





دانشکده علوم ریاضی

رساله دکتری احتمال

توزیع‌های دومدی-یکمدی و کاربردهای آنها

نگارنده: جمیل اونق

استادان راهنما

دکتر احمد نزاکتی
دکتر حسین باغیشنی

بهمن ۱۳۹۹

شماره: ۴۰۷-۲۲-۲
تاریخ: ۱۴۰۰/۲/۷
ویرایش:

باسمه تعالی



مدیریت تحصیلات تکمیلی

پیوست شماره ۲

دانشکده: ریاضی

گروه: آمار

رساله دکتری: آقای جمیل اونق

تحت عنوان: توزیع دومدی - یک مدی و کاربردهای آن

در تاریخ ۱۳۹۹/۱۱/۲۶ توسط کمیته تخصصی زیر جهت اخذ مدرک رساله دکتری ارزیابی گردید و با درجه عالی..... مورد پذیرش قرار گرفت.

امضاء	اساتید مشاور	امضاء	اساتید راهنما
-----	نام و نام خانوادگی: -----		نام و نام خانوادگی: دکتر احمد نزاکتی
-----	نام و نام خانوادگی: -----		نام و نام خانوادگی: دکتر حسین باغیشنی

امضاء	نماینده تحصیلات تکمیلی	امضاء	اساتید داور
	نام و نام خانوادگی: دکتر مهرداد غزنوی		نام و نام خانوادگی: دکتر داود شاهسونی
			نام و نام خانوادگی: دکتر محمدرضا ربیعی
			نام و نام خانوادگی: دکتر افشین فلاح



تقدیم بہ:

پدرم کہ راہ را بہ من نشان داد
مادرم کہ چگونہ رفتن را بہ من اموخت
خواہر و برادرم، کہ سایہ مہربانیشان سایہ سار زندگیم می باشد
ہمسرم، کہ حضورش ہمیشہ کرمانخش روح من بودہ است
پسر عزیزم الٹامی، امید بخش جانم کہ آسایش او آرامش من است.

سپاس‌گزاری

بر حسب وظیفه از تمامی اساتید بزرگواری که مرا در تحصیل علم و معرفت یاری کرده اند، از اساتید ارجمند جناب آقای دکتر احمد نزاکتی و جناب آقای دکتر حسین باغیشنی که با حسن خلق و فروتنی زحمت راهنمایی این رساله را برعهده گرفتند، از اساتید بزرگوار جناب آقای دکتر افشین فلاح، جناب آقای دکتر داود شاهسونی و جناب آقای دکتر محمدرضا ربیعی که زحمت داوری این رساله را متقبل شدند کمال تشکر و قدردانی را دارم.

جمیل اونق
بهمن ۱۳۹۹

تعهد نامه

اینجانب **جمیل اونق** دانشجوی دکتری رشته **آمار علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده رساله با عنوان **توزیع‌های دومدی-یک‌مدی و کاربردهای آن‌ها**، تحت راهنمایی **احمد نزاکتی** 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 خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- 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 شده است.

جمیل اونق

بهمن ۱۳۹۹

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) 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 نمی‌باشد.

چکیده

در عمل برای تحلیل داده‌های واقعی پذیرفتن نرمال بودن توزیع آنها، در موارد متعدد، چندان معتبر نیست و باید مدل‌هایی را به کار گرفت که قادر به در نظر گرفتن ساختار واقعی احتمالی داده‌ها باشند. برای این منظور، به توزیع‌های منعطف برای تحلیل داده‌های غیرنرمال مانند داده‌های چوله و دومی، دم‌سبک، نیمه‌دم‌سنگین و دم‌سنگین نیاز داریم. با این هدف در این رساله به معرفی توزیع‌های احتمالی می‌پردازیم که برای تحلیل داده‌های با ویژگی‌های اشاره شده مناسب هستند. به طور واضح، ابتدا دو رهیافت را برای توسعه توزیع‌های متقارن به توزیع‌های دومی-یک‌مدی متقارن و چوله ارائه می‌کنیم. سپس کاربرد آنها را در توسعه مدل‌های رگرسیون خطی استوار و فضایی تشریح و ارزیابی می‌کنیم. در بخش دوم رساله، یک مدل رگرسیون نیمه‌پارامتری مکان-مقیاس با دم‌های نیمه‌سنگین را بر پایه توزیع هایپربولیک سکانت معرفی می‌کنیم. تعمیم‌هایی چوله از توزیع متقارن هایپربولیک سکانت را توسعه و مدل رگرسیونی خطی مبتنی بر آنها را می‌سازیم. همچنین رفتار مجانبی برآوردگرهای ML مدل رگرسیونی معرفی‌شده را ارزیابی می‌کنیم. در پایان یک نسخه چندمتغیره از توزیع هایپربولیک سکانت نیز معرفی می‌کنیم.

کلمات کلیدی: توزیع‌های دومی-یک‌مدی، داده‌های پرت، چولگی، رگرسیون مکان-مقیاس، رگرسیون فضایی، مدل‌های خطی تنومند، رگرسیون مد.

فهرست مقالات مستخرج از رساله

۱. اونق، ج.، باغیشنی، ح. و نزاکی، ا. (۱۳۹۹)، مدل رگرسیون نیمه پارامتری مکان-مقیاس با دم‌های نیمه سنگین بر اساس توزیع هایپربولیک سکانت، مجله مدل‌سازی پیشرفته ریاضی، شماره ۱۰، دوره ۲، ص ۳۷۹.
2. Ownuk, J., Baghishani, H. and Nezakati, A. (2020). Heavy or semi-heavy tail, that is the question. *Journal of Applied Statistics*, 48, 1-23.
3. Ownuk, J., Nezakati, A. and Baghishani, H. (2018). Multivariate Hyperbolic Secant Distribution, The 14th Iranian Statistics Conference, Shahrood University of Technology, Shahrood, Iran.
۴. اونق، ج.، نزاکی، ا. و باغیشنی، ح. (۱۳۹۶ الف)، رفتار مجانبی برآوردهای درست‌نمایی ماکسیمم مدل رگرسیون خطی با جمله خطای دومی، سمینار احتمال و فرآیندهای تصادفی، ص ۲۲۸-۲۳۹، قزوین.
۵. اونق، ج.، باغیشنی، ح. و نزاکی، ا. (۱۳۹۶ ب)، رگرسیون فضایی برای داده‌های دومی، سمینار آمار فضایی و کاربردهای آن، ص ۱۷-۲۷، شاهرود.
6. Ownuk, J., Baghishani, H. and Nezakati, A. *On developing new robust regression models for skewed or/and outlier data*, (Under preparation)
7. Ownuk, J., Nezakati, A. and Baghishani, H. *Developing flexible classes of distributions to account for both skewness and bimodality*, (Under preparation)

پیشگفتار

در اکثر مطالعات کلاسیک معمولاً فرض بر این است که توزیع داده‌ها متقارن و دم‌سبک هستند و از توزیع نرمال پیروی می‌کنند. اما گذشت زمان و پیشرفت تکنولوژی در گردآوری داده‌های نوین باعث به‌وجود آمدن پیچیدگی‌هایی در داده‌ها شده است. در نتیجه در اکثر مواقع، پذیره نرمال بودن داده‌ها به نتایج غیرمعقولی منجر می‌شود. دلایلی از قبیل نامتقارن بودن داده‌ها، چولگی، نیمه‌دم‌سنگینی، دم‌سنگینی و داده‌های پرت، باعث ایجاد این نتایج غیرمعقول می‌شوند. از این‌رو، در سال‌های اخیر به‌کار بردن توزیع‌های غیرنرمال، و توسعه و تعمیم توزیع‌های احتمالی متقارن به نسخه‌های منعطف گسترش چشمگیری داشته است. در کنار مطالعاتی که در این سو انجام می‌شود باید محتاطانه عمل کرد تا پیچیدگی بیش از حد به توزیع‌های احتمالی نسبت به مزیتی که به ارمغان می‌آورند، وارد نشود. یکی دیگر از مشکلات تحلیل داده‌ها کمبود توزیع‌های مناسب برای تحلیل داده‌های دومدی است. گزینه‌های موجود انعطاف‌پذیری بالایی ندارند تا دومدی بودن یا نبودن تابع چگالی را با پارامترها کنترل کنند. از این‌رو، در این رساله، روش‌هایی برای تعمیم توزیع‌ها به حالت‌های با انعطاف‌پذیری بالا ارائه شده است که، بدون اضافه کردن پیچیدگی بیش از حد، انعطاف‌پذیری بالایی به توزیع‌ها می‌دهند. برخی از این تعمیم‌ها به ساختن توابع چگالی منجر می‌شوند که دارای پارامترهایی هستند که چولگی و دومدی بودن را کنترل می‌کنند. مثال‌هایی از این تعمیم‌ها برای توزیع‌های دم‌سبک، نیمه‌دم‌سنگین و دم‌سنگین ارائه و تشریح شده‌اند. با توجه به این مقدمه، ساختار این رساله به‌صورت زیر است:

- در فصل اول، تعاریف و مفاهیم مورد نیاز برای ورود به موضوع رساله را مطرح می‌کنیم.
- در فصل دوم، تعمیم‌هایی چوله از توزیع متقارن هایپربولیک سکانت معرفی و کاربرد آن‌ها شرح داده شده و مدل رگرسیون نیمه‌پارامتری مکان-مقیاس با دم‌های نیمه‌سنگین بر اساس توزیع هایپربولیک سکانت، توسعه داده شده و همچنین توزیع هایپربولیک سکانت به حالت چندمتغیره گسترش داده شده است.
- در فصل سوم، روش‌هایی برای توسعه توزیع‌های متقارن به توزیع‌های چوله و دومدی شرح داده شده و مثال‌هایی به همراه کاربردها مورد بررسی قرار گرفته است.
- در فصل چهارم، کاربرد توزیع‌های دومدی-یک‌مدی مستخرج از فصل چهارم در رگرسیون خطی تنومند و رگرسیون خطی فضایی شرح داده شده و مورد ارزیابی قرار گرفته‌اند.

فهرست مطالب

ف	فهرست تصاویر
ش	فهرست جداول
ث	فهرست نمادها
y	فهرست کلمات اختصاری
۱	۱ پیش‌نیازها و ملزومات
۱	۱.۱ مقدمه
۴	۲.۱ تعاریف
۵	۳.۱ مدل رگرسیون خطی
۵	۱.۳.۱ رگرسیون استوار
۷	۴.۱ مدل‌های خطی و جمعی تعمیم‌یافته
۸	۱.۴.۱ مدل‌های جمعی تعمیم‌یافته برای پارامترهای مکان، مقیاس و شکل
۱۰	۵.۱ اسپلاین‌ها
۱۰	۶.۱ آمار فضایی
۱۱	۱.۶.۱ داده‌های فضایی
۱۲	۲.۶.۱ میدان تصادفی
۱۳	۷.۱ توابع چگالی استفاده‌شده در این رساله
۱۷	۲ توزیع هایپربولیک سکانت و تعمیم‌های آن
۱۹	۱.۲ تعمیم‌هایی چوله از توزیع هایپربولیک سکانت
۱۹	۱.۱.۲ توزیع‌های پیشنهادی
۲۳	۲.۱.۲ مدل‌سازی رگرسیونی
۳۰	۳.۱.۲ مطالعات شبیه‌سازی
۳۴	۲.۲ مثال‌های کاربردی

۳۶	۱.۲.۲ داده‌های صوت‌شناسی
۳۸	۲.۲.۲ داده‌های الماس
۴۰	۳.۲.۲ داده‌های شرکت مارتین ماریتا
۴۲	۳.۲ مدل رگرسیون نیمه‌پارامتری مکان-مقیاس برای داده‌های پرت
۴۵	۱.۳.۲ خانواده توزیع‌های هایپربولیک سکانت
۴۸	۲.۳.۲ مطالعه شبیه‌سازی
۵۱	۳.۳.۲ مثال واقعی
۵۵	۴.۲ توزیع هایپربولیک سکانت چندمتغیره
۶۰	۱.۴.۲ تابع مولد گشتاور توزیع MHS
۶۰	۲.۴.۲ برآورد پارامترهای توزیع MHS
۶۱	۳.۴.۲ کاربرد توزیع MHS
۶۳	۵.۲ نتیجه‌گیری
۶۷	۳ توزیع‌های دومدی-یکمدی
۶۷	۱.۳ مقدمه
۶۸	۲.۳ توسعه خانواده‌های دومدی و چوله
۶۸	۱.۲.۳ روش اول
۷۰	۲.۲.۳ روش دوم
۷۳	۳.۳ برخی حالات خاص توزیع‌های دومدی-یکمدی
۷۳	۱.۳.۳ توزیع نرمال دومدی-یکمدی
۷۷	۲.۳.۳ توزیع t استیودنت دومدی-یکمدی
۸۳	۳.۳.۳ توزیع لاپلاس دومدی-یکمدی
۸۶	۴.۳ مثال واقعی
۹۰	۵.۳ نتیجه‌گیری
۹۱	۴ مدل‌های رگرسیونی مبتنی بر توزیع BUN
	۱.۴ خانواده‌ای از مدل‌های رگرسیون خطی برای مانده‌های شامل داده پرت یک
۹۱	طرفه یا چوله
۹۳	۱.۱.۴ مدل رگرسیون خطی با جمله خطای BUN
۹۹	۲.۱.۴ مطالعه شبیه‌سازی
۱۰۱	۳.۱.۴ مثال‌های واقعی
۱۰۶	۲.۴ رگرسیون فضایی برای داده‌های دومدی و چوله
۱۰۷	۱.۲.۴ مقدمه
۱۰۸	۲.۲.۴ توزیع نرمال دومدی-یکمدی چندمتغیره

