵.۳) مبنی بر این است که وابستگی فضایی اتورگرسیو شرطی ذاتی ، اثرات تصادفی مدل (۶.۳) را به‌وسیله‌ی ساختار همسایگی که از پیش تعیین شده است، تشریح کند. از این رو، اثرات غیرساختاری، همیشه این اجازه را برای یادگیری بیزی درباره‌ی وابستگی فضایی قوی در داده از طریق همبستگی نسبی $\mathbf{u}_i$ و $\mathbf{w}_i$ پسینی خواهد داد (بی‌سگ و همکاران، ۱۹۹۱ و کارلین و همکاران، ۲۰۰۰). درباره‌ی مدل (۷.۳) یادگیری بیزی درباره‌ی وابستگی فضایی قوی اثرات تصادفی $\mathbf{w}_i$ به‌طور مستقیم از طریق برآوردگر پسین پارامتر همبستگی ϕ در مدل (۸.۳) انجام می‌گیرد (مستلزم هیچ همبستگی فضایی نیست, $\phi \longrightarrow 0$). در این روش ممکن است اثر تصادفی مستقل جداگانه در مدل (۷.۳) باشد. اما در عمل، می‌تواند منجر به ضعیف شدن توزیع‌های پسین شود (دیاگ و همکاران، ۲۰۰۲). بنابراین برای همه مدل‌ها برآورد سطح ناحیه معقول می‌باشد.

$$\hat{\boldsymbol{\mu}}_{b,i} = E_{\circ|y}[\boldsymbol{\alpha} + \bar{\mathbf{X}}_i\boldsymbol{\beta} + \mathbf{z}_i] = \hat{\boldsymbol{\alpha}} + \bar{\mathbf{X}}_i\hat{\boldsymbol{\beta}} + \hat{\mathbf{z}}_i$$

$E_{.|y}[.]$ نشان‌دهنده‌ی امیدریاضی پسین و $\mathbf{z}_i$ نشان‌دهنده‌ی اثرات تصادفی است که هر دو با $\mathbf{z}_i = \mathbf{u}_i + \mathbf{v}_i$ یا $\mathbf{z}_i = \mathbf{w}_i$ مشخص شده‌اند. در این مورد به‌ترتیب میانگین‌های پسین $\hat{\boldsymbol{\alpha}}$، $\hat{\boldsymbol{\beta}}$ و $\hat{\mathbf{z}}_i$ و $\boldsymbol{\alpha}$، $\boldsymbol{\beta}$ و $\mathbf{z}_i$ محاسبه می‌شوند. از طرفی هم، فرض کنید که میانگین‌های سطح ناحیه متغیرهای کمکی $\bar{\mathbf{X}}_i$ در دسترس‌اند. اساسا مدل (۸.۳) توسط بست و همکاران در سال (۱۹۸۸) مطرح شده است که شامل انواع مختلفی از اثرات تصادفی می‌باشد. واریانس $\boldsymbol{\sigma}_e^2$ ناحیه‌ی داخلی را برای همه‌ی نواحی درنظر می‌گیرد که معمولا غیر واقعی است زیرا، تغییرات فردی بین نواحی متفاوت است. به‌طور کلی، گسترش این مدل در هر ناحیه، واریانس $\boldsymbol{\sigma}_i^2$ متفاوتی را در نظر می‌گیرد.

$$\mathbf{y}_{ij}|\boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_i^2 \sim P(\boldsymbol{\mu_{ij}}, \boldsymbol{\sigma_i^2}) \qquad (۹.۳)$$

$$\boldsymbol{\sigma}_i^2 \sim \pi(\boldsymbol{\sigma}_i^2),$$

$$Log(\boldsymbol{\mu}_{ij}) = \boldsymbol{\alpha} + \mathbf{x_{ij}}\boldsymbol{\beta} + \mathbf{w}_i$$

$\pi(\boldsymbol{\sigma}_i^2)$ یک توزیع پیشین مبهم است. در این مورد، هر واریانس تنها با استفاده از اطلاعات نمونه در ناحیه‌ی iام برآورد می‌شود. هنگامی که داده بررسی در داخل هر ناحیه به‌طور جداگانه باشد، می‌تواند منجر به برآوردگرهای ضعیف‌تری از $\boldsymbol{\sigma}_i^2$ شود. یک جایگزینی می‌تواند این باشد که، با به‌کارگیری از ساختار سلسله‌مراتبی بر روی واریانس‌ها و وام گرفتن از اطلاعات

نواحی بتوان برآوردگرهای استوار را به‌دست آورد. به‌طور ویژه، می‌توانید لگاریتم واریانس‌ها را به‌صورت زیر مدل‌سازی کنید:

$$\mathbf{y}_{ij}|\boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_i^2 \sim P(\boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_i^2) \qquad (۱۰.۳)$$

$$Log(\boldsymbol{\sigma}_i^2)|\boldsymbol{\sigma}^2 \sim P(0, \boldsymbol{\sigma}^2)$$

$$\boldsymbol{\sigma}_i^2 \sim \pi(\boldsymbol{\sigma}^2)$$

مدل آخر در استفاده از توابع واریانس تعمیم‌یافته در هموارسازی واریانس‌های سطح ناحیه مشابه است، و به‌طور کامل مدل‌پایه می‌باشد. بنابراین، عدم اطمینان درباره‌ی برآوردگرهای واریانس منجر به واریانس پسین در برآوردگرهای کوچک ناحیه‌ای می‌شود. مدل (۹.۳) اساسا همان مدلی است که توسط آرورا و لاهیری در سال (۱۹۹۷) با اثرات تصادفی پیشنهاد شده است. آرورا و لاهیری (۱۹۹۷) با به‌کارگیری از مدل سطح ناحیه مشابه با اثرات تصادفی مستقل و با هدف رویکرد بیز تجربی برای برآورد اثرات تصادفی پیشنهاداتی را مطرح کرده‌اند. رائو و همکاران (۱۹۹۲) تقریبی از MSE را در مدل سطح واحد با واریانس‌های سطح ناحیه که متفاوت با توزیع پیشین مشترک است، مطرح کرده‌اند.

۳.۴.۳ توزیع پیشین پارامترهای مدل

برای عرض از مبدا α و ضرایب متغیرهای کمکی β پیشین‌های تخت ناسره به‌کار گرفته می‌شود اما، پسین‌های سره به‌وجود آمده است. گامای وارونه را همانند توزیع پسین برای هر یک از واریانس‌های $\sigma_i^2, \sigma_v^2, \sigma_u^2, \sigma_w^2$ استفاده می‌کنند. به‌منظور اطلاعات پیشین مبهم و یادگیری مدل از داده، مقادیر کوچک را برای پارامترهای توزیع گاما استفاده می‌کنند. پیشین برای پارامتر ϕ وابسته به دامنه و مقیاس اندازه‌گیری فاصله‌های بین نواحی کوچک است. گلمن (۲۰۰۶) نشان داد که پیشین‌های معکوس گاما با پارامترهای مقیاس و شکل برای واریانس‌های اثرات تصادفی ممکن است، انقباض دروغین ایجاد کند. به‌خصوص زمانی که تعداد گروه‌ها کوچک باشد، تعداد اندکی مشاهدات در هر گروه وجود دارد، و از این رو، چندین جایگزین پیشنهاد شده است. همیشه از یک توزیع نیمه‌کوشی بر روی انحراف استاندارد از اثرات تصادفی استفاده می‌کنند. اما در برآوردگرهای کوچک ناحیه‌ای اختلاف معنی‌داری را نخواهد داشت.

مثال ۱.۴.۳. برای داده سطح ناحیه می‌توان مدل را در معادله (۸.۳) نشان داد. پس، به‌دلیل داشتن اثرات تصادفی و متغیرهای کمکی می‌توان آن را گسترش داد. برای مثال:

$$\hat{\bar{\mathbf{Y}}}_i|\boldsymbol{\mu}_i, \hat{\mathbf{V}}_i^2 \sim P(\boldsymbol{\mu}_i, \hat{\mathbf{V}}_i^2) \qquad (۱۱.۳)$$

$$\boldsymbol{\mu}_i, = \boldsymbol{\alpha}^* + \bar{\mathbf{X}}_i\boldsymbol{\beta}^* + \mathbf{u}_i^* + \mathbf{v}_i^*$$

$\mathbf{u}_i^*$ و $\mathbf{v}_i^*$ به‌ترتیب نشان‌دهنده‌ی پواسون و توزیع‌های CAR می‌باشند. واریانس‌های سطح ناحیه در برآوردگرهای میانگین ناحیه به‌طور فرض شناخته شده‌اند، لذا با یک مربع نشان داده می‌شوند. برآورد از میانگین ناحیه به‌صورت زیر ارائه می‌شود:

$$\hat{\boldsymbol{\mu}}_{B,i} = E_{\circ|y}[\boldsymbol{\alpha}^* + \bar{\mathbf{X}}_i\boldsymbol{\beta}^* + \mathbf{u}_i^* + \mathbf{v}_i^*] = \hat{\boldsymbol{\alpha}}^* + \bar{\mathbf{X}}_i\hat{\boldsymbol{\beta}}^* + \hat{\mathbf{u}}_i^* + \hat{\mathbf{v}}_i^*$$

همانند قبل نماد ˆ نشان‌دهنده‌ی میانگین پسین از پارامتر مربوط است. برای مدل فوق $\hat{\boldsymbol{\alpha}}^*$ و $\hat{\boldsymbol{\beta}}^*$ به‌جای $\boldsymbol{\alpha}$ و $\boldsymbol{\beta}$ به‌منظور نشان دادن این عمکلرد بیان شده است تا مدل‌های سطح ناحیه، برآوردگرهای متفاوتی را در مدل‌های سطح واحد ارائه دهند. به‌طور مشابه، برآوردگرهای اثرات تصادفی $\mathbf{u}_i^*$ و $\mathbf{v}_i^*$ مشابه‌اند ولی متفاوت از $\mathbf{u}_i$ و $\mathbf{v}_i$ خواهند بود.

۵.۳ ارزیابی کیفیت برآوردگرها

ارزیابی کیفیت برآوردگرهای کوچک ناحیه‌ای که با مدل‌های سطح ناحیه و واحد به‌دست آمده است، می‌تواند در عمل متفاوت باشد. معمولا به دست آوردن واریانس‌ها و میانگین توان دوم خطاهای پیش‌بینی[۷] مهم می‌باشد اما، به‌دست آوردن برآوردگرهای خوب از (MSPE) معمولا متفاوت بر روش‌های SAE است. از این‌رو، دارای فرم بسته‌ای می‌باشند که منجر به برآورد پارامترهای مدل نمی‌شوند. فرمول تقریبی مختلفی پیشنهاد شده است که عبارت‌اند از: برآوردگرهای بوت استرپ، جک نایف و ... (جیانگ و لاهیری، ۲۰۰۶ و رائو، ۲۰۰۳).

از سوی دیگر، اندازه‌گیری بیز طبیعی دقیق و واریانس پسین برآوردگرهای کوچک ناحیه‌ای به‌طور خودکار از پسین به‌دست آمده است و به‌طور کامل عدم اطمینان برای همه‌ی پارامترهای مدل محاسبه می‌شود. یک معیار که می‌تواند برای مقایسه مدل در آمار بیزی به‌کار رود، معیار انحراف اطلاعات[۸] (DIC) است (اسپیگل هالتر و همکاران، ۲۰۰۲). بر این اساس، انحراف از مدل تاوانیده برای پیچیدگی مدل و تفسیر آن مشابه[۹] (AIC) است. پس، مدل‌های ترجیح داده می‌شود که میزان DIC کم‌تری داشته باشند.

برای مطالعات شبیه‌سازی زمانی که مقدار میانگین واقعی (درست) $\bar{Y}_i$ شناخته شده باشد، می‌توانید اریبی نسبی[۱۰] (RB) و ریشه‌ی میانگین توان دوم خطای نسبی[۱۱] RRMSE را برای ارزیابی و درستی برآوردگرهای کوچک ناحیه‌ای محاسبه کنید. توجه کنید RRMSE یعنی، $RRMSE = \sqrt{MSE}/$(مقدار واقعی).

[۷] Mean squared error prediction
[۸] Information deviation criterion
[۹] Akaike information criterion
[۱۰] Relative bias
[۱۱] Relative root mean squared error

لذا همانند زیر تعریف شده است:

$$RB_i = \frac{۱}{\mathbf{K}} \frac{\sum_{k=۱}^{K} (\hat{\bar{\mathbf{Y}}}_i^{(k)} - \bar{\mathbf{Y}}_i)}{\bar{\mathbf{Y}}_i},$$

$$RRMSE_i = \frac{\sum_{k=۱}^{K} (\hat{\bar{\mathbf{Y}}}_i^{(k)} - \bar{\mathbf{Y}}_i)^۲}{\bar{\mathbf{Y}}_i}$$

$\mathbf{k}$ شاخص بررسی نمونه‌ها است.

همانند اندازه‌گیری فراموضعی، می‌توان میانگین مطلق اریبی نسبی و میانگین ریشه‌ی میانگین مربعات توان دوم خطای نسبی[۱۲] (MRRMSE) را در نظر گرفت:

$$MARB = \frac{۱}{m} \sum_{i=۱}^{m} |RB_i|, \qquad (۱۲.۳)$$

$$MRRMSE = \frac{۱}{m} \sum_{i=۱}^{m} RRMSE_i$$

بنابراین m آخرین شماره از نواحی است. در این روش، به ترتیب می‌توانید ارزیابی کنید که آیا برآوردگرها اریب هستند یا خیر. هریک از این دو معیار را می‌توانید برای تصمیم‌گیری بر روی بهترین در مدل استفاده کنید. پس، برآوردگری بهتر خواهد بود که مقادیر کوچکتری از میانگین قدرمطلق اریبی نسبی[۱۳] MARB و میانگین نسبی ریشه میانگین توان دوم خطا MRRMSE را ارائه دهد.

همچنین، اریبی نسبی RBvar و ریشه نسبی میانگین توان دوم خطا RRMSEvar برآوردگرهای واریانس را در اطلاعات شبیه‌سازی محاسبه می‌کنند. علاوه بر این، به بررسی خطای واقعی برآوردگر کوچک ناحیه‌ای که چه میزانی است هم می‌پردازد. لذا همانند زیر تعریف شده است:

$$RBvar_i = \frac{۱}{K} \frac{\sum_{k=۱}^{K} (\hat{var}(\bar{Y}_i^{(k)}) - EMSE_i)}{EMSE_i}$$

$$RRMSEvar_i = \frac{۱}{K} \frac{\sum_{k=۱}^{K} (\hat{var}(\bar{Y}_i^{(k)}) - EMSE_i)^۲}{EMSE_i}$$

و $(EMSE_i)$ خطای تجربی واقعی است. لذا خواهیم داشت:

$$EMSE_i = \frac{۱}{K} \sum_{k=۱}^{K} (\hat{var}(\bar{Y}_i^{(k)}) - \bar{Y}_i)^۲$$

میانگین مطلق اریبی نسبی MARBvar و میانگین نسبی ریشه میانگین توان دوم خطا MRRMSEvar برآوردگرهای واریانس نیز همانند مدل (۱۲.۳) تعریف می‌شود.

[۱۲] Mean relative root mean squared error

[۱۳] Mean absolute relative bias

۶.۳ برآورد کوچک ناحیه‌ای در صورت عدم اطلاع از برآوردگرهای مستقیم

به‌منظور کاهش هزینه‌ها، اغلب بررسی‌ها در یک زیر مجموعه از نواحی با نمونه‌گیری از جمعیت انجام می‌شود که نماینده کل نواحی مورد مطالعه است. این بدان معنی است که برآوردگرهای مستقیم می‌تواند تنها برای تعداد کمی نواحی ارائه شود و برآوردگرهایی که برای آن‌ها نمونه در نواحی وجود ندارد باید از روش‌های دیگری به‌دست آورده شوند. بنابراین می‌توانید نواحی را به دو قسمت تقسیم کنید : ۱. نواحی داخل نمونه ۲. نواحی خارج نمونه. در مدل‌های سطح ناحیه، مقادیر از دست رفته $\hat{\bar{\mathbf{Y}}}_i$ و $\hat{\sigma}_i^2$ برای نواحی خارج نمونه وجود خواهد داشت ضمن اینکه، در مدل‌های سطح واحد مقادیر از دست رفته $\mathbf{y}_{ij}$ و $\mathbf{x}_{ij}$ در نواحی خارج نمونه وجود دارد. از طرفی فرض می‌شود، متغیرهای کمکی سطح ناحیه $\bar{\mathbf{x}}_i$ برای همه‌ی نواحی در دسترس‌اند و لذا، در بررسی‌ها از منابع متفاوتی به‌دست آورده می‌شوند. این یک فرض کلیدی است تا بتوانید برآوردهای نواحی کوچک را برای همه نواحی ارائه کنید. توجه کنید در اینجا فقط به مشکلات داده‌ای که در طرح بررسی از دست رفته است، پرداخته می‌شود.

برای مثال، عدم پاسخ در بررسی‌ها، یکی دیگر از منابع مشترک داده از دست رفته در برآوردهای کوچک ناحیه‌ای است اما در اینجا این مسئله برطرف نمی‌شود. یک رویکرد ساده در برخورد با مشکل عدم مشاهدات مستقیم در نواحی مختلف، به‌کارگیری از یک مدل رگرسیونی بر روی برخی متغیرهای کمکی است که از داده بررسی استفاده می‌شود. برآوردها برای نواحی خارج نمونه با تکیه بر مدل برآورده شده‌اند و اطلاعات اضافی محاسبه می‌شوند (همانند، متغیرهای کمکی سطح ناحیه). مشکل اصلی این روش این است که مقادیر جانهی شده برای عدم اطمینان در برآورد ضرایب رگرسیونی یا همبستگی فضایی بین متغیر هدف در نواحی مختلف محاسبه نمی‌شوند.

۱.۶.۳ همبستگی اثرات تصادفی فضایی

اگر هر یک از مدل‌های بخش (۴.۳) و مدل‌های مثال (۱.۴.۳) را در نظر بگیرید، اثرات تصادفی همبستگی فضایی می‌تواند علاوه بر متغیرهای کمکی هنگام پیش‌گویی برآوردهای کوچک ناحیه‌ای در نواحی خارج نمونه مورد محاسبه قرار بگیرد. این رویکرد هنگامی اهمیت پیدا می‌کند که، وام گرفتن اطلاعات نواحی همسایه‌ها به کمک برآوردگرهای مستقیم از دست‌رفته باشد.