۳.۲.۴	رگسیون فضایی پاسخ‌های دومدی	۱۰۹
۴.۲.۴	ارزیابی کارایی مدل پیشنهادی	۱۱۲
۵.۲.۴	نتیجه‌گیری	۱۱۵

۵ پیشنهادهای برای پژوهش‌های آینده ۱۱۷

آ	اثبات قضایا	۱۱۹
۱.آ	اثبات قضایای فصل سوم	۱۱۹
۲.آ	اثبات قضایای فصل چهارم	۱۲۵
۳.آ	اثبات قضایای فصل پنجم	۱۳۱

مراجع ۱۳۵

۱۴۷	واژه‌نامه فارسی به انگلیسی
۱۴۹	واژه‌نامه انگلیسی به فارسی

فهرست تصاویر

۲۰	۱.۲ چگالی‌های SHS ₁ برای مقادیر مختلف ν
۲۲	۲.۲ چگالی‌های SHS ₂ برای مقادیر مختلف α
۳۲	۳.۲ مقایسه دم‌ها برای چگالی‌های کوشی (نقطه-خط)، نرمال (نقطه چین)، SHS ₁ (خط ممتد) و SHS ₂ (خط چین)
۴۲	۴.۲ میانگین آریبی به اضافه و منهای میانگین sd برای سناریوها و حجم نمونه‌های مختلف: SHS ₁ (چپ)، SHS ₂ (وسط) و SSt (راست). SHS ₁ (خط ممتد)، SHS ₂ (خط چین)، SSt (نقطه چین) و SN (نقطه-خط).
۵۲	۵.۲ نرخ‌های پوشش برآوردشده برای هر یک از پارامترهای رگرسیون برای سناریوها و حجم نمونه‌های مختلف: SHS ₁ (چپ)، SHS ₂ (وسط) و SSt (راست). SHS ₁ (خط ممتد)، SHS ₂ (خط چین)، SSt (نقطه چین) و SN (نقطه-خط).
۶۲	۶.۲ نمودارهای جعبه‌ای مقادیر DAIC برای سناریوها و حجم نمونه‌های مختلف: SHS ₁ (چپ)، SHS ₂ (وسط) و SSt (راست).
۷۲	۷.۲ نمودارهای جعبه‌ای مقادیر DBIC برای سناریوها و حجم نمونه‌های مختلف: SHS ₁ (چپ)، SHS ₂ (وسط) و SSt (راست).
۸۲	۸.۲ بافت نگار مانده‌ها و تابع چگالی برآوردشده داده‌های اصوات شناسی برای توزیع‌های مختلف برای عبارت خطا
۹۲	۹.۲ نمودار پراکندگی داده و خطوط رگرسیون منطبق با توزیع‌های مختلف برای عبارت خطا برای مجموعه داده مارتین ماریتا
۱۰۲	۱۰.۲ نمودار توابع چگالی توزیع‌های HS، نرمال و کوشی (سمت راست) و دم‌های سمت راست آنها (سمت چپ)
۱۱۲	۱۱.۲ نمودارهای چندک-چندک (چپ) و کرمی (راست) باقی‌مانده‌های حاصل از سه مدل نرمال (بالا)، t استیودنت (وسط) و HS (پایین) برای یک مجموعه داده شبیه‌سازی‌شده با حجم نمونه ۵۰۰
۱۲۲	۱۲.۲ اثرات خطی و غیرخطی برآوردشده (ممتد) و واقعی (خط چین) پارامترهای مکان و مقیاس مدل نرمال برای حجم نمونه ۵۰۰

۱۳.۲	اثرات خطی و غیرخطی برآوردشده (ممتد) و واقعی (خطچین) پارامترهای مکان
۵۳	و مقیاس مدل، t استیودنت برای حجم نمونه 500
۱۴.۲	اثرات خطی و غیرخطی برآوردشده (ممتد) و واقعی (خطچین) پارامترهای مکان
۵۴	و مقیاس مدل HS برای حجم نمونه 500
۱۵.۲	نمودارهای پراکنش کناری پاسخ و متغیرهای تبیینی برای داده‌های دریاچه‌های
۵۴	اسیدی
۱۶.۲	نمودارهای چندک-چندک و کرمی مدل‌های نرمال (بالا)، HS (وسط) و t استیودنت
۵۶	(پایین) برای مدل‌های (۷.۲) (الف) و (۸.۲) (ب)
۱۷.۲	منحنی‌های اثرات برآوردشده پارامترهای مکان (بالا) و مقیاس (پایین) در مدل
۵۶	نرمال برای مدل (۷.۲)
۱۸.۲	منحنی‌های اثرات برآوردشده پارامترهای مکان (بالا) و مقیاس (پایین) در مدل
۵۷	HS برای مدل (۷.۲)
۱۹.۲	منحنی‌های اثرات برآوردشده پارامترهای مکان (بالا) و مقیاس (پایین) در مدل t
۵۷	استیودنت برای مدل (۷.۲)
۲۰.۲	منحنی‌های توزیع دومتغیره برای مقادیر مختلف پارامترها. $\text{vech}(\Sigma) =$
	$(1, 0/5, 1)'$ (بالا سمت راست)، $\text{vech}(\Sigma) = (1, -0/5, 1)'$ (بالا سمت چپ)،
	$\text{vech}(\Sigma) = (1/5, 0/5, 1/5)'$ (پایین سمت راست)، $\text{vech}(\Sigma) = (1/5, -0/5, 1/5)'$
۵۹	(پایین سمت چپ) و در همه نمودارها بردار پارامتر $\mu = (0, 0)'$ است.
۱.۳	منحنی چگالی (۱.۳) برای برخی مقادیر k
۶۹	۲.۳ منحنی‌های چگالی‌های (۳.۳) (سمت راست) و (۴.۳) (سمت چپ) برای
۷۱	برخی مقادیر k
۳.۳	منحنی‌های چگالی‌های (۶.۳) (سمت راست) و (۷.۳) (سمت چپ) برای
۷۲	برخی مقادیر k
۷۲	۴.۳ منحنی‌های چگالی (۸.۳) برای برخی مقادیر k
۷۴	۵.۳ منحنی‌های چگالی (۹.۳) برای برخی مقادیر پارامترهای شکل
۶.۳	منحنی‌های (۱۲.۳) برای برخی مقادیر پارامترها. $\mu = 0$ و $\sigma = 1$ و $k = 0$
	و $(\nu = 2)$ (بالا سمت راست)، $(\mu = 0$ و $\sigma = 1$ و $k = 1$ و $\nu = 2)$ (بالا سمت
	چپ)، $(\mu = 0$ و $\sigma = 1$ و $a = 0/1$ و $\nu = 2)$ (پایین سمت راست) و $(\mu = 0$
۷۸	و $\sigma = 1$ و $k = -1$ و $\nu = 5)$ (پایین سمت چپ)
۷.۳	منحنی‌های (۱۴.۳) برای برخی مقادیر پارامترها. $(\mu = 0$ و $\sigma = 1$ و $a = 0)$
	(بالا سمت راست)، $(\mu = 0$ و $\sigma = 1$ و $k = 0)$ (بالا سمت چپ)، $(\mu = 0$ و
	$\sigma = 1$ و $a = 0/5)$ (پایین سمت راست) و $(\mu = 0$ و $\sigma = 1$ و $k = 1)$ (پایین
۸۵	سمت چپ).

- ۸.۳ بافت نگار داده‌ها و منحنی توابع چگالی برازش شده مدل‌های رقیب به‌ازای
 ۸۷ مقادیر پارامترهای برآورد شده برای مجموعه داده مثال ۱.۴.۳
- ۹.۳ بافت نگار داده‌ها و منحنی توابع چگالی برازش شده مدل‌های رقیب به‌ازای
 ۸۸ مقادیر پارامترهای برآورد شده برای مجموعه داده مثال ۲.۴.۳
- ۱۰.۳ بافت نگار داده‌ها و منحنی توابع چگالی برازش شده مدل‌های مختلف به‌ازای
 ۸۹ مقادیر پارامترهای برآورد شده برای مجموعه داده مثال ۳.۴.۳
- ۱.۴ منحنی های تابع چگالی BUN برای $\mu = 1$ ، $\sigma = 0.5$ ، $k = 4$ و $a = 0.15$
 (بالا سمت راست)، $\mu = 1$ ، $\sigma = 0.5$ ، $k = -3$ و $a = -2$ (بالا سمت چپ)،
 $\mu = 1$ ، $\sigma = 0.5$ ، $k = -3$ و $a = 2$ (پایین سمت راست) و $\mu = 1$ ، $\sigma = 0.5$ ،
 ۹۳ $k = -3$ و $a = 1$ (پایین سمت چپ)
- ۲.۴ پراکنش و خط‌های رگرسیونی مثالی از رگرسیون مد
 ۹۷
- ۳.۴ شکل تابع زیان (۲.۴)، تابع ψ (۳.۴) و تابع وزن (۴.۴) ساخته شده از هسته
 ۹۸ BUN به ازای $k = -0.3165$ و $c = 1/9078$
- ۴.۴ بافت نگار مانده‌ها و منحنی های تابع چگالی برآورد شده برای داده‌های صوت شناسی
 ۱۰۳ در مدل های رگرسیونی مختلف پیشنهادی
- ۵.۴ بافت نگار و منحنی های تابع چگالی برآورد ده بری داده‌های فشار خون
 ۱۰۵ سیستمولیک بر حسب مدل های پیشنهادی مختلف
- ۶.۴ منحنی تابع چگالی توزیع MBUN برای $p = 2$ به‌ازای مقادیر مختلف پارامترهای
 توزیع. $k = (1, -1)'$ و $a = (0, 0)'$ (بالا سمت راست)، $k = (1, -1)'$ و $a =$
 $(0.5, 0.5)'$ (بالا سمت چپ)، $k = (-2, -2)'$ و $a = (-2, -2)'$ (پایین سمت
 راست)، $k = (-2, -2)'$ و $a = (2, 2)'$ (پایین سمت چپ) و در همه نمودارها
 ۱۱۰ پارامترهای $\mu = (0, 0)'$ و $\text{vech}(\Sigma) = (1, 0.5, 1)'$ هستند.
- ۷.۴ مقادیر AIC (ستون سمت راست) و BIC (ستون سمت چپ) برای دو مدل MBUN
 و MN به ازای حجم‌های نمونه مختلف ۱۱۴
- ۸.۴ برازش توابع چگالی برآورد شده MBUN و MN بر مانده‌های استاندارد شده دو مدل
 ۱۱۶ در دو مثال واقعی. مجموعه داده اول (بالا) و مجموعه داده دوم (پایین)

فهرست جداول

۱.۲	اریبی و MSE برآورد پارامترها برای مدل صحیح و نسبت اربیی و MSE برای	۳۱
۲.۲	اریبی و MSE برآورد پارامترها برای مدل صحیح و نسبت اربیی و MSE برای	۳۱
۳.۲	اریبی و MSE برآورد پارامترها برای مدل صحیح و نسبت اربیی و MSE برای	۳۲
۴.۲	بخش‌های شاخص ارزیابی عینی (رنجش)	۳۷
۵.۲	نتایج MLE تحت مدل‌های مختلف برای داده‌های صوت‌شناسی	۳۸
۶.۲	آزمون KS برای داده‌های صوت‌شناسی	۳۸
۷.۲	نتایج MLE داده الماس برای توزیع‌های مختلف برای عبارت خطا	۴۱
۸.۲	آزمون KS برای داده الماس	۴۱
۹.۲	نتایج MLE داده شرکت مارتین ماریتا برای توزیع‌های مختلف برای عبارت خطا	۴۳
۱۰.۲	آزمون KS برای داده شرکت مارتین ماریتا	۴۳
۱۱.۲	مقادیر MSE پارامترهای خطی و IMSE اثر جمعی برای داده‌های شبیه‌سازی شده	۵۰
۱۲.۲	نتایج برازش مدل‌های (۷.۲) و (۸.۲) بر روی داده‌های دریاچه‌های اسیدی	۵۵
۱۳.۲	اریبی و MSE برآوردگرهای ML و MM تحت مدل $MHS_2(\mu_1, \Sigma_1)$	۶۲
۱۴.۲	اریبی و MSE برآوردگرهای ML و MM تحت مدل $MHS_2(\mu_1, \Sigma_2)$	۶۳
۱۵.۲	اریبی و MSE برآوردگرهای ML و MM تحت مدل $MHS_2(\mu_2, \Sigma_1)$	۶۴
۱۶.۲	اریبی و MSE برآوردگرهای ML و MM تحت مدل $MHS_2(\mu_2, \Sigma_2)$	۶۵
۱۷.۲	نتایج برآورد ML برای مجموعه داده اول	۶۵
۱۸.۲	نتایج برآورد ML برای مجموعه داده دوم	۶۶
۱.۳	نتایج برآورد ML برای مجموعه داده مثال ۱.۴.۳	۸۷
۲.۳	نتایج برآورد ML برای مجموعه داده مثال ۲.۴.۳	۸۸
۳.۳	نتایج برآورد ML برای مجموعه داده مثال ۳.۴.۳	۸۹