روش اطلاعات برای نواحی که با مشاهدات غیرمستقیم از سایر نواحی وام گرفته است، به شرح زیر می‌باشد. همچنین اگر می‌خواهید برآورد در نواحی خارج نمونه را داشته باشید از مدل‌های سطح ناحیه که در معادله‌ی (۱۱.۳) مطرح شده است، استفاده کنید، بنابراین

می‌تواند همانند مدل زیر بیان شود:

$$\begin{bmatrix}\hat{\bar{\mathbf{Y}}}_s\\ \boldsymbol{\mu}_s\end{bmatrix} = \boldsymbol{\alpha} + \begin{bmatrix}\bar{\mathbf{X}}_s\\ \bar{\mathbf{X}}_{\underline{s}}\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\mathbf{z}_s\\ \mathbf{z}_{\underline{s}}\end{bmatrix} + \begin{bmatrix}\mathbf{e}_s\\ 0\end{bmatrix} \qquad (۱۳.۳)$$

$\mathbf{z}$ نشان‌دهنده‌ی همبستگی اثرات تصادفی فضایی و زیرنویس s نشان‌دهنده‌ی مشاهدات (داخل نمونه) نواحی و $\underline{s}$ نواحی مشاهده نشده است. مقدار $\mathbf{z}_{\underline{s}}$ را می‌توان توسط همبستگی (فضایی) با $\mathbf{z}_s$ برآورد کرد.

۲.۶.۳ مشخص‌سازی نرمال چندمتغیره

هنگامی که بردار کاملی از اثرات تصادفی فضایی $\mathbf{z} = \mathbf{w}$ دارای توزیعی باشد، آن‌گاه همانند مدل (۸.۳) نشان داده می‌شود. توزیع شرطی از $\mathbf{w}_{\underline{s}}|\mathbf{w}_s$ را طبق رابطه‌ی زیر نشان می‌دهند (دیاگ و همکاران، ۱۹۹۸):

$$MVN(\boldsymbol{\Sigma}_{\underline{s}s}\boldsymbol{\Sigma}_{ss}^{-1}\mathbf{w}_s, \boldsymbol{\Sigma}_{\underline{s}\underline{s}} - \boldsymbol{\Sigma}_{\underline{s}s}^T\boldsymbol{\Sigma}_{ss}\boldsymbol{\Sigma}_{\underline{s}s})$$

برآورد از $\hat{\mathbf{w}}_{\underline{s}}$ هنگامی است که امیدریاضی پسین میانگینی از MVN شرطی باشد، $E_{.|y}[\boldsymbol{\Sigma}_{\underline{s}s}\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{w}_s]$. پس، برآوردگر نهایی برای مجموعه‌ای از نواحی در $\underline{s}$ همانند مدل زیر است:

$$\hat{\bar{\mathbf{Y}}}_{\underline{s}} = \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\beta}}\bar{\mathbf{X}}_{\underline{s}} + \hat{\mathbf{w}}_{\underline{s}}$$

این برآوردگر می‌تواند به‌عنوان نسخه‌ی بهبودیافته‌ی برآوردگر ترکیبی و احتمال اینکه اریبی در این برآوردگر کاهش یابد، درنظر گرفته شود.

۳.۶.۳ مشخص‌سازی اتورگرسیو شرطی

پیش‌گویی اثرات تصادفی، $\mathbf{w}_{\underline{s}}$، در نواحی خارج نمونه با استفاده از مشخص‌سازی مدل نرمال چندمتغیره (۸.۳) غیرمبهم است. می‌توان به این علت بیان کرد که توزیع پیوند (ربط) از $\mathbf{w} = (\mathbf{w}_s, \mathbf{w}_{\underline{s}})$ و پیش‌گویی از $\mathbf{w}_{\underline{s}}|\mathbf{w}_s$ منحصر به‌فرد تعیین شده است. در مقابل، پیش‌گویی اثرات تصادفی CAR $\mathbf{v}_{\underline{s}}$ است، بنابراین توزیع پیوند برای بردار کامل $\mathbf{v} = (\mathbf{v}_s, \mathbf{v}_{\underline{s}})$ تعریف نشده است (بانرجی و همکاران، ۲۰۰۴). در عوض، پیش‌گویی فرآیندهای که به‌طور مستقیم با توزیع‌های شرطی $\mathbf{v}_{\underline{s}}|\mathbf{v}_s$ انجام شده است، به خوبی نیز تعریف می‌شود، اما با این تفاوت که در عمل منحصر به‌فرد نمی‌باشد. این نکته توسط بانرجی و همکاران (۲۰۰۴) مطرح شد. برای این مورد یک مدل اتورگرسیو شرطی بر روی نقطه‌ای از سطح داده برازش داده شده است.

پیش‌گویی در یک مکان جدید می‌تواند در ساختار یک مدل CAR برای مجموعه‌ای از مکان‌های مشاهده شده انجام شود. و به‌طور جداگانه، توزیع شرطی تعیین شده از مکان جدید با توجه به مشاهدات یا با ایجاد یک مدل CAR برای مجموعه کاملی از مکان‌های مشاهده شده

جدید نیز انجام گیرد. هر دو روش معتبر است، اما پیش‌گویی توزیع‌های مختلف مقدم‌تر است. در مورد داده سطح ناحیه، منطقی نیست که رویکرد قبلی را در نظر گرفت، بنابراین مشخص نیست که چطور ساختار مجاورت فقط نواحی خارج نمونه را در نواحی چشم‌پوشی می‌کند. بنابراین مدل CAR (۶.۳) برای مجموعه کاملی از اثرات تصادفی فضایی که در نواحی داخل نمونه و در خارج نمونه که $\mathbf{v} = (v_s; v_{\underline{s}})$ است، را مشخص می‌کنند. و به سادگی پاسخ داده که در نواحی خارج نمونه ازدست رفته باشد را مورد بررسی قرار می‌دهد. این عمل منجر به مجموعه تغییریافته از توزیع‌های شرطی کامل برای اثرات تصادفی فضایی در نواحی خارج نمونه در طرح MCMC و نیز با به‌کارگیری از برآوردگر توزیع پسین انجام می‌شود.

اگرچه مشخص‌سازی برای پیش‌گویی اثرات تصادفی خارج نمونه مشخص است ولی زمانی که مشاهدات مستقیم از چندین نواحی باشند، اغلب با مشکلات زیر مواجه می‌شوید:

۱. برازش v_i در نواحی داخل نمونه ممکن است مشکل‌ساز باشد، اگر بیشتر یا همه‌ی همسایگی نواحی خارج نمونه باشد.

۲. به‌طور مشابه هنگامی که v_i در نواحی خارج نمونه پیش‌گویی می‌شود، ممکن است نواحی با مقدار کم یا همسایگی‌های داخل نمونه وجود نداشته باشند.

در اصل مدل می‌تواند برازش داده شود ولی، هنگامی اتفاق می‌افتد که در آن‌جا نواحی با همسایگی از دست رفته باشد، اما اگر تعدادی از آن‌ها از دست رفته باشد شیوه‌ی برآورد خیلی پایدار نخواهد بود. اثرات تصادفی غیرفضایی $\mathbf{u}$ هم می‌تواند در مدل اثر گذارد، اما باید تمامی اعضای $u_{\underline{s}}$ به صفر برسد تا از مشکلات شناسایی‌پذیری با $v_{\underline{s}}$ دوری کند. به هر حال، اختلاف (تفاوت) اندازه‌گیری اثرات تصادفی بین پیش‌گویی اثرات داده و پاسخ در مدل غیرساختاری است. اگر پاسخ نقض شود، اثرات تصادفی $\mathbf{u}$ نمی‌تواند همبستگی بین نواحی را داشته باشد. پس برآورد مقیاس مقادیر در $u_{\underline{s}}$ غیرممکن می‌شود.

۴.۶.۳ وام گرفتن اطلاعات در سطوح اجرایی بالاتر

همان‌طوری که قبلا اشاره شد، هنگامی که تعدادی داده در داخل نمونه نباشد، ممکن است پیش‌گویی اثرات تصادفی فضایی در سطح بالا با استفاده از یک مدل CAR مشکل‌ساز باشد زیرا، اطلاعات از همسایگی نواحی وام گرفته است و به اندازه‌ی کافی این احتمال وجود دارد که همسایگی داخل نمونه در دسترس برآورد طرح فضایی نباشد. علاوه بر این مشکل، مشخص‌سازی مدل CAR در سطح جغرافیایی بالاتر می‌تواند با تنک بودن در داده استفاده شود. فرض می‌شود، همه‌ی نواحی از یک قسمت از M سطوح اجرایی بالاتر و همچنین هر کدام از آن‌ها از چند نواحی ساخته شده باشد. از طرفی، اثر تصادفی $\mathbf{r_k}$، $k = ۱, ..., M$ را به‌جای اثر فضایی v_i اضافه می‌کنند، بنابراین میانگین در مدل‌های سطح ناحیه همانند زیر تجزیه

می‌شود، پس خواهیم داشت:

$$\boldsymbol{\mu}_i = \boldsymbol{\alpha} + \beta\bar{\mathbf{X}}_i + \mathbf{u}_i + \mathbf{r}_{k(i)} \qquad (۱۴.۳)$$

شاخص $k(i)$ نشان‌دهنده‌ی سطوح اجرایی بالاتر در ناحیه‌ی iام است. یک رابطه مشابه را می‌توان برای ساختن $\boldsymbol{\mu}_{ij}$ در مدل سطح ناحیه پیشنهاد کرد. $\mathbf{r}_{k(i)}$ ساختار فضایی با ویژگی تغییرات فضایی بزرگ مقیاس را در بر می‌گیرد. یک پیشین CAR مشابه آنچه که برای v_i، تعیین کرده‌ایم را فرض می‌کنیم، اما مجاورت، مطابق سطح اجرایی بالاتر تعریف می‌شود و فرض می‌شود حداقل یک ناحیه در هر منطقه نمونه‌برداری می‌شود. حال اگر الگوی فضایی در این سطح بالاتر، خیلی ضعیف یا اصلا وجود نداشته باشد اثرات تصادفی $\mathbf{r}_{k(i)}$ ممکن است، یک توزیع غیرفضایی همانند $P(0, \boldsymbol{\sigma}_r^2)$ را تعیین کنند.

جمع‌بندی

به‌طور خلاصه می‌توان گفت در این فصل، بنابر اهمیت مدل‌های بیزی و ساختار سلسه‌مراتبی، می‌توان آن‌ها را در برآورد کوچک ناحیه‌ای لحاظ نمود. با توجه به اینکه، برآورد کوچک ناحیه‌ای نیز، به‌دلیل وابستگی بین نواحی حالت فضایی بودن را در خودش دارد این موضوع کمکی بر آن خواهد داشت که از داده‌های مشبکه‌ای استفاده کنیم. لذا مشخصاتی که در این مدل وجود دارد را به‌طور کامل بیان نمودیم و به کمک مدل اتورگرسیو شرطی توانسته‌ایم در فصل بعدی داده‌های واقعی خود را به کمک مدل سلسه‌مراتبی برازش دهیم. نتایج حاصل از داده‌ها نیز در فصل آتی بیان می‌کنیم.

فصل ۴

کاربرد مدل فضایی بیزی برای برآورد کوچک ناحیه‌ای داده‌های بیمه استان گیلان

۱.۴ مقدمه

هدف از این فصل، تحلیل تعداد بیمه‌شدگان اجباری در شهرستان‌های استان گیلان است. با توجه به ماهیت شمارشی پاسخ در این مجموعه داده و وجود اطلاعات مکانی همسایگی شهرستان‌ها، از رده مدل‌های آمیخته خطی تعمیم‌یافته فضایی استفاده می‌کنیم. آماردان‌های مختلفی از این مدل‌ها در بحث نواحی کوچک استفاده کرده‌اند. به‌عنوان دو نمونه می‌توان به کلینشمیت و همکاران (۲۰۰۰) و ژانگ و لی (۲۰۱۱) اشاره کرد. استنباط مبتنی بر درستنمایی در رده SGLMMs، به‌جز نسخه‌های خطی آن، به دلیل وجود انتگرال‌های پیچیده و معمولا با بعد بالا برای محاسبه تابع درستنمایی مدل، بسیار مشکل و چالش‌برانگیز است (بریسلو و کلیتون، ۱۹۹۳؛ مک کالاک، ۱۹۹۷) . به همین دلیل، معمولا برای تحلیل و استنباط در این مدل‌ها از دیدگاه بیزی استفاده می‌شود (دیگل و همکاران، ۱۹۹۸). دلیل معمول برای انتخاب دیدگاه بیزی، برازش ساده مدل به خاطر وجود الگوریتم‌های مونت کارلوی زنجیر مارکوفی MCMC است (رابرت و کسلا، ۲۰۰۵). در این فصل، از الگوریتم‌های MCMC برای

برازش یک مدل SGLM در نواحی کوچک (شهرستان‌های استان گیلان) برای تحلیل تعداد بیمه‌شدگان اجباری، استفاده می‌کنیم.

در ادامه، در بخش (۲.۴) به معرفی داده‌ها و در بخش بعدی نیز به معرفی مدل بیزی و در آخر نیز به برازش و تحلیل مدل پیشنهادی می‌پردازیم.

۲.۴ معرفی داده‌ها

در صنعت بیمه، بنا بر مبنای قانونی، اهداف و مقررات حاکم، دو دسته بیمه به نام‌های، بیمه‌های اجتماعی و بیمه‌های بازرگانی وجود دارند.

تعریف ۱.۲.۴. (بیمه‌های اجتماعی یا بیمه‌های اجباری) این نوع بیمه، به بیمه‌های ناشی از قانون نیز معروف هستند که بیشتر به کارگران و اقشار کم‌درآمد جامعه مربوط می‌شوند. قانون‌گذار برای اقشاری از جامعه که از یک سو نیروی تولیدی جامعه محسوب می‌شوند و از سوی دیگر خود به فکر آینده و معیشت خود نیستند یا برای آینده خود نمی‌توانند برنامه مناسبی داشته باشند، دولت را موظف کرده است که برای حمایت از آن‌ها، بیمه‌های اجتماعی را تعریف کند و این اقشار را زیر چتر این نوع بیمه‌ها قرار دهد.

تعریف ۲.۲.۴. (بیمه‌های بازرگانی) این نوع بیمه، به بیمه‌های اختیاری نیز معروف هستند، به میل و اراده خود بیمه‌گذار تعیین می‌شوند. در این نوع بیمه‌ها تعهد دوطرفه است؛ یعنی بیمه‌گر به ازای دریافت حق بیمه از بیمه‌گذار، تأمین پوشش‌های بیمه‌ای مشخصی را در اختیار وی قرار می‌دهد.

بنا به اهمیت نوع بیمه اجباری، در این پایان‌نامه داده‌های بیمه اجباری استان گیلان را بر اساس یک مدل بیزی نواحی کوچک فضایی تحلیل می‌کنیم. در این مجموعه داده، متغیر پاسخ تعداد بیمه‌شدگان اجباری در ۱۶ شهرستان استان گیلان است. همچنین متغیرهای تبیینی شامل تعداد خانوارها، جمعیت و مستمری تبعی در هر شهرستان هستند. نمایی از ناحیه فضایی تحت مطالعه و نقشه همسایگی شهرستان‌های استان گیلان در شکل ۱.۴ مشاهده می‌شود.

شکل ۱.۴: نقشه همسایگی شهرستان‌های استان گیلان

۳.۴ معرفی مدل بیزی

با توجه به شمارشی بودن متغیر پاسخ، از توزیع پواسون با تابع پیوند لگاریتمی برای مدل‌بندی آن استفاده کردیم. چرا که همان‌طوری که اشاره داشتیم، داده‌های بیمه نیز ماهیت شمارشی دارند. بنابراین با استفاده از مدل (۳۵.۲) برای داده‌های مورد نظر می‌توان مدل پیشنهادی زیر را در نظر گرفت.

$$\log(\boldsymbol{\mu}(s_i)) = \boldsymbol{\eta}(s_i) = \boldsymbol{\beta}_\circ + \boldsymbol{\beta}_۱\mathbf{x}_۱(s_i) + \boldsymbol{\beta}_۲\mathbf{x}_۲(s_i) + \boldsymbol{\beta}_۳\mathbf{x}_۳(s_i) + \mathbf{v}(s_i) + \boldsymbol{\epsilon}(s_i), \quad i = ۱, ..., ۱۶ \tag{۱.۴}$$

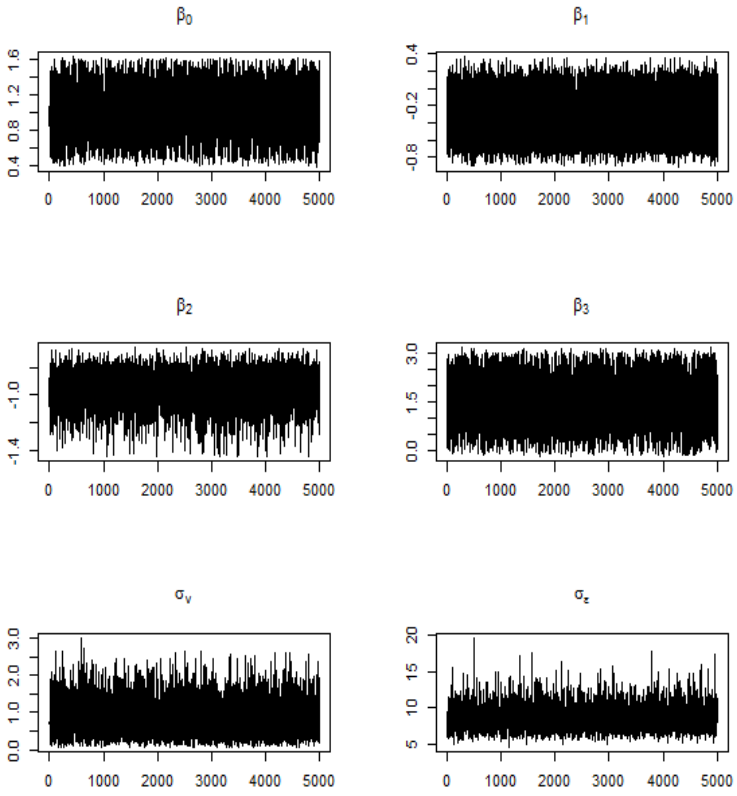
و در آن $\mathrm{x}_۱$ تعداد خانوار، $\mathrm{x}_۲$ تعداد جمعیت و $\mathrm{x}_۳$ تعداد مستمری تبعی را نشان می‌دهد. برای برازش مدل همه متغیرهای تبیینی را استاندارد کردیم. برای تشکیل مدل بیزی، باید توزیع‌های پیشین پارامترها را مشخص کنیم. در مدل (۱.۴) برای همه پارامترهای رگرسیونی توزیع پیشین نرمال $\mathrm{N}(0, 100)$ را در نظر گرفتیم و برای دو پارامتر رگرسیونی $\sigma_v^۲$ و $\sigma_\epsilon^۲$ توزیع گامای معکوس با مقادیر ابرپارامتر ۰/۰۰۱ و ۰/۰۰۱ را تعیین کردیم. بنابراین مدل بیزی کامل