۱۰۰	۱.۴	مقادیر MSE کل بردار ضرایب در مثال شبیه‌سازی شده به ازای روش‌های برآورد مختلف
۱۰۱	۲.۴	مقادیر MSE فردی پارامترها در مثال شبیه‌سازی شده به ازای روش‌های برآورد مختلف
۱۰۲	۳.۴	مقادیر اریبی فردی پارامترها در مثال شبیه‌سازی شده به ازای روش‌های برآورد مختلف
۱۰۴	۴.۴	نتایج برآورد ML مدل‌های رگرسیون با جمله خطای BUN برای داده‌های صوت شناسی
۱۰۶	۵.۴	نتایج برآورد ML مدل‌های رگرسیون با جمله خطای BUN برای داده‌های فشار خون سیستولیک
۱۱۳	۶.۴	مقادیر اریبی و MSE برآوردهای پارامترها در مثال شبیه‌سازی
۱۱۳	۷.۴	مقادیر اریبی و MSE برآوردهای ضرایب رگرسیونی در مثال شبیه‌سازی
۱۱۵	۸.۴	برآورد پارامترهای رگرسیونی و معیارهای انتخاب مدل دو مجموعه داده واقعی

فهرست نمادها

$\text{sign}(\cdot)$	تابع علامت
$\text{Cov}(\cdot)$	کواریانس
$\ell^{(2)}$	ماتریس هسیان
$\tilde{\ell}^{(2)}$	ماتریس هسیان
B	ماتریس اطلاع فیشر
$\tilde{B}$	ماتریس اطلاع فیشر
$S(\cdot)$	بردار امتیاز
$C(\cdot)$	تابع مشخصه
$\text{Beta}(a, b)$	توزیع بتا با پارامترهای شکل a و b
$M(\cdot)$	تابع مولد گشتاور
$\text{SHS}_2(\mu, \sigma, \alpha)$	توزیع هایپربولیک سکانت چوله نوع دوم با پارامتر مکان μ ، مقیاس σ و شکل α
$\text{SHS}_1(\mu, \sigma, \nu)$	توزیع هایپربولیک سکانت چوله نوع اول با پارامتر مکان μ ، مقیاس σ و شکل ν
$\ell(\cdot)$	تابع لگاریتم درستنمایی
$g(\cdot)$	تابع چگالی
$f(\cdot)$	تابع چگالی
$G(\cdot)$	تابع توزیع
$F(\cdot)$	تابع توزیع
$\phi(\cdot)$	تابع چگالی نرمال استاندارد
$\Phi(\cdot)$	تابع توزیع نرمال استاندارد
$d_\nu(\cdot)$	تابع چگالی t استیودنت استاندارد با ν درجه آزادی
$D_\nu(\cdot)$	تابع توزیع t استیودنت استاندارد با ν درجه آزادی
γ	توزیع هایپربولیک سکانت تعمیم یافته چوله با پارامتر مکان μ ، مقیاس σ و شکل های t و γ
$\text{SGSH}(\mu, \sigma, t, \gamma)$	
$\text{GSH}(\mu, \sigma, t)$	توزیع هایپربولیک سکانت تعمیم یافته با پارامتر مکان μ ، مقیاس σ و شکل t
$\text{BHS}(\mu, \sigma, a, b)$	توزیع بتا هایپربولیک سکانت با پارامتر مکان μ ، مقیاس σ و شکل های a و b
$\text{HS}(\mu, \sigma)$	توزیع هایپربولیک سکانت با پارامتر مکان μ و مقیاس σ

$St(\mu, \sigma, \lambda, \nu)$	توزیع t استیودنت چوله با پارامتر مکان μ ، مقیاس σ ، شکل λ و درجه آزادی ν
$SN(\mu, \sigma, \lambda)$	توزیع نرمال چوله با پارامتر مکان μ ، مقیاس σ و شکل λ
	توزیع نرمال نامتقارن دومی-یکمدی با پارامتر شکل های α و β ، مکان μ و مقیاس σ
$ABPN(\alpha, \beta, \mu, \sigma)$	
$OLLN(\alpha, \mu, \sigma)$	توزیع آدلگ-لوجستیک نرمال با پارامتر شکل α ، مکان μ و مقیاس σ
$PLD(\nu, \rho, \zeta)$	توزیع لوگ-داگوم توانی با پارامترهای شکل ν ، ρ و ζ
$BN(\mu, \sigma, a, b)$	توزیع بتا نرمال با پارامتر مکان μ ، مقیاس σ و شکل های a و b
$BUN(\mu, \sigma, k, a)$	توزیع نرمال دومی-یکمدی با پارامتر مکان μ ، مقیاس σ و شکل های k و a
$BUST(\mu, \sigma, k, a, \nu)$	توزیع t استیودنت دومی-یکمدی با پارامتر مکان μ ، مقیاس σ ، شکل های k و a و درجه آزادی ν
$BUL(\mu, \sigma, k, a)$	توزیع لاپلاس دومی-یکمدی با پارامتر مکان μ ، مقیاس σ و شکل های k و a
$N(0, 1)$	توزیع با میانگین ۰ و واریانس ۱
$I(\cdot)$	تابع نشانگر مجموعه
$E(\cdot)$	امید ریاضی
$Var(\cdot)$	واریانس
$Bias(\cdot)$	اریبی
$MSE(\cdot)$	میانگین توان های دوم خطا
$\xrightarrow{d}$	همگرایی در توزیع
$\xrightarrow{P}$	همگرایی در احتمال
max	بیشینه
min	کمینه

فهرست کلمات اختصاری

ABPN (Uni-bi-modal Asymmetric Power Normal)

AI (Articulation Index)

ASPL (A-weighted Sound Pressure Level)

BHS (Beta Hyperbolic Secant)

BN (Beta Normal)

BUL (Bimodal–Unimodal Laplace)

BUN (Bimodal–Unimodal Normal)

BUSt (Bimodal–Unimodal Student t)

CSR (Complete Spatial Randomness)

GAM (Generalized Additive Model)

GHS (Generalized Secant Hyperbolic)

GLM (Generalized Linear Model)

GOLL (Generalized Odd Log-Logistic)

HS (Hyperbolic Secant)

IF (Internally Flawless)

KS (Kolmogorov–Smirnov)

LSPL (Linear Sound Pressure Level)

MBUN (Multivariate Bimodal–Unimodal Normal)

MHS (Multivariate Hyperbolic Secant)

ML (Maximum Likelihood)

MSE (Mean Square Error)

OLL (Odd Log-Logistic)

OLLN (Odd Log-Logistic Normal)

OLS (Ordinary Least Squares)

PLD (Power Log-Dagum)

SBP (Systolic Blood Pressure)

sd (Standard Deviation)

SE (Standard Error)

SGHS (Skew Generalized Secant Hyperbolic)

SL (Skew Laplace)

SN (Skew Normal)

SSt (Skew Student t)

St (Student t)

VSI (Very Slightly Imperfect)

VVSI (Very Very Slightly Imperfect)

فصل ۱

پیش‌نیازها و ملزومات

۱.۱ مقدمه

کاربران آمار معمولاً هنگامی از توزیع‌های دم‌سنگین استفاده می‌کنند که استواری با وجود داده‌های پرت مد نظر باشد. در توزیع‌های متقارن، به منظور استفاده از تکنیک‌های برآورد استوار، معمولاً یک روش استاندارد برای به‌کارگیری استنباط آماری وجود دارد. اما، روش‌های استوار برای توزیع‌های نامتقارن موجب برآوردهای با اریبی بالا می‌شود (تاکوچی^۱ و همکاران (۲۰۰۲)). بنابراین وجود ویژگی‌های چوله و دم‌سبکی داده‌های واقعی یک ویژگی جذاب برای مدل‌سازی و استنباط آماری به‌ویژه در زمینه مدل‌های رگرسیونی است. روش‌های متعددی برای به‌دست آوردن توزیع‌های دم‌سنگین چوله در نوشته‌های علمی ارائه شده‌اند که از جمله می‌توان، فرناندز و استیل^۲ (۱۹۹۸)، آزالینی و کاپیتانیو^۳ (۲۰۰۳)، جونز و فادی^۴ (۲۰۰۳)، ما و جنتن^۵ (۲۰۰۴)، فراریه^۶ و استیل (۲۰۰۶) را نام برد.

در طیف گسترده‌ای از زمینه‌های کاربردی مانند اقتصاد، مالی و آب‌شناسی، کلفتی دم توزیع، بسیار سنگین نیست، بلکه نیمه‌سنگین است. علاوه‌براین، می‌توان شرایطی را در نظر

¹Takeuchi

²Fernandez and Steel

³Azzalini and Capitanio

⁴Jones and Faddy

⁵Ma and Genton

⁶Ferreira

گرفت که در آن داده‌ها علاوه بر نیمه‌دم‌سنگینی، چوله نیز باشند. در واقع، حالتی را بیان می‌کنیم که در آن یک توزیع چوله با دم نیمه‌سنگین در رگرسیون خطی برای عبارت خطا، نسبت به چوله دم‌سنگین گزینه بهتری باشد. به همین منظور، در فصل دوم، دو خانواده جدید از توزیع‌های چوله را ارائه می‌کنیم که به نوعی دم نیمه‌سنگین دارند. عملکرد مدل‌های معرفی‌شده را از طریق مطالعات شبیه‌سازی با معیارهای مقایسه AIC و BIC، ارزیابی کرده و با توزیع‌های جایگزین چوله دم‌سنگین مقایسه می‌کنیم. همه مدل‌های جدید ارائه‌شده در فصل دوم توسط بسته shs در محیط نرم‌افزار R، که در صفحه اینترنتی^۷ در دسترس است، اجرا شده‌اند. بحث دیگری که در تحلیل داده‌ها وجود دارد این است که داده‌ها ممکن است دم سبک، چوله و دارای دو مد باشند یا واقعا داده‌ها دارای دم سنگین باشند که دم‌سنگینی آنها ممکن است به علت وجود داده‌های پرت باشد. برای تحلیل این نوع داده‌ها توزیع‌های آماری دائما در حال به‌روز رسانی هستند تا انعطاف‌پذیری بیشتری داشته باشند. در سال‌های اخیر تعمیم‌های زیادی برای توزیع‌های آماری معرفی شده‌اند. این تعمیم‌ها در عمل انعطاف‌پذیری بالایی به توزیع‌های اصلی داده‌اند. برخی از این تعمیم‌ها در زیر فهرست شده‌اند:

الف) آزالینی (۱۹۸۵) خانواده‌ای چوله از توزیع‌های متقارن را به شکل $g(x)F(\lambda x)$ معرفی کرد که در آن $g(\cdot)$ تابع چگالی متقارن حول صفر، $F(\cdot)$ با تابع توزیع پیوسته است که مشتق اول آن حول صفر متقارن است و λ پارامتر چولگی توزیع با مقادیر حقیقی است. به عنوان یک مثال از این خانواده می‌توان توزیع نرمال چوله^۸ (SN) را نام برد.

ب) توزیع‌های نمایی شده^۹ که توسط گوپتا^{۱۰} و همکاران (۱۹۹۸) معرفی شده‌اند دارای توابع چگالی

$$\alpha (1 - G(x))^{\alpha-1} g(x)$$

هستند که در آن $g(\cdot)$ تابع چگالی دلخواه، $G(\cdot)$ تابع توزیع متناظر g و $\alpha > 0$ پارامتر شکل است.

پ) خانواده توزیع‌های بتا- G را یوجین^{۱۱} و همکاران (۲۰۰۲) به شکل

$$\frac{1}{B(a,b)} G(x)^{a-1} (1 - G(x))^{b-1} g(x)$$

معرفی کردند که در آن $g(\cdot)$ تابع چگالی دلخواه، $G(\cdot)$ تابع توزیع متناظر g ، $a > 0$ ، $b > 0$ پارامترهای شکل و $B(a,b)$ تابع بتا است. توزیع بتا نرمال^{۱۲} (BN) مثالی از این خانواده است.