برای داده‌های بیمه را می‌توان به صورت مدل سلسله‌مراتبی زیر نوشت:

$$\begin{aligned} \mathbf{y}(s_i) &\sim Pois(\boldsymbol{\mu}(s_i)) \\ \log(\boldsymbol{\mu}(s_i)) &= \boldsymbol{\eta}(s_i) \\ \boldsymbol{v} &\sim \text{CAR}(0, \boldsymbol{\sigma}_v^2 A^{-1}) \\ \boldsymbol{\epsilon}(s_i) &\sim \text{N}(0, \boldsymbol{\sigma}_\epsilon^2) \end{aligned}$$

۴.۴ برازش مدل و تحلیل نتایج

برنامه BUGS در سال (۱۹۸۹) توسط اندرو توماس و دیوید اشپیگل‌هالتر معرفی شد. برای برازش مدل، از یک الگوریتم MCMCs استفاده کردیم که در نرم‌افزار Openbugs اجرا شد. یک نمونه مونت کارلویی به حجم ۱۰۰۰۰ تولید کردیم که دوره سوزاندن الگوریتم را ۵۰۰۰ انتخاب کردیم. همچنین طول گام انتخاب نمونه‌ها را ۱۰ در نظر گرفتیم که از وابستگی نمونه‌ها بکاهیم. بنابراین همه نتایج بر اساس ۱۰۰۰ نمونه نهایی گزارش شده‌اند. نمودارهای اثر نمونه‌های تولید شده برای پارامترهای توزیع پسین در شکل ۲.۴ نمایش داده شده‌اند. به سادگی می‌توان دریافت رفتار آمیختگی زنجیر برای همه پارامترهای مدل خوب است.



شکل ۲.۴: نمودارهای اثر پارامترهای مدل بیزی داده‌های بیمه

برآورد پارامترهای مدل به همراه نواحی اعتبار (CI) ۹۵ درصد متناظر هر کدام در جدول ۱.۴ گزارش شده‌اند. با توجه به اعداد جدول، می‌توان گفت متغیر تعداد خانوار بر پاسخ تاثیر معنی‌داری ندارد. اما در مقابل جمعیت هر شهرستان اثر منفی معنی‌دار بر تعداد بیمه‌های اجباری دارد و تعداد مستمری تبعی اثر مثبت معنی‌دار دارد. از طرفی، حضور اثر فضایی در مدل لازم است. اما واریانس بزرگ جمله خطا در مقابل اثر فضایی، حاکی از آن است که اطلاعات مهم دیگری هستند که بر تعداد بیمه‌های اجباری اثر دارند ولی نادیده گرفته شده‌اند. بنابراین، تحلیل این داده‌ها و برآوردهای نواحی کوچک برای تعداد کل بیمه‌شدگان در سطح شهرستان‌ها نیازمند دقت بیشتر و در نظر گرفتن سایر اطلاعات مهم است.

جدول ۱.۴: نتایج برازش مدل بیزی بر روی داده‌های بیمه اجباری استان گیلان

اثر	برآورد	انحراف معیار	کران پایین CI	کران بالا CI
β_0	۱/۰	۰/۳	۰/۵	۱/۶
β_1	−۰/۴	۰/۳	−۰/۸	۰/۲
β_2	−۰/۹	۰/۱	−۱/۳	−۰/۷
β_3	۱/۳	۰/۸	۰/۰	۲/۹
σ_v^2	۰/۹	۰/۵	۰/۱	۱/۹
σ_ϵ^2	۸/۶	۱/۶	۶/۰	۱۲/۴

۵.۴ نتیجه‌گیری و آینده تحقیق

در این پایان‌نامه، از یک مدل CAR برای مدل‌بندی داده‌های فضایی مشبکه‌ای استفاده کردیم. برازش مدل بیزی مطرح شده نیز به کمک الگوریتم‌های MCMC انجام شد. اهمیت این نوع مدل‌ها برای مدل‌بندی داده‌های فضایی نواحی کوچک، به دلیل عملی و مناسب بودن برازش و تفسیر ساده مدل، در کاربردهای واقعی قابل توجه است. با توجه به مطالبی که در این پایان‌نامه مطرح شد، برای آینده تحقیق می‌توان استنباط بیزی برآورد کوچک ناحیه‌ای را با مدل‌های ICAR و MCAR انجام داد که به نوبه خود می‌تواند جالب توجه و اهمیت داشته باشد. و همچنین استفاده از مدل‌های شمارشی مانند مدل‌های دوجمله‌ای، بتا ـ دوجمله‌ای، برنولی، دوجمله‌ای منفی و وایبل گسسته و نیز به‌کارگیری از الگوریتم INLA می‌تواند مفید و پرکاربرد باشد.

مراجع

[۱] محمدزاده م، (۱۳۹۴)، ،«آمارفضایی» چاپ اول، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران.

[2] Anderson, T.W., and Hsiao, C. (1981), Formulation and estimation of dynamic models using panel data, *Journal of Econometrics 18*, 67–82.

[3] Arora, V., and Lahiri, P. (1997), On the superiority of the bayesian method over the BLUP in small area estimation problems, Statistica Sinica, 7, 1053–1063.

[4] Arora, V., Lahiri, P. and Mukherjee, K. (1997), Empirical bayes estimation of finite population means from complex surveys, *Journal of the American Statistical Association, 92(440),* 1555–1562.

[5] Brewer, K.R.W. (1963), Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process, *Australian Journal of Statistzcs, 5,* 5–13.

[6] Brackstone, G.J. (1987), Small area data: Policy issues and technical challenges, in R. Platek, J.N.K. Rao, C.E., Sarndal and M.P., Singh (Eds.,), Small Area Statistics, New York: Wiley, pp. 3–20.

[7] Battese, G.E., Harter, R.M., and Fuller, W.A. (1988), An error-components model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association,83,* 28–36.

[8] Besag, J., J.C., York, and A., Molline (1991), Bayesian image restoration, with two applications in spatial statistics, Annals of the Institute of Statistical Mathematics 43, 1-59.

[9] Breslow, N., and Clayton, D. (1993), Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association, 88,* 9–25.

[10] Bell, W.R., (1999), Accounting for uncertainty about variances in small area estimation, Bulletin of The International Statistical Institute,

[11] Banerjee, S., B.P., Carlin, and A.E., Gelfand (2004), Hierarchical Modeling and Analysis for Spatial Data, Chapman Hall-CRC, Boca Raton, Florida.

[12] Best, N., Richardson, S., and Thomson, A., (2005), A comparison of bayesian spatial models for disease mapping, Statistical Methods in Medical Research, 14, 35–59.

[13] Banerjee, S., Carlin, B.P., and Gelfand, A.E., (2014), Hierarchical Modeling and Analysis for Spatial Data. CRC Press, New york.

[14] Cressie, N. (1991), Small-area prediction of undercount using the general linear model, Proceedings of Statistics Symposium 90: Measurement and Improvement of Data Quality, Ottawa: Statistics Canada, pp. 93–105.

[15] Casella, G., and George, E.I., (1992), Explaininig the gibbs sampler, The American Statistician 46(3), 167-174, 39, 41.

[16] Cressie, N. (1993), Statistics for Spatial Data, New York: Wiley.

[17] Chaudhuri, A. (1993), Small domain statistics: A review, Statistica Neerlandica ,48, 215–236.

[18] Chib, S., and Greenberg, E. (1995), Understanding the metropolis-hastings algorithm, American Statistician, 49, 327–335.

[19] Cochran, W.G., (1977), Sampling techniques (3rd Eds.,), John Wiley, Sons, Inc., New York.

[20] Carlin, B.P., and Louis, T.A., (2000), Bayes and Empirical Bayes Methods for Data Analysis (2nd Eds.), London: Chapman and Hall.

[21] Datta, G.S., and Ghosh, M., (1991), Bayesian prediction in linear models: Applications to small area estimation, The Annals of Statistics, 17, 48–1770.

[22] Datta, G.S., Ghosh, M., Nangia, N., and Natarajan, K., (1996), Estimation of median income of four-person families: A bayesian approach, in W.A. Berry, K.M., Chaloner, and J.K., Geweke (Eds.), Bayesian Analysis in Statistics and Econometrics1,New York: Wiley, pp. 129–140.

[23] Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics, *Journal of the Royal Statistical Society: Series C Applied Statistics, 47(3),* 299–350.

[24] Datta, G.S., Lahiri, P., Maiti, T., and Lu, K.L. (1999), Hierarchical bayes estimation of unemployment rates for the u.s., states, *Journal of the American Statistical Association, 94,* 1074–1082.

[25] Datta, G.S., Kubakawa, T., and Rae, J.N.K., (2002), Estimation of MSE in small area estimation, Technical Report, Department of Statistics, University of Georgia, Athens.

[26] Diggle, P., Moyeed, R., Rowlingson, B., and Thomson, M., (2002), Childhood malaria in the gambia: A case-study in model-based geostatistics, *Journal of the Royal Statistical Society*: Series C (Applied Statistics), 51(4), 493–506.

[27] Eberly, L.E., and Carlin, B.P., (2000). Identifiability and convergence issues for markov chain monte carlo fitting of spatial models, Statistics in Medicine, 19(17–18), 2279–2294.

[28] EURAREA Consortium (2004), Project reference volume. Technical Report, EURAREA Consortium .

[29] Fuller, W.A., and Battese, G.E., (1973), Transformations for estimation of linear models with nested-error structure, *Journal of the American Statistical Association, 68,* 626–632.

[30] Fay, R.E., and Herriot, R.A., (1979), Estimates of income for small places: An application of james-stein procedures to census data, *Journal of the American Statistical Association, 74(366a),* 269–277.

[31] Fuller, W.A., (1989), Prediction of true values for the measurement error model, in conference on statistical analysis of measurement error models and applications, Humboldt State University.

[32] Fuller, W.A., (1999), Environmental surveys over time, *Journal of Agricultural, Biological and Environmental Statistics, 4,* 331–345.

[33] Gonzalez, J.F., Placek, P.J., and Scott, C., (1966), Synthetic estimation of followback surveys at the national center for health statistics, in W.L., Schaible (Eds.,), In direct Estimators in U.S, Federal Programs, Springer-Verlag: New York, pp. 16–27.

[34] Gonzalez, M.E., and Wakesberg, J., (1973), Estimation of the error of synthetic estimates, Paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

[35] Gonzalez, M.E., (1973), Use and evaluation of synthetic estimates, proceedings of the social statistics section, American Statistical Association, pp. 33–36.

[36] Gonzalez, M.E., and Hoza, C., (1978), Small-area estimation with application to unemployment and housing estimates, *Journal of the American Statistical Association, 73,* 7–15.

[37] Gelfand, A.E., and Smith, A.F.M., (1990), Sample-based approaches to calculating marginal densities, *Journal of the American Statistical Association, , 85,* 972–985.

[38] Gelfand, A.E., and Smith, A.F.M., (1991), Gibbs sampling for marginal posterior expectations, Communications in Statistics-Theory and Methods, 20, 1747–1766.

[39] Gilks, W.R., and Wild, P., (1992), Adaptive rejection sampling for gibbs sampling, Applied Statistics, 41, 337–348.

[40] Ghosh, M., (1992), Constrained bayes estimation with applications, *Journal of the American Statistical Association, 87(418),* 533–540.

[41] Ghosh, M., and Nangia, N., (1993), Estimation of median income of fourpers on families: A bayesian time series approach, technic report, depar, tment of Statistics, University of Florida, Gainsville.

[42] Ghosh, M., and J.N.K., Rao (1994), Small area estimation: An Appraisal, In: Statistical Science 9 (1), pp. 55–-93.

[43] Greenland, S. and J., Robins (1994), Ecologic studies biases, misconceptions, and counterexamples, *American Journal of Epidemiology, 139 (8),* 747–760.

[44] Gilks, W.R., Richardson, S., and Spiegelhalter, D. (Eds.,) (1995), Markov chain Monte Carlo in practice, CRC press.

[45] Gilks, W.R., Richardson, S., and Spiegelhalter, D.J., (Eds.,) (1996), Markov chain monte carlo in practice, London: Chapman and Hall.

[46] Ghosh, M., Nangia, N., and Kim, D. (1996), Estimation of median income of four-person families: A bayesian time series approach, *Journal of the American Statistical Association*, 91, 1423–1431.

[47] Ghosh, M., Natarajan, K., Stroud, T.W.F., and Carlin, B.P., (1998), Generalized linear models for small-area estimation, *Journal of the American Statistical Association, 93(441),* 273–282.

[48] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B., (2003), Bayesian data analysis, Chapman and Hall/CRC Texts in Statistical Science.

[49] Gelman, A. (2006), Prior distributions for variance parameters in hierarchical models comment on article by browne and draper. Bayesian Analysis, 1(3), 515–534.

[50] Hastings, W.K., (1970), Monte carlo sampling methods using markov chains and their applications, Biometrika, 57, 97–109.

[51] Hansen, M.H., Madow, W.G., and Tepping, B.J., (1983), An evaluation of model-dependent and probability sampling inferences in sample surveys, *Journal of the American Statistical Association, 78,* 776–793.

[52] Hedayat, A,S., and Sinha, B.K., (1991), Design and inference in finite population sampling, New York: Wiley.

[53] Harville, D.A., (1991), Comment, Statistical Science, 6, 35-39.

[54] Hodges J, Sargent D., (2001), Counting degrees of freedom in hierarchical and other richly parameterized models. Biometrika, 88, 367–379.

[55] Isaki, C.T., Tsay, J.H., and Fuller, W.A., (2000), Estimation of census adjustment factors, Survey Methodology, 26, 31–42.

[56] Jiang, J. and Lahiri, P., (2006)., Mixed Model Prediction and Small Area Estimation. Test, 15, 1–96.

[57] Jackson, C., Best, N., and Richardson, S., (2006)., Improving ecological inference using individual-level data. Statistics in medicine, 25(12), 2136–2159.

[58] Jiang, J., and P., Lahiri (2006)., Mixed Model Prediction and Small Area Estimation, In: Test 15 (1), pp. 1–96. Annals of Statistics , 30, 1782–1810.

[59] Kleffe, J., and Rao, J.N.K., (1992), Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model. *Journal of Multivariate Analysis, 43(1),* 1–15.

[60] Kleinschmidt I., Bagayako M., Clarke GPY, et al(2000), A Spatial Statistical Approach to Malaria Mapping, *International Journal of Epidemiology, 29,* 355–-361.

[61] Levy, P.S., (1971), The TJse of Mortality data in evaluating synthetic estimats, Proceedings of the social statistics section, American Statistical Association, pp. 328–331.

[62] Louis, T.A., (1984), Estimating a population of parameter values using bayes and empirical bayes methods, *Journal of the American Statistical Association, 79(386),* 393–398.

[63] Lahiri, P. (1990), Adjusted bayes and empirical bayes estimation in finite population sampling, Sankhyā: *The Indian Journal of Statistics, Series B,* 50–66.

[64] Lohr, S.L., (1999), Sampling: Design and Analysis, Pacific Grove; CA:Duxbury.

[65] LeSage, J.P., and Pace, R.K., (2004), Models for spatially dependent missing data, *The Journal of Real Estate Finance and Economics, 29(2),* 233–254.

[66] Lin, R., Louis, T.A., Paddock, S.M., and Ridgeway, G., (2009), Ranking USRDS provider specific SMRs from 1998–2001, Health Services and Outcomes Research Methodology, 9(1), 22–38.

[67] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E., (1953), Equations of state calculations by fast computing machines, *Journal of Chemical Physics, 21,* 1087–1091.

[68] McCulloch, C.E., (1997), Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association, 92,* 162–170.

[69] Mukhopadhyay, P. (1998), Small area estimation in survey sampling, New Delhi: Narosa Publishing House.

[70] Marker, D.A., (1999), Organization of small area estimators using a generalized linear regression framework, *Journal of Official Statistics, 15,* 1–24.

[71] McCullogh CE, Searle SR (2001). Generalized, Linear, and Mixed Models. John Wiley and Sons, Inc., New York.

[72] Moura, F.A.S., and Meung, D. (2002), Small area estimation using multilevel models, Survey Methodology, 25, 73–80.

[73] Milana Karaganis, (2009), A hierarchical bayes approach, department of statistics, The University of Manitoba.

[74] Normand, S.L.T., Glickman, M.E., and Gatsonis, C.A., (1997), Statistical methods for profiling providers of medical care: issues and applications, *Journal of the American Statistical Association, 92(439),* 803–814.

[75] Purcell, N.J., and Kish, L., (1980), Postcensal estimates for local areas (or Domains), International Statistical Review, 48, 3–18.

[76] Platek, R. and Singh, M.P., (Eds.,) (1986), Small area statistics: Contributed papers, Laboratory for research in statistics and probability, Carleton university, Ottawa, Canada.

[77] Platek, R., Rao, J.N.K., Sarndal, C.E., and Singh, M.P., (Eds.,) (1987), Small Area Statistics, New York: Wiley.

[78] Pfeffermann, D., and Burck, L., (1990), Robust small area estimation combining time series and cross-sectional data, Suruey Methodology, 16, 217–237.

[79] Pfeffermann, D. (2002), Small area estimation-new developments and directions, In: International Statistical Review, 70 (1), pp. 125–-143.

[80] Petrucci, A., and Salvati, N. (2006), Small area estimation for spatial correlation in watershed erosion assessment, *Journal of Agricultural, Biological, and Environmental Statistics, 11,* 2–169.

[81] Paddock, S.M., Ridgeway, G., Lin, R., and Louis, T.A., (2006), Flexible distributions for triple-goal estimates in two-stage hierarchical models, Computational Statistics and Data Analysis, 50(11), 3243–3262.

[82] Pratesi, M. Salvati, N. (2008), Small area estimation: The EBLUP estimator based on spatially correlated random area effects, Statistical Methods and Applications, 17, 113–141.

[83] Pfeffermann, D. (2013), New important developments in small area estimation, Statistical Science, 28, 40–68.

[84] Royal, R.M., (1970), On finite population sampling theory under certain linear regression, Biometrika, 57, 377–387.