⁷[git://github.com/jamilownuk/shs.git](https://github.com/jamilownuk/shs.git)

⁸Skew Normal

⁹Exponentiated

¹⁰Gupta

¹¹Eugene

¹²Beta Normal

ت) خانواده توزیع‌های چوله-متقارن^{۱۳} با تابع چگالی احتمال $\pi(x)g(x)$ که توسط آرنولد و لین^{۱۴} (۲۰۰۴) معرفی شد که در آن $\pi(\cdot)$ یک تابع اندازه‌پذیر لبگ، با شرط $0 \leq \pi(x) \leq 1$ و $\pi(x) + \pi(-x) = 1$ است و $g(\cdot)$ تابع چگالی متقارن حول صفر است.

ث) خانواده تابع چگالی آد لوگ - لوجستیک^{۱۵} (OLL) توسط گلیتون و لینچ^{۱۶} (۲۰۱۰) تعمیم داده شده است که تابع چگالی آن

$$\frac{\alpha G(x)^{\alpha-1} (1 - G(x))^{\alpha-1}}{[G(x)^{\alpha} + (1 - G(x))^{\alpha}]^2} g(x)$$

است که در آن $\alpha > 0$ پارامتر شکل است. توزیع آد لوگ-لوجستیک نرمال^{۱۷} (OLLN) مثالی از خانواده OLL است که توسط دوواریت^{۱۸} و همکاران (۲۰۱۸) معرفی شد. تعمیم این خانواده یعنی آد لوگ-لوجستیک تعمیم‌یافته^{۱۹} (GOLL) نیز توسط کوردیرو^{۲۰} و همکاران (۲۰۱۶) مورد بررسی قرار گرفت که دارای تابع چگالی

$$\frac{\alpha \theta G(x)^{\alpha\theta-1} (1 - G(x)^{\theta})^{\alpha-1}}{[G(x)^{\alpha\theta} + (1 - G(x)^{\theta})^{\alpha}]^2} g(x)$$

است، که در آن $\alpha > 0$ و $\theta > 0$ پارامترهای شکل توزیع هستند.

اکثر این توزیع‌ها انعطاف‌پذیری بالایی به توزیع‌های اصلی خود داده‌اند اما با معایبی نیز همراه هستند. به عنوان مثال، دارای توابع چگالی و توزیع پیچیده‌ای هستند که به دست آوردن ویژگی‌های توزیع مانند گشتاورها را با مشکل مواجه می‌کنند. اکثر این خانواده‌ها دومی نیستند یا در صورت دومی بودن پارامتری ندارند که فاصله دو مد را کنترل کند. همچنین بعضی از این خانواده‌ها به سختی به حالت‌های چندمتغیره گسترش‌پذیر هستند. برای رفع این معایب، در فصل سوم این رساله، به بیان و اثبات قضیه‌های جدیدی برای گسترش و تعمیم توزیع‌ها، به خانواده‌های جدیدی از توزیع‌ها که چوله و دومی هستند، می‌پردازیم. در واقع این تعمیم‌ها، حالت‌های مختلفی از قبیل دومی یا یک‌مدی بودن، متقارن یا چوله بودن، دو مدی متقارن و دومی چوله بودن را به توزیع‌های یک‌مدی متقارن اضافه می‌کنند. به عبارت دیگر، یک توزیع از این خانواده‌ها همه حالت‌های ذکر شده را در خود دارد. در این نوع تعمیم‌ها یک یا چند پارامتر شکل به توزیع اصلی اضافه می‌شود که برخی از پارامترها چولگی و برخی دومی بودن و فاصله دو مد را کنترل می‌کنند. ویژگی‌های گفته‌شده در خانواده‌های

¹³Skew-Symmetric

¹⁴Arnold and Lin

¹⁵Odd Log-Logistic

¹⁶Gleaton and Lynch

¹⁷Odd Log-Logistic Normal

¹⁸Duarte

¹⁹Generalized Odd Log-Logistic

²⁰Cordeiro

پیشنهادی در طیف گسترده‌ای از تحلیل داده‌ها قابل استفاده هستند که در این ساله، کاربرد یکی از حالات خاص این خانواده‌ها را، در فصل چهارم، در مدل رگرسیون خطی با مانده‌های چوله یا مانده‌های شامل داده پرت یک‌طرفه تشریح می‌کنیم. رفتار مجانبی برآوردگرهای مدل رگرسیونی پیشنهادی را نیز بررسی کرده و نشان می‌دهیم برآوردگرهای ماکسیمم درست‌نمایی حاصل سازگار و دارای توزیع مجانبی نرمال هستند. علاوه بر این، از هسته تابع چگالی توزیع پیشنهادی تابع زیانی می‌سازیم که برای ساخت برآوردگرهای M استوار استفاده می‌شود. رفتار مجانبی برآوردگرهای استوار و توزیع مجانبی آنها را نیز بررسی می‌کنیم. همچنین این توزیع منعطف دومی چوله را به حالت چندمتغیره گسترش داده و کاربرد آن را در تحلیل داده‌های فضایی شرح می‌دهیم. در ادامه مباحث مقدماتی لازم برای ورود به محتوی رساله را مطرح خواهیم کرد.

۲.۱ تعاریف

تعریف ۱.۲.۱. متغیر تصادفی X دارای چگالی دم‌سنگین است اگر تابع چگالی آن یعنی $f(\cdot)$ ، در رابطه زیر صدق کند:

$$\lim_{x \rightarrow \infty} \frac{f(xy)}{f(x)} = y^\gamma \quad \gamma \in \mathbb{R}.$$

تعریف ۲.۲.۱. تابع چگالی $f(x)$ را یک تابع با دم نیمه‌سنگین می‌نامند اگر برای $\eta > 0$ به شکل $f(x) = e^{-\eta x} h(x)$ باشد که $h(x)$ یک تابع چگالی دم‌سنگین است.

در نوشته‌های علمی تعاریف متفاوتی برای توزیع‌های دم سنگین و نیمه‌سنگین وجود دارد که در این رساله تعاریف ارائه‌شده در امی^{۲۱} و همکاران (۲۰۱۸) را در نظر گرفتیم.

تعریف ۳.۲.۱. مد یک توزیع احتمال گسسته مقداری همچون x می‌باشد که بیشترین مقدار تابع جرم احتمال را می‌گیرد. مد برای توزیع احتمال پیوسته مقدار x می‌باشد که تابع چگالی احتمال حداکثر مقدار (ماکسیمم مطلق) است بنابراین مد در قله است.

طبق تعریف رسمی مد، تنها x هایی که به ازای آنها تابع چگالی ماکسیمم مطلق خود را می‌گیرند، به عنوان مد در نظر گرفته می‌شوند اما در سراسر این رساله، x هایی که توابع چگالی به ازای آنها ماکسیمم نسبی خود را نیز بگیرند، به عنوان مد در نظر گرفته می‌شوند.

²¹Omey

۳.۱ مدل رگرسیون خطی

مدل رگرسیون خطی، تقریبی برای مدل‌بندی ارتباط بین متغیر پاسخ Y و یک یا چند متغیر تبیینی^{۲۲} X است. در مدل رگرسیون خطی، داده‌ها با تابع پیشگوی خطی مدل‌بندی و پارامترهای نامعلوم مدل با استفاده از داده‌ها برآورد می‌شوند. یک مدل رگرسیون خطی با p متغیر تبیینی را می‌توان به صورت

$$Y_i = \mathbf{X}_i\beta + \epsilon_i, i = 1, 2, \dots, n \quad (1.1)$$

نوشت، که در آن Y_i متغیر پاسخ مشاهده i -ام، $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$ بردار ضرایب رگرسیونی، $\mathbf{X}_i$ بردار معلوم p بعدی متغیرهای تبیینی مشاهده i -ام و ϵ_i مولفه خطای تصادفی است که معمولاً دارای توزیع نرمال فرض می‌شود. بر طبق روش‌های پارامتری معمول برای برآورد و استنباط مدل‌های رگرسیون خطی، ϵ_i را متغیر تصادفی نرمال فرض کرده و با بیشینه کردن تابع درستنمایی، پارامترهای مدل را برآورد می‌کنند. فرض نرمال بودن ϵ_i زمانی درست است و کارایی دارد که مانده‌های مدل، متقارن و دارای دم‌های سبک باشند. اما در عمل همیشه اینگونه نیست. در بعضی مواقع با داده‌های نامتقارنی مواجه می‌شویم که به شدت چوله یا شامل داده‌های پرت هستند. زمانی که مانده‌ها دارای چولگی هستند، توزیع ϵ_i را به جای نرمال بودن یک توزیع چوله در نظر می‌گیرند یا برای داده‌های شامل داده پرت نیز توزیع ϵ_i را توزیع‌های دم‌سنگین انتخاب می‌کنند. برای برآورد پارامترهای مدل (۱.۱) روش‌های ناپارامتری نیز وجود دارند. از روش‌های ناپارامتری معمول در رگرسیون خطی، روش کمترین توان‌های دوم^{۲۳} (OLS) است. برآوردگر OLS جواب معادله

$$\min \sum_{i=1}^n \epsilon_i^2$$

با مشتق گرفتن نسبت به β_j ‌ها برای $j = 0, 1, \dots, p$ و برابر صفر قرار دادن آن حاصل می‌شود. روش OLS نیز به مانده‌های چوله یا مانده‌های شامل داده پرت بسیار حساس است و در عمل کارایی ندارد. یکی از روش‌های ناپارامتری مناسب برای خطاهای غیرنرمال، رگرسیون استوار^{۲۴} است که در ادامه به اختصار چارچوب کلی آنها را در حالت خطی تشریح می‌کنیم.

۱.۳.۱ رگرسیون استوار

در مدل رگرسیون خطی (۱.۱) اگر داده‌ها شامل داده‌های پرت باشند، روش OLS عملکرد خوبی نخواهد داشت. در چنین موقعیت‌هایی، روش‌های استوار رگرسیون خطی مورد استفاده قرار

^{۲۲}Explanatory variable

^{۲۳}Ordinary least squares

^{۲۴}Robust

می‌گیرند. انواع مختلفی از روش‌های استوار وجود دارند که ما در این قسمت برآوردهای M ^{۲۵} را مورد بررسی قرار می‌دهیم. در روش برآوردهای M به دنبال کمینه کردن تابع زیان $\rho(\cdot)$ نسبت به ضرایب رگرسیون هستیم یعنی

$$\hat{\beta}_M = \operatorname{argmin}_{\beta} \sum_{i=1}^n \rho \left(y_i - \sum_{j=0}^p x_{ij} \beta_j \right). \quad (2.1)$$

همچنین ممکن است معادله هدف به صورت زیر باشد:

$$\min_{\beta} \sum_{i=1}^n \rho(u_i) = \min_{\beta} \sum_{i=1}^n \rho \left(\frac{y_i - \sum_{j=0}^p x_{ij} \beta_j}{\sigma} \right)$$

در این صورت، برای مسئله (۲.۱)، نیاز به برآورد σ به وسیله

$$\hat{\sigma} = \frac{\operatorname{median}|e_i - \operatorname{median}(e_i)|}{0.6745}$$

داریم که در آن $e_i = y_i - \sum_{j=0}^p x_{ij} \beta_j$ (سوسانتی و پراتیوی^{۲۶} (۲۰۱۴)). مشتق اول تابع هدف (۲.۱) نسبت به β برابر

$$\sum_{i=1}^n x_{ij} \psi \left(\frac{y_i - \sum_{j=0}^p x_{ij} \beta_j}{\hat{\sigma}} \right) = 0, \quad j = 0, 1, \dots, p \quad (3.1)$$

است. دراپر و اسمیت^{۲۷} (۱۹۹۸) معادله (۳.۱) را با معرفی تابع وزن حل کردند. در واقع می‌توان نوشت

$$\sum_{i=1}^n w_i x_{ij} \left(y_i - \sum_{j=0}^p x_{ij} \beta_j \right) = 0, \quad j = 1, 2, \dots, p \quad (4.1)$$

که در آن $w_i = w \left(\frac{y_i - \sum_{j=0}^p x_{ij} \beta_j}{\hat{\sigma}} \right)$. معادله (۴.۱) را با روش IRLS حل می‌کنند. بر طبق مارونا^{۲۸} و همکاران (۲۰۱۹)، برآوردهای M با شرایط

۱. $\rho(\epsilon)$ تابع نزولی بر حسب $|\epsilon|$ است

۲. $\rho(0) = 0$

۳. $\rho(\epsilon)$ برای $\epsilon > 0$ صعودی است به طوری که $\rho(\epsilon) < \rho(\infty)$

۴. اگر ρ متناهی باشد، فرض $\rho(\infty) = 1$ نیز برقرار است

²⁵M-estimator

²⁶Susanti and Pratiwi

²⁷Draper and Smith

²⁸Maronna

۵. $\psi(\epsilon) \geq 0, \epsilon \geq 0$ تابع فرد است و برای

سازگار و دارای توزیع مجانبی نرمال با میانگین β و واریانس $\frac{\nu}{n}$ هستند به طوری که

$$\nu = \frac{E_F(\psi^2(\epsilon))}{E_F(\psi'(\epsilon))^2}.$$

که در آن F تابع توزیع دلخواه است. در حالت خاص اگر کارایی مجانبی برآوردگر M نسبت به نرمال مد نظر باشد، F را توزیع نرمال در نظر می‌گیرند.