[85] Rao, J.N.K., (1979), On deriving mean square errors and their non-negative unbiased estimators in finite population sampling, *Journal of the Indian Statistical Association, 17,* 125–136.

[86] Rao, J.N.K., (1986), Synthetic estimators, SPREE and best model based predictors, Proceedings of the conference on survey research methods in agriculture, US, Department of Agriculture, Washington, DC, pp. 1–16.

[87] Robinson, G.K., (1991), That BLUP is a good thing: The estimation of random effects, Statistical science, 6(1), 15–32.

[88] Rao, J.N.K., and Yu, M. (1992), Small area estimation by combining time series and cross-sectional data, Proceedings of the section on survey research method, American Statistical Association, pp. 1–9.

[89] Rao, J.N.K., (1992), Estimating totals and distribution functions using auxiliary information at the estimation stage, Proceedings of the workshop on uses of auxiliaq information in surveys, Statistics Sweden.

[90] Rao, J.N.K., and Yu, M. (1994), Small area estimation by combining time series and cross-sectional data, *Canadian Journal of Statistics, 22,* 51 1–528.

[91] Rao, J.N.K., and Choudhry, G.H., (1995), Small area estimation: Overview and empirical study, in B.G., Cox, D.A., Binder, B.N., Chinnappa, A. Christianson, M.J., Colledge, and P.S., Kott (Eds.,), Business survey methods, New York: Wiley, pp. 527–542.

[92] Rao, J.N.K., (2001), EB and EBLUP in small area estimation, in S.E., Ahmed and N. Reid (Eds.,), Empirical bayes and likelihood inference, Lecture notes in statistics 148, New York: Springer, pp. 33–43.

[93] Rao, J.N.K., (2003), Small Area Estimation, Wiley, Hoboken, New Jersy.

[94] Robert, C.P., and Casella, G., (2005), Monte carlo statistical methods, Second Ed., Springer-Verlag, New York.

[95] S. Richardson, G. Li., (2010), Bayesian statistics for small area estimation, Department of epidemiology and public health imperial college london st, Mary's campus, Norfolk place w2 1PG London-United Kingdom.

[96] Rao, J.N.K., and Molina, I., (2015), Small Area Estimation, 2nd Edition, Wiley, New York.

[97] Smith, S.K., and Lewis, B.B., (1980), Some New Techniques for Applying the Housing Unit Method of Local Population Estimations, Demography, 17, 349–340.

[98] Smith, T.M.F., (1983), On the validity of inferences from non-random samples, *Journal of the Royal Statistical Society, Series A, 146,* 394–403.

[99] Särndal, C.E., B. Swensson and J.H., Wretman (1992), Model Assisted Survey Sampling, New York: Springer.

[100] Sardnal, C.E., Swensson, B., and Wretman, J.H., (1992), Model assisted survey sampling, New York Springer-Verlag.

[101] Singh, A.C., Mantel, J.H., and Thomas, B.W., (1994), Time series EBLUPs for small areas using survey data, Survey Methodology, 20, 33–43.

[102] Skinner, C.J., (1994), Sample models and weights, Proceedings of the section on survey research methods, American statistical association, Washington, DC, pp. 133–142.

[103] Spiegelhalter, D.J., N.G., Best, B.P., Carlin, and A. van Der Linde (2002), Bayesian measures of model complexity and

t with discussion, J.R., Statist. Soc. B., 64, 583–639.

[104] Salvati, N. (2004), Small area estimation by spatial models: the spatial empirical best linear unbiased prediction spatial EBLUP, Working Paper 2004/03, University of Florence.

[105] Singh, B. B., G.K., Shukla, and D. Kundu (2005), Spatio-Temporal models in small area estimation, Survey Methodology 31 (2), 183–196.

[106] Tanner, M.A., and Wong, W.H., (1987), The calculation of posterior distributions by data augmentation with discussion, *Journal of the American Statistical Association, 82,* 528–550.

[107] Thompson, M.E., (1997), Theory of sample survey, London: Chapman and Hall.

[108] You, Y. (1999), Hierarchical bayes and related methods for model-based small area estimation, Unpublished Ph.D. Thesis, Carleton University, Ottawa, Canada.

[109] Yar M, Hennell S, Clarke P, Meltzer H, Gatward R., (2002), Childhood mental disorder in england: Ward estimates, Technical report, Office for National Statistics, United Kingdom.

[110] Valliant, R., Dorfmap, A.H., and Royall, R.M., (2001), Finite Population Sampling and Inference: A Prediction Approach, New York: Wiley.

[111] Zhang W. Li, N., (2011), Prevalence, risk factors, and management of prehypertension, *International Journal of Hypertension, 60,* 53–59.

پیوست آ

توزیع‌های شرطی کامل برای مشخص‌سازی CAR با مشاهدات گم‌شده

فرض کنید یک مدل سطح ناحیه دارید و داده از نواحی lام اول وجود دارد بنابراین، مقادیر برآوردگرهای مستقیم و واریانس‌های نمونه‌گیری برای این نواحی نیز وجود داشته باشد. در این زمینه، $s = ۱, ..., l$ و $\underline{s} = l + ۱, ..., m$ است. توزیع‌های شرطی کامل برای این مدل با استفاده از نمونه‌گیری گیبز به صورت زیر است:

$$\Pi(\alpha, \beta|...) = N((\bar{X}^T V^{-۱} \bar{X})^{-۱} \bar{X}^T V^{-۱} (\hat{\bar{Y}} - v), (\bar{X}^T V^{-۱} \bar{X})^{-۱})$$

پس،

$$\hat{\bar{Y}} = (\hat{\bar{Y}}_۱, ..., \hat{\bar{Y}}_l)^T, v = (v_۱, ..., v_l)^T, V = diag(\hat{V}_۱^۲, ..., \hat{V}_l^۲)$$

و

$$\bar{X}^T = \begin{pmatrix} ۱ & ... & ۱ \\ \bar{X}_۱ & ... & \bar{X}_l \end{pmatrix}$$

$$\Pi(v_i|...) \propto exp\{\frac{-۱}{\hat{\sigma}_i^۲}(\hat{\bar{Y}}_i - \alpha - \beta \bar{X}_i - v_i)^۲ + \frac{۱}{\sigma_v^۲} n_i (v_i - \bar{v}_i^۲)\}, \quad i \epsilon s$$

$$\Pi(v_j|...) \propto exp\{\frac{-۱}{\sigma_v^۲} n_j(v_j - \bar{v}_j)^۲\}, \quad j \epsilon \underline{s}$$

لذا به ترتیب n_i و n_j نشان‌دهنده‌ی تعدادی از همسایگی‌های نواحی i و j می‌باشند.

$$\Pi(\sigma_v^۲|...) = Ga^{-۱}(\epsilon + \frac{n}{۲}, \epsilon + \frac{۱}{۲}\Sigma_{k\sim l}(v_k - v_l)^۲)$$

از این رو، شرطی کامل برای هر v_i مستقیما وابسته بر مشاهدات داده و همسایگی‌های آن است، ضمن اینکه v_j تنها وابسته بر روی همسایگی‌های $\bar{v}_j$ و غیرمستقیم بر روی مشاهدات داده می‌باشد. علاوه بر این، همان‌گونه که انتظار می‌رفت α و β تنها اطلاعاتی هستند که بوسیله داده از نواحی تحت بررسی به‌دست آورده می‌شوند.