۴.۱ مدل‌های خطی و جمعی تعمیم‌یافته

فرض کنید مشاهدات y_i ، برای $i = 1, \dots, n$ ، به‌طور مستقل دارای تابع (چگالی) احتمال $f(y_i|\theta^i)$ با میانگین $E(Y_i) = \mu_i$ باشند. بنا بر نلدر و ودربرن^{۲۹} (۱۹۷۲) چارچوب یک مدل خطی تعمیم‌یافته^{۳۰} (GLM) به صورت زیر تعریف می‌شود.

فرض کنید تابع $g(\cdot)$ ، یک تابع پیوند معلوم یکنوا باشد که میانگین متغیر پاسخ، μ ، را به متغیرهای تبیینی مرتبط می‌کند. بنابراین می‌توان نوشت:

$$g(\mu) = X\beta$$

که در آن، $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$ بردار پارامترهای $p+1$ بعدی و X ماتریس طرح معلوم $n \times p$ است. در این مدل، به‌طور کلی پیوند پارامتر μ به متغیرهای تبیینی خطی نیست و یک تابع پیوند مانند $g(\cdot)$ ، ارتباط بین μ و تابع خطی $X\beta$ را مشخص می‌کند. به‌عنوان مثالی از رده مدل‌های خطی تعمیم‌یافته می‌توان رگرسیون لجستیک^{۳۱} را نام برد. بنابر مطالب بیان‌شده رده مدل‌های GLM زمانی کاربرد دارد که رابطه بین $g(\mu)$ و متغیرهای تبیینی خطی باشد. اگر برای برخی از متغیرهای تبیینی این رابطه غیرخطی باشد رده مدل‌های GLM تقریب درستی نخواهد بود. در چنین موقعیتی برای مدل‌بندی صحیح، هیستی و تبشیرانی^{۳۲} (۱۹۹۰)، مدل جمعی تعمیم‌یافته^{۳۳} را به‌صورت زیر تعریف کردند:

$$g(\mu) = X\beta + \sum_{j=1}^J h_j(x_j)$$

که در آن، $\beta = (\beta_1, \beta_2, \dots, \beta_{J'})$ بردار پارامترهای J' بعدی، $X = (X_1, X_2, \dots, X_{J'})$ ماتریس طرح معلوم $n \times J'$ بعدی، $h_j(\cdot)$ تابع ناپارامتری هموار از متغیر تبیینی x_j و x_j ‌ها برای $j = 1, \dots, J'$

²⁹Nelder and Wedderburn

³⁰Generalized linear model

³¹Logistic Regression

³²Hastie and Tibshirani

³³Generalized additive model (GAM)

$1, 2, \dots, J$ و X_i ها برای $i = 1, 2, \dots, J'$ بردارهایی با طول n از مشاهدات متغیرهای تبیینی با اثرات غیرخطی هستند. رده مدل‌های GAM با استفاده از تابع پیوند برای مدل‌بندی میانگین (پارامتر مکان) متغیر پاسخ استفاده می‌شود. رده مدل GAM برای پارامترهای مقیاس و شکل نیز ارایه شده است که در بخش بعدی شرح خواهیم داد.

۱.۴.۱ مدل‌های جمعی تعمیم‌یافته برای پارامترهای مکان، مقیاس و شکل

رده مدل‌های جمعی تعمیم‌یافته برای پارامترهای مکان، مقیاس و شکل^{۳۴} یک چارچوب مدل‌بندی کلی و کاملاً منعطف برای یک متغیر پاسخ فراهم کرده است. یک طیف وسیع از توزیع‌های پیوسته و گسسته، که توزیع‌های چوله و دم‌سنگین را نیز شامل می‌شوند، برای توزیع متغیر پاسخ معرفی و در بسته‌افزار `gamlss` در محیط `R` تعبیه شده‌اند. برای مشاهده فهرست این توزیع‌ها می‌توانید به استاسینوپولوس و همکاران (۲۰۱۷) مراجعه کنید. برخی از این توزیع‌ها تا چهار پارامتر را شامل می‌شوند که می‌توان آنها را با μ, σ, ν و τ نشان داد، که به ترتیب بیانگر پارامترهای مکان (مثل میانگین)، مقیاس (مثل انحراف معیار) و شکل (مثل چولگی و کشیدگی) هستند. در چارچوب مدل‌بندی `GAMLSS` همه پارامترهای توزیع پاسخ را می‌توان به صورت پارامتری یا ترکیبی جمعی از توابع ناپارامتری هموار از متغیرهای تبیینی مدل‌بندی کرد. البته می‌توان آمیخته‌ای از دو بخش پارامتری و ناپارامتری را به صورت مدلی نیمه‌پارامتری نیز در نظر گرفت.

برای شروع فرض کنید مشاهدات y_i ، برای $i = 1, \dots, n$ ، به‌طور مستقل دارای تابع (چگالی) احتمال $f(y_i|\theta^i)$ با بردار پارامترهای $\theta^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i})' = (\mu_i, \sigma_i, \nu_i, \tau_i)'$ باشند. بنا بر ریگبی و استاسینوپولوس (۲۰۰۵) چارچوب یک مدل `GAMLSS` به صورت زیر تعریف می‌شود.

فرض کنید تابع $g_k(\cdot)$ ، برای $k = 1, 2, 3, 4$ ، یک تابع پیوند معلوم یکنوا باشد که پارامتر θ_k را به متغیرهای تبیینی مرتبط می‌کند. بنابراین می‌توان نوشت:

$$g_k(\theta_k) = \mathbf{X}_k \beta_k + \sum_{j=1}^{J_k} h_{jk}(x_{jk}) \quad (5.1)$$

که در آن، $\beta_k = (\beta_{1k}, \beta_{2k}, \dots, \beta_{J'_k})'$ بردار پارامترهای J'_k بعدی، $\mathbf{X}_k = (\mathbf{X}_{1k}, \mathbf{X}_{2k}, \dots, \mathbf{X}_{J'_k})$ ماتریس طرح معلوم $n \times J'_k$ بعدی، $h_{jk}(\cdot)$ تابع ناپارامتری هموار از متغیر تبیینی x_{jk} و x_{jk} ها برای $j = 1, 2, \dots, J_k$ و X_{ik} ها برای $i = 1, 2, \dots, J'_k$ بردارهایی با طول n از مشاهدات متغیرهای تبیینی با اثرات غیرخطی هستند. مدل (۵.۱) یک مدل نیمه‌پارامتری را در چارچوب مدل‌های GAM برای پارامتر θ_k تشکیل می‌دهد.

³⁴Generalized Additive Models for Location, Scale, and Shape (GAMLSS)

در مدل (۵.۱) اثرات هموار متغیرها را می‌توان با کمک رهیافت بسط توابع پایه به صورت یک اثر تصادفی، مشابه $h(x) = Z\gamma$ ، نمایش داد که در آن Z یک ماتریس پایه است که بر اساس مقادیر x ساخته می‌شود و γ برداری از ضرایب (اثرات) تصادفی است که فرض می‌شود دارای توزیع نرمال چندمتغیره است. با این نمایش، واضح است که می‌توان اثرات تصادفی را نیز به پیشگوی (۵.۱)، برای هر پارامتر، به صورت جمعی اضافه کرد. بنابراین مدل‌بندی داده‌های وابسته مثل داده‌های طولی، سری زمانی و فضایی در چارچوب رده مدل‌های GAMLSS نیز امکان‌پذیر است.

با این مقدمه، مدل (۵.۱) را برای هر کدام از پارامترهای مکان، مقیاس و شکل می‌توان به صورت

$$\begin{aligned} g_1(\mu) &= X_1\beta_1 + \sum_{j=1}^{J_1} Z_{j1}\gamma_{j1} \\ g_2(\sigma) &= X_2\beta_2 + \sum_{j=1}^{J_2} Z_{j2}\gamma_{j2} \\ g_3(\nu) &= X_3\beta_3 + \sum_{j=1}^{J_3} Z_{j3}\gamma_{j3} \\ g_4(\tau) &= X_4\beta_4 + \sum_{j=1}^{J_4} Z_{j4}\gamma_{j4} \end{aligned}$$

نمایش داد که در آن μ, σ, ν, τ بردارهای n بعدی هستند و β_k و X_k مشابه قبل تعریف می‌شوند. همچنین ماتریس پایه معلوم $n \times q_{jk}$ بعدی وابسته به متغیرهای تبیینی هستند و γ_{jk} بردار تصادفی q_{jk} بعدی با پذیره $\gamma_{jk} \sim N_{q_{jk}}(0, G_{jk}^{-1})$ است که در آن G_{jk}^{-1} وارون (تعمیم‌یافته) $q_{jk} \times q_{jk}$ بعدی ماتریس متقارن $G_{jk} = G_{jk}(\lambda_{jk})$ است که ممکن است به ابرپارامتر λ_{jk} وابسته باشد. اگر ماتریس G_{jk} نامفرد باشد، بردار ضرایب γ_{jk} دارای یک توزیع ناسره است که چگالی آن متناسب است با $\exp(-\frac{1}{2}\lambda_{jk}\gamma'_{jk}G_{jk}\gamma_{jk})$. به عنوان مثال، می‌توان به اثر تصادفی اتورگرسیو شرطی فضایی^{۳۵} (CAR)، رو و هلد^{۳۶} ((۲۰۰۵)) برای داده‌های فضایی شبکه‌ای اشاره کرد که یک توزیع ناسره دارد.

تعیین ساختارهای مختلف برای ماتریس‌های Z و G انواع مختلفی از اثرات جمعی را نتیجه می‌دهد. این اثرات شامل مولفه‌های تصادفی، هموار متغیرها، سری زمانی و اثرات فضایی می‌شوند. برای مولفه‌های هموار انتخاب‌های مختلفی مانند اسپلاین‌های توانیده^{۳۷} (ایلرز و مارکس^{۳۸} (۱۹۹۶))، اسپلاین‌های درجه سوم^{۳۹}، و چندجمله‌ای‌های موضعی وجود

^{۳۵}Spatial Conditional Autoregressive

^{۳۶}Rue and Held

^{۳۷}Penalized splines

^{۳۸}Eilers and Marx

^{۳۹}Cubic splines

دارند.

چارچوب مدل‌بندی GAMLSS این امکان را فراهم می‌کند که توزیع شرطی متغیر پاسخ (به شرط اثرات تصادفی)، هر توزیع دلخواهی چه از خانواده توزیع‌های نمایی چه خارج از آن باشد. این یک ویژگی خیلی منعطف و خوب محسوب می‌شود. البته به‌عنوان یک محدودیت، توزیع اثرات تصادفی باید نرمال باشد.

۵.۱ اسپلاین‌ها

اصطلاح اسپلاین‌ها، به‌وسیله‌ای برای رسم نمودارها اطلاق می‌شود. هموارکننده‌های بر پایه اسپلاین‌ها، شیوه‌هایی از رگرسیون ناپارامتری هستند. اسپلاین‌ها توابع رگرسیون قطعه‌ای هستند که در نقاطی که گره نامیده می‌شوند، به هم متصل می‌شوند. در واقع، توابع رگرسیونی متفاوتی در داخل ناحیه بین دو گره برازش داده می‌شوند و این قطعات در گره‌ها به یکدیگر متصل می‌شوند. برای برازش مدل به وسیله اسپلاین‌ها تعداد و مکان گره‌ها توسط تحلیل‌گر مشخص می‌شود. در عمل انواع مختلفی از اسپلاین‌ها مانند اسپلاین‌های رگرسیونی^{۴۰}، هموارساز^{۴۱}، مکعبی^{۴۲} و طبیعی^{۴۳} وجود دارند. همچنین ترکیب انواع مختلفی از اسپلاین‌ها به‌عنوان یک نوع اسپلاین مورد استفاده قرار می‌گیرند (ایلرز و مارکس (۱۹۹۶)). مدل

$$y_i = h(x_i) + \epsilon_i$$

را برای $i = 1, \dots, n$ در نظر بگیرید که در آن y_i متغیر پاسخ، x_i متغیر تبیینی و ϵ_i عبارت جمله خطا است. مسئله برآورد تابع هموار $h(\cdot)$ است که با حل

$$\min_h \sum_{i=1}^n (y_i - h(x_i))^2$$

به‌دست می‌آید. همچنین این مسئله با اضافه کردن قسمت تاوان به‌صورت

$$\min_h \sum_{i=1}^n (y_i - h(x_i))^2 + \alpha \int_a^b (h''(x))^2 dx$$

بیان می‌شود که در آن $\alpha > 0$ پارامتر هموار است.

۶.۱ آمار فضایی

داده‌هایی که اغلب با آنها سر و کار داریم، مستقل از یکدیگر نیستند و بین آنها وابستگی وجود دارد که می‌تواند ناشی از موقعیت جغرافیایی آنها باشد. به این نوع داده‌ها که نسبت به

⁴⁰Regression splines

⁴¹Smoothing

⁴²Cubic

⁴³Natural

موقعیت قرار گرفتن شان بین آنها وابستگی وجود دارد، داده‌های فضایی^{۴۴} می‌گویند. به‌عنوان مثال‌هایی از این نوع داده‌ها می‌توان میزان آلودگی هوا در مناطق مختلف شهر تهران، درصد خلوص طلا در نقاط مختلف یک معدن و مقدار آب‌های زیرزمینی در نقاط مختلف یک استان را نام برد.

۱.۶.۱ داده‌های فضایی

از ویژگی‌های داده‌های فضایی، نمایش هر داده با موقعیت جغرافیایی آن در فضای مورد مطالعه است. هرچه قدر داده‌های فضایی جمع‌آوری شده نزدیکتر باشند، وابستگی بیشتری دارند و با افزایش فاصله بین موقعیت داده‌ها، وابستگی آنها کاهش می‌یابد. داده‌های فضایی به سه گروه داده‌های زمین‌آماري^{۴۵}، داده‌های شبکه‌ای^{۴۶} و الگوهای نقطه‌ای^{۴۷} تقسیم می‌شوند.

داده‌های زمین‌آماري

داده‌های زمین‌آماري در موقعیت‌های ثابت و مشخص در ناحیه‌ای پیوسته مشاهده می‌شوند. یعنی ممکن است بین دو موقعیت جغرافیایی، موقعیت‌های دیگری هم وجود داشته باشد. البته متغیر مورد مطالعه ممکن است پیوسته یا گسسته باشد.

داده‌های شبکه‌ای

داده‌های شبکه‌ای مختص مکان‌های ناحیه‌ای هستند، یعنی بین موقعیت‌های مشاهده‌شده امکان یافتن یک موقعیت جدید وجود ندارد. ممکن است این مکان‌ها منظم یا نامنظم باشند. تعداد سرقت‌ها در استان‌های مختلف یک کشور و نرخ رشد اقتصادی هر استان، مثال‌هایی از این نوع داده‌ها هستند.

الگوی نقطه‌ای

داده‌های الگوی نقطه‌ای داده‌هایی هستند که در آنها موقعیت مشاهده‌شده، خود یک متغیر تصادفی است. این داده‌ها عمدتاً به سه دسته به‌طور کامل تصادفی فضایی^{۴۸} (CSR)، منظم^{۴۹} و خوشه‌ای^{۵۰} تقسیم می‌شوند. یک مثال برای این نوع مشاهدات، موقعیت مراکز زلزله است.

⁴⁴Spatial data

⁴⁵Geostatistical data

⁴⁶Lattice data

⁴⁷Point patterns

⁴⁸Complete spatial randomness

⁴⁹Regular

⁵⁰Cluster

با توجه به کاربردهای ذکر شده در این رساله در بخش بعد میدان تصادفی را به اختصار تشریح می‌کنیم.

۲.۶.۱ میدان تصادفی

میدان تصادفی $Z(\cdot)$ را به صورت

$$Z(s) = \mu(s) + \delta(s), \quad s \in D \subset \mathbb{R}^d; d \geq 2 \quad (۶.۱)$$

تجزیه می‌کنیم که در آن $\mu(\cdot)$ تغییرات مقیاس بزرگ^{۵۱} یا روند^{۵۲} و $\delta(\cdot)$ تغییرات مقیاس کوچک^{۵۳} یا فرآیند خطای میدان تصادفی نامیده می‌شود. در تحلیل رگرسیون فضایی با n مشاهده از میدان تصادفی و با حضور متغیرهای تبیینی، می‌توان روند را به صورت تابعی خطی از ضرایب رگرسیونی و متغیرهای تبیینی به صورت

$$\mu(s_i) = x'(s_i)\beta \quad i = 1, \dots, n$$

نشان داد که در آن $\beta = (\beta_0, \dots, \beta_p)'$ و بردار متغیرهای تبیینی p بعدی با درایه‌های $x_j(s_i)$ ، برای $j = 1, \dots, p$ ، است. برای در نظر گرفتن پارامتر عرض از مبدا در مدل، می‌توان قرار داد $x_1(s_i) = 1$.

برای مدل‌بندی داده‌های فضایی اغلب پذیره‌های ساده‌تری بر ساختار وابستگی میدان تصادفی Z اعمال می‌کنند. یکی از آنها مانا بودن میدان تصادفی است. در ادبیات آمار فضایی انواع مانایی وجود دارند که یکی از عمومی‌ترین آنها که معمولاً در عمل استفاده می‌شود، مانای ذاتی است.

تعریف ۱.۶.۱ (مانای ذاتی). میدان تصادفی $Z(\cdot)$ مانای ذاتی است، هرگاه

$$۱) \quad E[Z(s)] = \mu$$

$$۲) \quad Var(Z(s_1) - Z(s_2)) = 2\gamma(s_1 - s_2), \quad s_1, s_2 \in D \subset \mathbb{R}^d.$$

که در آن $\gamma(\cdot)$ تابعی مثبت است و واریوگرام نامیده می‌شود. یعنی میانگین میدان تصادفی ثابت و مستقل از s و واریانس عبارت $(Z(s_1) - Z(s_2))$ فقط تابعی از فاصله موقعیت‌های s_1 و s_2 است.

معمولاً در اغلب کاربردها برای $\delta(s)$ از پذیره نرمال بودن استفاده می‌شود.

⁵¹Large Scale Variation

⁵²Trend

⁵³Small Scale Variation

۷.۱ توابع چگالی استفاده شده در این رساله

در این بخش توابع چگالی توزیع های مختلفی که در این رساله مورد استفاده قرار گرفته اند را فهرست می کنیم.

توزیع هایپربولیک سکانت

تابع چگالی توزیع هایپربولیک سکانت^{۵۴} (HS) به صورت زیر تعریف می شود:

$$f(x) = \frac{1}{\pi\sigma} \frac{2 \exp\left(\frac{x-\mu}{\sigma}\right)}{\exp\left(2\frac{x-\mu}{\sigma}\right) + 1}, \quad x \in \mathbb{R}$$

که در آن μ پارامتر مکان و σ پارامتر مقیاس است. برای نمایش توزیع از نماد $X \sim \text{HS}(\mu, \sigma)$ استفاده می کنیم.

توزیع نرمال چوله

تابع چگالی نرمال چوله^{۵۵} (SN) به صورت زیر تعریف شده است:

$$f(x) = \frac{2}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right) \Phi\left(\alpha \left(\frac{x-\mu}{\sigma}\right)\right), \quad x \in \mathbb{R}$$

که در آن $\phi(\cdot)$ و $\Phi(\cdot)$ به ترتیب توابع چگالی و توزیع نرمال استاندارد هستند. در این تابع چگالی، μ پارامتر مکان، σ پارامتر مقیاس و α پارامتر چولگی هستند (آزالینی (۱۹۸۵)). برای نمایش توزیع از نماد $X \sim \text{SN}(\mu, \sigma, \alpha)$ استفاده می کنیم.

توزیع t استیودنت چوله

تابع چگالی t استیودنت چوله^{۵۶} (SSt) به صورت زیر داده شده است:

$$f(x) = \frac{2}{\sigma} t\left(\left(\frac{x-\mu}{\sigma}\right); \nu\right) T\left(\alpha \left(\frac{x-\mu}{\sigma}\right) \sqrt{\frac{\nu+1}{\nu + \left(\frac{x-\mu}{\sigma}\right)^2}}; \nu+1\right), \quad x \in \mathbb{R}$$

که در آن $t(\cdot; \nu)$ و $T(\cdot; \nu)$ به ترتیب توابع چگالی و توزیع تجمعی برای توزیع t استیودنت^{۵۷} (St) استاندارد با درجه آزادی ν هستند. همچنین μ پارامتر مکان، σ پارامتر مقیاس و α پارامتر چولگی هستند (آرلانووال^{۵۸} و آزالینی (۲۰۱۳)). برای نمایش توزیع از نماد $SSt \sim \text{HS}(\mu, \sigma, \alpha, \nu)$ استفاده می کنیم.

⁵⁴Hyperbolic Secant

⁵⁵Skew Normal

⁵⁶Skew Student t

⁵⁷Student t

⁵⁸Arellano-Valle

توزیع بتا هایپربولیک سکانت

توزیع بتا هایپربولیک سکانت^{۵۹} (BHS)، توسط فیشر و وگان^{۶۰} (۲۰۱۰) پیشنهاد شده است به صورت زیر داده می‌شود:

$$f(x) = \frac{1}{\sigma B(a, b)} G\left(\frac{x-\mu}{\sigma}\right)^{(a-1)} \left(1 - G\left(\frac{x-\mu}{\sigma}\right)\right)^{(b-1)} g\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R}$$

که در آن $g(\cdot)$ و $G(\cdot)$ به ترتیب توابع چگالی و توزیع HS می‌باشند. در این تابع چگالی μ پارامتر مکان، σ پارامتر مقیاس، a و b پارامترهای شکل هستند. برای نمایش توزیع از نماد $X \sim \text{BHS}(\mu, \sigma, a, b)$ استفاده می‌کنیم.

توزیع هایپربولیک سکانت تعمیم‌یافته

تابع چگالی هایپربولیک سکانت تعمیم‌یافته^{۶۱} (GSH) که توسط وگان (۲۰۰۲) مورد بررسی قرار گرفت برابر است با:

$$f(x) = c_1(t) \frac{\exp\left(c_2(t) \left(\frac{x-\mu}{\sigma}\right)\right)}{\exp\left(2c_2(t) \left(\frac{x-\mu}{\sigma}\right)\right) + 2a(t) \exp\left(c_2(t) \left(\frac{x-\mu}{\sigma}\right)\right) + 1}, \quad x \in \mathbb{R}$$

که در آن

$$\begin{cases} a(t) = \cos(t), \quad c_2(t) = \sqrt{(\pi^2 - t^2)/3}, \quad c_1(t) = c_2(t) \sin(t)/t, & -\pi < t \leq 0 \\ a(t) = \cosh(t), \quad c_2(t) = \sqrt{(\pi^2 + t^2)/3}, \quad c_1(t) = c_2(t) \sinh(t)/t, & t > 0. \end{cases}$$

در این توزیع μ پارامتر مکان، σ پارامتر مقیاس و t پارامتر کشیدگی هستند. برای نمایش توزیع از نماد $X \sim \text{GSH}(\mu, \sigma, t)$ استفاده می‌کنیم.

توزیع هایپربولیک سکانت تعمیم‌یافته چوله

تابع چگالی هایپربولیک سکانت تعمیم‌یافته چوله^{۶۲} (SGSH) که توسط فیشر و وگان (۲۰۰۲) معرفی شد به صورت زیر داده می‌شود:

$$f(x) = \frac{2}{\gamma + \frac{1}{\gamma}} \left(\frac{1}{\sigma} g\left(\left(\frac{x-\mu}{\gamma\sigma}\right)\right) I(x \leq \mu) + \frac{1}{\sigma} g\left(\gamma \left(\frac{x-\mu}{\sigma}\right)\right) I(x > \mu) \right), \quad x \in \mathbb{R}$$

که در آن $f(x)$ تابع چگالی GSH و $I(\cdot)$ تابع نشانگر می‌باشند. در این تابع چگالی μ پارامتر مکان، σ پارامتر مقیاس، t پارامتر کشیدگی و λ پارامتر چولگی هستند. برای نمایش توزیع از نماد $X \sim \text{SGHS}(\mu, \sigma, t, \gamma)$ استفاده می‌کنیم.

⁵⁹Beta Hyperbolic Secant

⁶⁰Fischer and Vaughan

⁶¹Generalized Secant Hyperbolic

⁶²Skew Generalized Secant Hyperbolic

توزیع بتا نرمال

توزیع بتا نرمال^{۶۳} (BN)، توسط یوجین و همکاران (۲۰۰۲) معرفی شده است و دارای تابع چگالی به شکل

$$f(x) = \frac{1}{\sigma B(a, b)} \Phi\left(\frac{x-\mu}{\sigma}\right)^{(a-1)} \left(1 - \Phi\left(\frac{x-\mu}{\sigma}\right)\right)^{(b-1)} \phi\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R}$$

است که در آن $\phi(\cdot)$ و $\Phi(\cdot)$ به ترتیب توابع چگالی و توزیع نرمال استاندارد هستند. در این تابع چگالی، μ پارامتر مکان، σ پارامتر مقیاس، a و b پارامترهای شکل هستند. برای نمایش توزیع از نماد $X \sim \text{BN}(\mu, \sigma, a, b)$ استفاده می کنیم.