پیوست ب

دستورات نرم‌افزار R و OPENBUGS

در این قسمت دستورات مربوط به کدنویسی داده‌های بیمه استان گیلان ارائه شده است.

```
model
{
for(i in 1:N)
{
y[i] ~ dpois(mu[i])
log(mu[i]) <- alpha + beta1*x1[i] + beta2*x2[i] +beta3*x3[i] + u[i] + v[i]
u[i] ~ dnorm(0, precu)
}
v[1:N] ~ car.normal(adj[], weights[], num[], precv)
for (k in 1:sumNumNeigh){ weights[k] <- 1}
alpha ~ dflat()
beta1 ~ dnorm(0,1.0E-5)
beta2 ~ dnorm(0,1.0E-5)
beta3 ~ dnorm(0,1.0E-5)
```

```
precu ~ dgamma(0.001, 0.001)
precv ~ dgamma(0.1, 0.1)
sigmau<-1/precu
sigmav<-1/precv
}

#liatticels
#Data

list(N=16,
num=c(1 ,2, 3, 2, 6, 4, 3, 4 ,3, 3, 6, 3,
 4, 4, 7, 3),
adj=c(4 ,
15, 8,
13, 11, 15,
13, 1,
15, 10, 7, 12, 14, 8,
10, 9, 11, 15,
5, 12, 14,
2, 15, 5, 14,
16, 11, 6,
6, 15, 5,
15, 3, 6, 9, 16, 13,
7, 5, 14,
4, 16, 11, 3,
8, 5, 12, 7,
3, 11, 6, 10, 5, 8, 2,
13, 11, 9
),
y=c(5547,4822, 17454,
6317,
3538,
4112,
5195 ,
```

```
14081,
5,573,
12170,
7422,
3138,
5410,
7977,
115632,
2283 ),


x1=c(28742, 38824, 48193, 61055, 16351, 18416, 51586, 58378,

31209, 13220, 41975, 15306, 22246, 49351, 179456, 16901),

x2=c(91257, 108130, 139016, 200649, 46975, 54226, 147399, 167544,

92310, 42408, 125074, 43225, 69865, 140686, 523749, 52649),

x3=c(3286, 2417,9657,4272, 1094,1647, 3,069,6538,2606, 8985, 3591,
 1184, 4720, 4578, 55630, 1129),
sumNumNeigh = 58)
list(alpha=0, beta1=0.5, beta2=0.5, beta3=0.5, precv=0.01, precu=0.01),
precu ~ dgamma(0.001, 0.001)
precv ~ dgamma(0.1, 0.1)
sigmau<-1/precu

rm(list=ls())

library(R2OpenBUGS)

l.model <- function(){
## Likelihood
for(i in 1:N){
```

```
y[i] ~ dpois(mu[i])
log(mu[i]) <- beta[1] + beta[2]*x1[i] + beta[3]*x2[i] +
beta[4]*x3[i] + v[i] + u[i]
u[i] ~ dnorm(0, precu)
# Area-specific relative risk (for maps)
# RR[i] <- exp(beta0 + beta1 * x1[i] +
# beta2 * x2[i] + beta3 * x3[i] + v[i])
}
# CAR prior distribution for random effects:
v[1:N] ~ car.normal(adj[], weights[], num[], precv)
for (k in 1:sumNumNeigh){
weights[k] <- 1
}
precu ~ dgamma(0.01, 0.01)
precv ~ dgamma(0.01, 0.01)
sigmau <- sqrt(1/precu)
sigmav <- sqrt(1/precv)
#
for (j in 1:4){
beta[j] ~ dnorm(0,.01)
}
}

# Loading data

l.data <- list(N = 16,
num=c(1 ,2, 3, 2, 6, 4, 3, 4 ,3, 3, 6, 3, 4, 4, 7, 3),
adj=c(4 ,
15, 8,
13, 11, 15,
13, 1,
15, 10, 7, 12, 14, 8,
10, 9, 11, 15,
5, 12, 14,
```

```
2, 15, 5, 14,
16, 11, 6,
6, 15, 5,
15, 3, 6, 9, 16, 13,
7, 5, 14,
4, 16, 11, 3,
8, 5, 12, 7,
3, 11, 6, 10, 5, 8, 2,
13, 11, 9
),
y=c(5547,4822, 17454
, 6317,
3538,
4112,
5195 ,
14081,
5,573 ,
12170 ,
7422 ,
3138,
5410,
7977,
115632,
2283 ),


x1=c(-0.363253508829662, -0.109955722412744, 0.125428820429955,
0.448570670151676, -0.674562065811612, -0.622681493608584, 0.210673750931106,
0.381314354516079, -0.301273183587593, -0.753224570202742, -0.0307907427460408,
-0.70081639896036, -0.526457478336138, 0.154522138885697, 3.42324942531429,
-0.660743995733324),

x2=c(-0.314534366653528, -0.16939426940249, 0.0962844506648097,
0.626446224734223, -0.695444328513923, -0.633071849099043, 0.168394296883944,
```

```
0.341679857194627, -0.305476551066179, -0.734729270382322, -0.0236434362476521,
-0.727701506531549, -0.498546513894339, 0.110649647275325, 3.40572468273282,
-0.646637067694722),

x3=c(-0.252182437726062, -0.319255229612263, 0.239556062972422,
-0.176079131857485, -0.421369503105293, -0.378686817359529, -0.505577116393952,
-0.500482980301329, -0.00118045934410152, -0.304667476256115,
0.187688495483899, -0.228641354267729, -0.41442295388808, -0.141500753531803,
-0.152460864518961, 3.78793058700498, -0.418668067298599),
sumNumNeigh = 58)

l.inits <- function() {
list(beta = c(0, 0, 0, 0), precv=0.1, precu=0.1,
v= c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
u= c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
}

### Fitting model

l.out <- bugs(data = l.data, inits = l.inits,
parameters.to.save = c("beta", "sigmav", "sigmau"),
model.file = l.model,
n.chains = 1, n.thin=10, n.iter = 10000, debug=FALSE)

l.out

l.out$summary

par(mfrow=c(2,3))
#Intercept
plot(density(l.out$sims.list$beta[,1]), main=expression(alpha), xlab="")
#abline(v = seq(-0.5, 1), col = "grey")
abline(v = mean(l.out$sims.list$beta[,1]), col = "red", lwd = 1)
abline(v = median(l.out$sims.list$beta[,1]), col = "red", lwd = 1, lty = 2)
```

```
plot(density(l.out$sims.list$beta[,2]), main=expression(beta[1]), xlab="")
abline(v = mean(l.out$sims.list$beta[,2]), col = "red", lwd = 1)
abline(v = median(l.out$sims.list$beta[,2]), col = "red", lwd = 1, lty = 2)

plot(density(l.out$sims.list$beta[,3]), main=expression(beta[2]), xlab="")
abline(v = mean(l.out$sims.list$beta[,3]), col = "red", lwd = 1)
abline(v = median(l.out$sims.list$beta[,3]), col = "red", lwd = 1, lty = 2)

plot(density(l.out$sims.list$beta[,4]), main=expression(beta[3]), xlab="")
abline(v = mean(l.out$sims.list$beta[,4]), col = "red", lwd = 1)
abline(v = median(l.out$sims.list$beta[,4]), col = "red", lwd = 1, lty = 2)

plot(density(l.out$sims.list$sigmav), main=expression(sigma[v]), xlab="")
abline(v = mean(l.out$sims.list$sigmav), col = "red", lwd = 1)
abline(v = median(l.out$sims.list$sigmav), col = "red", lwd = 1, lty = 2)

plot(density(l.out$sims.list$sigmau), main=expression(sigma[u]), xlab="")
abline(v = mean(l.out$sims.list$sigmau), col = "red", lwd = 1)
abline(v = median(l.out$sims.list$sigmau), col = "red", lwd = 1, lty = 2)

par(mfrow=c(3,2))
acf(l.out$sims.list$beta[,1], main=expression(alpha), xlab="")
acf(l.out$sims.list$beta[,2], main=expression(beta[1]), xlab="")
acf(l.out$sims.list$beta[,3], main=expression(beta[2]), xlab="")
acf(l.out$sims.list$beta[,4], main=expression(beta[3]), xlab="")
acf(l.out$sims.list$sigmav, main=expression(sigma[v]), xlab="")
acf(l.out$sims.list$sigmau, main=expression(sigma[u]), xlab="")

par(mfrow=c(3,2))
ts.plot(l.out$sims.list$beta[,1], main=expression(alpha), xlab="")
ts.plot(l.out$sims.list$beta[,2], main=expression(beta[1]), xlab="")
ts.plot(l.out$sims.list$beta[,3], main=expression(beta[2]), xlab="")
ts.plot(l.out$sims.list$beta[,4], main=expression(beta[3]), xlab="")
```

```
ts.plot(l.out$sims.list$sigmav, main=expression(sigma[v]), xlab="")
ts.plot(l.out$sims.list$sigmau, main=expression(sigma[u]), xlab="")
```

واژه‌نامه فارسی به انگلیسی

اتورگرسیو شرطی ..... Conditional autoregressive
برآوردگر کوچک ناحیه‌ای..... Small area estimation
برآوردگر مستقیم..... Direct estimates
برآوردگر غیرمستقیم ..... Indirect estimators
داده الگو نقطه‌ای ..... Point data
رگرسیون لجستیک..... Logistic regression
ماتریس همسایگی ..... Neighborhood matrix
مدل سطح واحد..... Unit level model
میانگین مربع خطا ..... Mean squared error
میدان تصادفی ..... Random field
هوروتیز_ تامپسون ..... Horwitz-thompson

واژه‌نامه انگلیسی به فارسی

اتورگرسیو شرطی ..... Conditional autoregressive
برآوردگر مستقیم ..... Direct estimates
هوروتیز_ تامپسون ..... Horwitz-thompson
برآوردگر غیرمستقیم ..... Indirect estimators
رگرسیون لجستیک ..... Logistic regression
میانگین مربع خطا ..... Mean squared error
ماتریس همسایگی ..... Neighborhood matrix
داده الگو نقطه‌ای ..... Point data
میدان تصادفی ..... Random field
برآوردگر کوچک ناحیه‌ای ..... Small area estimation
مدل سطح واحد ..... Unit level model

Abstract

In recent years, the problem of small-scale regionalization has to be considered because of the need reliable statistics have been very much considered. The major problem here is the impossibility of measuring the target variable for each individual in the area in question. Even sampling from all areas under study could lead to high financial and time costs. Several statistical models have been proposed to achieve this goal. The main purpose of these models is the use of auxiliary information to upgrade direct estimates. To this end, the importance of using spatial information of areas within the framework of spatial models plays an effective role in analyzing small areas. The similarity of space between adjacent areas is also useful information that small space-based models have been proposed to use for this information. In view of the fact that in many analyzes of small regions with nonnormal responses such as numeric responses, we classify the generalized linear space mixing models in the framework of Bayesian inference for the analysis of small area data. By studying the simulation and analysis of the insurance data package in Gilan province, we show the performance of these models.

Keywords: Small area estimation, Spatial models, Generalized linear Mixed models, Bayesin inference.

Shahrood University of Technology

Faculty of Mathematical Sciences

MSc Thesis in: Mathematical Statistics

Bayesian Spatial and non-Spatial Random Effects Models for Small Area Estimation

By: Motahareh Yousefi

Supervisors

Dr. Mohammad Reza Rabiei
Dr. Hossein Baghishani

September, 2018