توزیع لوگ-داگوم توانی

لوگ-داگوم توانی^{۶۴} (PLD) توسط باکوچ^{۶۵} و همکاران (۲۰۱۹) ارائه شده است. تابع چگالی این توزیع عبارتست از

$$f(x) = \zeta \left(\nu + \rho \left| \frac{x-\mu}{\sigma} \right|^{\nu-1} \right) \frac{\exp - \left(\nu \frac{x-\mu}{\sigma} + \text{sign} \left(\frac{x-\mu}{\sigma} \right) \left(\frac{\rho}{\nu} \right) \left| \frac{x-\mu}{\sigma} \right|^{\nu} \right)}{\left(1 + \exp - \left(\nu x + \text{sign} \left(\frac{x-\mu}{\sigma} \right) \left(\frac{\rho}{\nu} \right) \left| \frac{x-\mu}{\sigma} \right|^{\nu} \right) \right)^{\zeta+1}}, \quad x \in \mathbb{R}$$

که در آن μ پارامتر مکان، σ پارامتر مقیاس و ν ، ρ و ζ پارامترهای شکل هستند. برای نمایش توزیع از نماد $X \sim \text{PLD}(\mu, \sigma, \nu, \rho, \zeta)$ و هنگامی که $\mu = 0$ و $\sigma = 1$ از نماد $X \sim \text{PLD}(\nu, \rho, \zeta)$ استفاده می کنیم.

توزیع نرمال توانی نامتقارن دومی-یکمدی

بولفرونی^{۶۶} و همکاران (۲۰۱۸) نرمال توانی نامتقارن دومی-یکمدی^{۶۷} (ABPN) را به صورت زیر معرفی کردند:

$$f(x) = \frac{2\alpha}{\alpha-1} \phi\left(\frac{x-\mu}{\sigma}\right) \left(\Phi\left(\left|\frac{x-\mu}{\sigma}\right|\right)\right)^{\alpha-1} \Phi\left(\beta \frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R}$$

که در آن $\phi(\cdot)$ و $\Phi(\cdot)$ به ترتیب توابع چگالی و توزیع نرمال استاندارد هستند. پارامترهای شکل عبارتند از α و β ، μ پارامتر مکان و σ پارامتر مقیاس هستند. برای نمایش توزیع از نماد $X \sim \text{ABPN}(\alpha, \beta, \mu, \sigma)$ استفاده می کنیم.

⁶³Beta Normal

⁶⁴Power Log-Dagum

⁶⁵Bakouch

⁶⁶ Bolfarine

⁶⁷Uni-bi-modal Asymmetric Power Normal

توزیع آد لوگ-لوجستیک نرمال

تابع چگالی توزیع آد لوگ-لوجستیک نرمال (OLLN) توسط دوواریت و همکاران (۲۰۱۸) به صورت زیر معرفی شد:

$$\frac{\alpha \Phi\left(\frac{x-\mu}{\sigma}\right)^{\alpha-1} \left(1 - \Phi\left(\frac{x-\mu}{\sigma}\right)\right)^{\alpha-1}}{\left[\Phi\left(\frac{x-\mu}{\sigma}\right)^{\alpha} + \left(1 - \Phi\left(\frac{x-\mu}{\sigma}\right)\right)^{\alpha}\right]^2} \phi\left(\frac{x-\mu}{\sigma}\right)$$

که در آن $\phi(\cdot)$ و $\Phi(\cdot)$ به ترتیب توابع چگالی و توزیع نرمال استاندارد هستند. همچنین α پارامتر شکل، μ پارامتر مکان و σ پارامتر مقیاس هستند. برای نمایش این توزیع از نماد $X \sim \text{OLLN}(\alpha, \mu, \sigma)$ استفاده می‌کنیم.

توزیع نیمه St

تابع چگالی توزیع نیمه St به صورت زیر می‌باشد:

$$f(x) = \frac{2\Gamma(\nu+1)}{\sigma\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{\left(\frac{x-\mu}{\sigma}\right)^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R}$$

که در آن ν درجه آزادی، μ پارامتر مکان و σ پارامتر مقیاس هستند.

توزیع نیمه کوشی

تابع چگالی توزیع نیمه کوشی به صورت زیر است:

$$f(x) = \frac{2}{\sigma\pi} \frac{1}{1 + \left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}$$

که در آن μ پارامتر مکان و σ پارامتر مقیاس هستند.

فصل ۲

توزیع هایپربولیک سکانت و تعمیم‌های آن

در این فصل توزیع هایپربولیک سکانت را مورد بررسی قرار می‌دهیم. در بخش اول این فصل دو تعمیم جدید از توزیع هایپربولیک سکانت، با ویژگی‌های جالب، معرفی می‌کنیم. ویژگی نیمه‌دم‌سنگینی داده‌ها را به یک مدل رگرسیون خطی بسط می‌دهیم. به طور خاص، ویژگی‌های مجانبی برآوردهای ML پارامترهای رگرسیون را هنگامی که عبارت خطا دارای توزیع‌های جدید معرفی شده است را بررسی می‌کنیم. مطالعات شبیه‌سازی را با مقایسه برآوردهای ML پارامترهای رگرسیون تحت پذیره‌های مختلفی برای توزیع جمله خطا، انجام می‌دهیم. همچنین سه مثال واقعی ارائه می‌دهیم تا برتری دم‌نیمه‌سنگین بودن عبارت خطا را نسبت به دم‌سنگین بودن نشان دهیم. در بخش دوم این فصل، با فرض اینکه تعداد نقاط پرت نسبت به کل داده‌ها کم باشد، انعطافی را در مدل‌بندی رگرسیونی ایجاد می‌کنیم. بدین منظور یک مدل رگرسیونی جدید نیمه‌پارامتری مکان - مقیاس معرفی می‌کنیم که پاسخ‌ها دارای دم نیمه‌سنگین هستند و بر اساس توزیع متقارن هایپربولیک سکانت می‌باشند. همچنین در بخش سوم این فصل توزیع هایپربولیک سکانت را به حالت چندمتغیره توسعه می‌دهیم تا در داده‌های چندمتغیره نیز بتوانیم از توزیع‌های نیمه‌دم‌سنگین استفاده کنیم. از این فصل مقالات اونق، نزاکتی و باغیشنی (۲۰۱۸)، اونق، باغیشنی و نزاکتی (۱۳۹۹) و اونق، باغیشنی و نزاکتی (۲۰۲۰) استخراج شده‌اند.

۱.۲ تعمیم‌هایی چوله از توزیع هایپربولیک سکانت

درحالی‌که مطالعات قابل توجهی روی تحلیل مقادیر کرانگین و داده‌های پرت با استفاده از توزیع‌های دم سنگین انجام شده‌اند، اطلاعات اندکی درباره رفتار داده‌های دم نیمه‌سنگین وجود دارد؛ به‌ویژه هنگامی که داده‌ها شامل چند داده مشکوک به پرت است. همچنین مطالعات کمی روی توزیع‌های متقارن دم‌نیمه‌سنگین وجود دارند که به عنوان نمونه می‌توان توزیع HS که براساس فیشچر (۲۰۱۳)، از (فیشر (۱۹۲۱)) نشات می‌گیرد را نام برد. وگان (۲۰۰۲) مدل چوله توزیع HS را با نام هایپربولیک سکانت تعمیم‌یافته^۱ (GSH) معرفی کرد و فیشر و وگان (۲۰۰۲) یک جایگزین دیگر با عنوان SGSH اضافه کردند. مدل دیگر چوله از یک توزیع با دم نیمه‌سنگین، توزیع BHS است که توسط فیشر و وگان (۲۰۱۰) پیشنهاد شده است. این مدل‌های چوله دارای چگالی‌های احتمال و توابع توزیع تجمعی پیچیده‌ای هستند. مثلاً در محاسبه گشتاورهای آنها سری نامتناهی وجود دارد که پیچیدگی محاسباتی بالایی دارد. علاوه بر آن، رفتار مجانبی برآوردهای ماکسیمم درست‌نمایی پارامترهای مربوطه در مدل رگرسیون خطی که برای این توزیع‌ها پرکاربرد هستند، مورد بحث واقع نشده‌اند. توزیع‌های پیشنهادی ما در این فصل چندین مزیت دارند: از جمله (۱) فرمول ساده برای توابع چگالی، (۲) سادگی محاسبه گشتاورها، (۳) انطباق بهتر با داده‌ها و (۴) داشتن نمایش تصادفی که بتوان با الگوریتم‌های کارا نمونه تصادفی تولید کرد. برای به‌کارگیری آسان توزیع‌های جدید توسط کاربران آمار بسته‌ای تحت نرم‌افزار پرکاربرد R با نام SHS، توسط نویسندگان این اثر، مهیا شده است که در صفحه اینترنتی زیر^۲ قابل دسترس است.

۱.۱.۲ توزیع‌های پیشنهادی

توزیع HS متقارن و چگالی آن به شکل زنگوله مشابه با توزیع نرمال است اما دم‌های نسبتاً سنگین‌تری دارد. اگر V دارای تابع چگالی نیمه کوشی باشد، آن‌گاه $U = \log(V)$ دارای توزیع HS است (فیشر ۲۰۱۳).

پیشنهاد اول

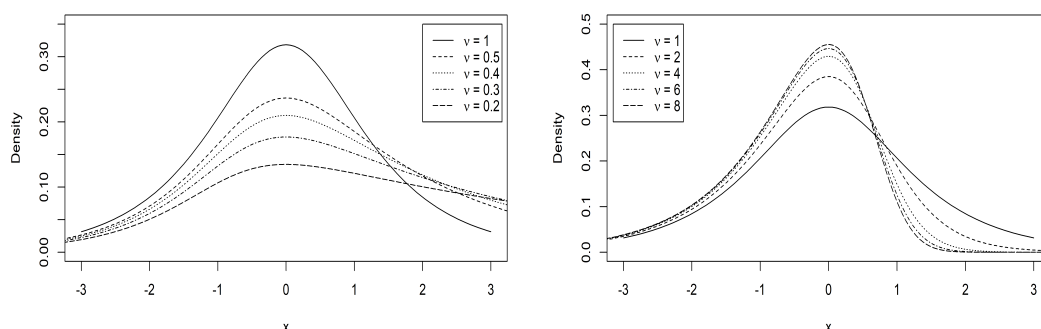
با استفاده از این حقیقت که توزیع کوشی حالت خاصی از توزیع St با درجه آزادی ۱ است، می‌توانیم توزیع HS چوله بسازیم. فرض کنید Y یک توزیع نیمه St با درجه آزادی ν باشد، آن‌گاه متغیر تصادفی $Z = \log(Y)$ دارای تابع چگالی

$$f(z) = c_\nu \exp(z) \left(1 + \frac{1}{\nu} \exp(2z)\right)^{-\frac{\nu+1}{2}}, \quad c_\nu = \frac{2\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \quad (1.2)$$

¹Generalized Hyperbolic Secant

²[git://github.com/jamilownuk/SHS.git](https://github.com/jamilownuk/SHS.git)

است. این توزیع را توزیع هایپربولیک سکانت چوله^۳ نوع اول (SHS_1) می‌نامند. اگر متغیر تصادفی Z از توزیع استاندارد SHS_1 پیروی کند آن را با نماد $SHS_1(0, 1, \nu)$ نشان می‌دهند. با اضافه کردن پارامتر مکان و مقیاس به این متغیر تصادفی به صورت $X = \sigma Z + \mu$ حالت کلی‌تر SHS_1 را خواهیم داشت که در این حالت نماد $SHS_1(\mu, \sigma, \nu)$ را به کار می‌بریم. شکل ۱.۲ چگالی‌های SHS_1 را برای برخی مقادیر ν نشان می‌دهد. با در نظر گرفتن تابع چگالی (۱.۲)



شکل ۱.۲: چگالی‌های SHS_1 برای مقادیر مختلف ν

داده شده برای متغیر تصادفی Z ، به راحتی می‌توان نشان داد هنگامی که $\nu = 1$ ، $f(x)$ به توزیع HS با پارامتر مکانی μ و پارامتر مقیاس σ تبدیل می‌شود. هرگاه $\nu < 1$ یا $\nu > 1$ ، توزیع SHS_1 به ترتیب چوله منفی یا چوله مثبت است.

گزاره ۱.۱.۲. تابع توزیع تجمعی متغیر تصادفی استاندارد SHS_1 به صورت زیر داده می‌شود:

$$F(z) = I_{\frac{\exp(\gamma z)}{\nu + \exp(\gamma z)}} \left(\frac{1}{\gamma}, \frac{\nu}{\gamma} \right)$$

که در آن $I_y(a, b)$ تابع بتا ناکامل است.

□

برهان. برای برهان به پیوست ۱.آ مراجعه کنید.

عبارت زیر را برای تابع مولد گشتاور Z می‌توان بیان کرد.

گزاره ۲.۱.۲. برای توزیع SHS_1 با چگالی (۱.۲) تابع مولد گشتاور و مشخصه به صورت زیر است:

$$M(t) = \frac{\nu^{\frac{t}{\gamma}} \Gamma\left(\frac{1+t}{\gamma}\right) \Gamma\left(\frac{\nu-t}{\gamma}\right)}{\sqrt{\pi} \Gamma\left(\frac{\nu}{\gamma}\right)}.$$

$$C(t) = \frac{\nu^{\frac{it}{\gamma}} \Gamma\left(\frac{1+it}{\gamma}\right) \Gamma\left(\frac{\nu-it}{\gamma}\right)}{\sqrt{\pi} \Gamma\left(\frac{\nu}{\gamma}\right)}.$$

³Skew Hyperbolic Secant

برهان. برای برهان به پیوست ۱.آ مراجعه کنید. □

بنابراین به راحتی می‌توان گشتاورهای توزیع را به صورت زیر نوشت:

$$\begin{aligned} E(Z) &= \frac{1}{\nu} (\log(\nu) + \psi_\nu) \\ E(Z^2) &= (E(Z))^2 + \frac{1}{\nu} \psi'_\nu \\ E(Z^3) &= (E(Z))^3 + \frac{3}{\nu} E(Z) \psi'_\nu + \frac{1}{\nu} \psi''_\nu \\ E(Z^4) &= (E(Z))^4 + \frac{24}{16} (E(Z))^2 \psi'_\nu + \frac{8}{16} E(Z) \psi''_\nu + \frac{3}{16} \psi'''_\nu + \frac{1}{16} \psi''''_\nu \end{aligned}$$

که در آنها $\psi_\nu = \psi\left(\frac{1}{\nu}\right) - \psi\left(\frac{\nu}{\nu}\right)$, $\psi'_\nu = \psi'\left(\frac{1}{\nu}\right) + \psi'\left(\frac{\nu}{\nu}\right)$, $\psi''_\nu = \psi''\left(\frac{1}{\nu}\right) - \psi''\left(\frac{\nu}{\nu}\right)$ و $\psi'''_\nu = \psi'''\left(\frac{1}{\nu}\right) + \psi'''\left(\frac{\nu}{\nu}\right)$ و $\psi''''_\nu = \psi''''\left(\frac{1}{\nu}\right) - \psi''''\left(\frac{\nu}{\nu}\right)$ تابع دای گاما^۴ و $\psi(\cdot)$ و $\psi''(\cdot)$ و $\psi'''(\cdot)$ به ترتیب مشتق اول، دوم و سوم آن هستند. برای $\nu = 1$ می‌توان بررسی کرد که $E(Z) = 0$ و $\text{Var}(Z) = \frac{\pi^2}{6}$ که به ترتیب میانگین و واریانس توزیع HS استاندارد هستند.

پیشنهاد دوم

در این قسمت، تعمیم دیگری از توزیع HS به عنوان توزیع هایپربولیک سکانت چوله^۵ نوع دوم (SHS_2) را معرفی می‌کنیم. برای این توزیع، تابع چگالی احتمال به صورت زیر است:

$$f(x) = \frac{\cos\left(\frac{\pi}{2}\alpha\right)}{\pi\sigma} \frac{e^{\alpha\left(\frac{x-\mu}{\sigma}\right)}}{\cosh\left(\frac{x-\mu}{\sigma}\right)}, \quad -1 < \alpha < 1, \quad \sigma > 0, \quad \mu \in \mathbb{R}, \quad x \in \mathbb{R}. \quad (2.2)$$

اگر متغیر تصادفی X دارای چگالی (۲.۲) باشد، آن را با نماد $\text{SHS}_2(\mu, \sigma, \alpha)$ نشان می‌دهیم. گزاره زیر، تایید می‌کند که SHS_2 دارای ویژگی متقارن نسبت به پارامتر چولگی است.

گزاره ۳.۱.۲. اگر $X \sim \text{SHS}_2(\mu, \sigma, \alpha)$ آن گاه $-X \sim \text{SHS}_2(-\mu, \sigma, -\alpha)$.

برهان. به پیوست ۱.آ مراجعه کنید. □

هنگامی که $\alpha = 0$ است، توزیع تبدیل به توزیع HS می‌شود. هنگامی که $\alpha < 0$ یا $\alpha > 0$ ، توزیع SHS_2 به ترتیب دارای چولگی مثبت یا منفی است. شکل ۲.۲ تابع چگالی SHS_2 را برای برخی مقادیر α نشان می‌دهد.

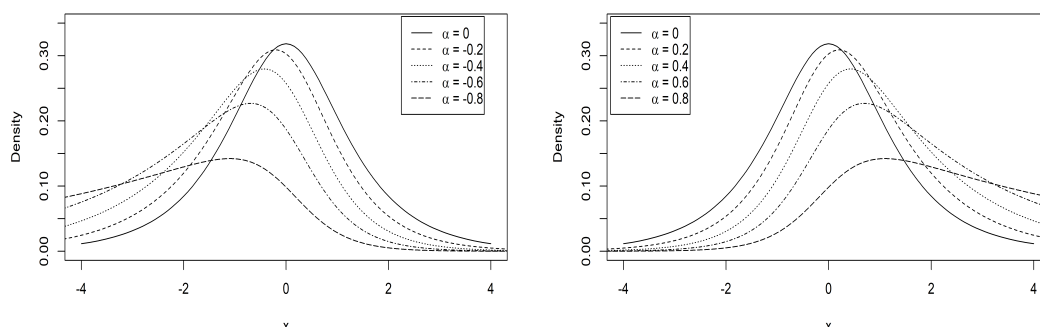
گزاره ۴.۱.۲. تابع توزیع تجمعی $Z \sim \text{SHS}_2(0, 1, \alpha)$ به صورت زیر است:

$$F(z) = I_{\frac{e^{\frac{1}{2}z}}{1+e^{\frac{1}{2}z}}} \left(\frac{1+\alpha}{2}, \frac{1-\alpha}{2} \right).$$

برهان. به پیوست ۱.آ مراجعه کنید. □

^۴Digamma function

^۵Skew hyperbolic secant



شکل ۲.۲: چگالی های SHS₂ برای مقادیر مختلف α

می توان یک نمایش تصادفی ساده برای توزیع SHS₂ ارائه کرد. اگر $Y \sim \text{Beta}\left(\frac{1+\alpha}{2}, \frac{1-\alpha}{2}\right)$ ، آن گاه $X = \frac{1}{\pi} \log\left(\frac{Y}{1-Y}\right)$ از توزیع SHS₂ با چگالی (۲.۲) پیروی می کند. اگر داشته باشیم $Z \sim \text{SHS}_2(0, 1, \alpha)$ ، گزاره های زیر برای توابع مولد گشتاور و مشخصه متغیر تصادفی Z به دست می آیند.

گزاره ۵.۱.۲. توابع مولد گشتاور و مشخصه Z به ترتیب به صورت زیر هستند:

$$C(t) = \frac{\cos\left(\frac{\pi}{2}\alpha\right)}{\cos\left(\frac{\pi}{2}(\alpha + it)\right)}$$

$$M(t) = \frac{\cos\left(\frac{\pi}{2}\alpha\right)}{\cos\left(\frac{\pi}{2}(\alpha + t)\right)}$$

□

برهان. به پیوست ۱.آ مراجعه کنید.

با استفاده از گزاره ۵.۱.۲ می توان نشان داد

$$E(Z) = \frac{\pi}{2} \tan\left(\frac{\pi}{2}\alpha\right)$$

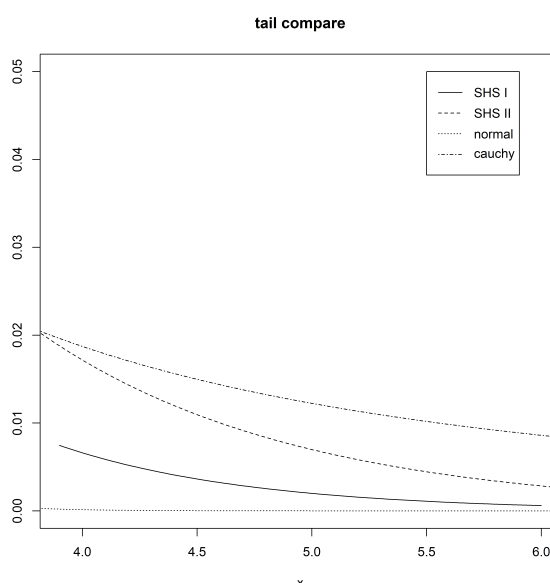
$$E(Z^2) = \frac{\pi^2}{2} \left(1 + 2 \tan^2\left(\frac{\pi}{2}\alpha\right)\right)$$

$$E(Z^3) = \frac{\pi^3}{2} \left(5 \tan\left(\frac{\pi}{2}\alpha\right) + 6 \tan^3\left(\frac{\pi}{2}\alpha\right)\right)$$

$$E(Z^4) = \frac{\pi^4}{2} \left(5 + 28 \tan^2\left(\frac{\pi}{2}\alpha\right) + 24 \tan^4\left(\frac{\pi}{2}\alpha\right)\right).$$

رفتار دم مدل های پیشنهادی

در شکل ۳.۲ دم های چگالی های (۱.۲) و (۲.۲) را با توابع چگالی نرمال و کوشی مقایسه شده اند. توزیع کوشی، از توزیع های دم سنگین و توزیع نرمال، توزیع دم سبک است. به طور شهودی می توانیم بینیم که توزیع های پیشنهادی ما دارای دم های سبک تر از توزیع کوشی و سنگین تر از نرمال هستند و بنابراین دارای دم های نیمه سنگین هستند. این نتیجه را می توان



شکل ۳.۲: مقایسه دم‌ها برای چگالی‌های کوشی (نقطه-خط)، نرمال (نقطه چین)، SHS_1 (خط ممتد) و SHS_2 (خط چین)

با مطالعه نظری رفتار دم‌ها برهان کرد. قضیه زیر نشان می‌دهد که توزیع‌های SHS دارای دم‌های نیمه‌سنگین هستند. این قضیه برای حالت استاندارد ارائه شده است.

قضیه ۱.۱.۲. توزیع‌های SHS_1 و SHS_2 ، هر دو دارای دم نیمه‌سنگین هستند.

□

برهان. به پیوست ۱.آ مراجعه کنید.

۲.۱.۲ مدل‌سازی رگرسیونی

در مدل‌سازی رگرسیون خطی، پذیره معمول نرمال بودن توزیع خطا در مواردی که مشاهدات می‌توانند چوله یا دارای دم‌های سنگین یا نیمه‌سنگین باشند، مناسب نیست. درواقع، انحراف از نرمال بودن خطاها می‌تواند تأثیرات مغایری روی نتایج استنباط داشته باشد (بارتولوچی و ساسیا^۶ (۲۰۰۵)). راهکارهای متعددی برای مقابله با داده‌های غیرنرمال وجود دارند که به عنوان مثال می‌توان رگرسیون SSt (لانگ^۷ و همکاران (۱۹۸۹))، توزیع‌های آمیخته مقیاس نرمال (فراریه^۸ و لاجس (۲۰۱۶))، رگرسیون نرمال چوله (زلی^۹ و همکاران (۲۰۱۶))، آمیخته متناهی از توزیع‌های نرمال (سوفریتی و گالیمبرتی^{۱۰} (۲۰۱۱)) و رگرسیون بیزی با توزیع خطای منعطف (سیلوا^{۱۱} و همکاران (۲۰۲۰)). را نام برد. با این حال، هیچکدام از این

⁶Bartolucci and Scaccia

⁷Lange

⁸Ferreira

⁹Zeller

¹⁰Soffritti and Galimberti

¹¹Silva

مدل ها توزیع هایی با دم نیمه سنگین را برای قسمت خطای رگرسیون خطی برازش نداده اند. همچنین، قابلیت شناسایی^{۱۲} و هزینه محاسباتی روش های مبتنی بر توزیع های آمیخته، که به عنوان راه حل مقابله با داده های مشکوک ارائه شده است، می تواند چالش برانگیز باشد. در این بخش مدلی را در نظر می گیریم که توزیع خطای آن نامتقارن، چوله و دارای دم نیمه سنگین است. فرض کنید مشاهدات $y_i, i = 1, \dots, n$ ، از مدل خطی زیر پیروی کنند:

$$y_i = \beta' x_i + \sigma \epsilon_i, \quad (3.2)$$

که در آن $x_i \in \mathbb{R}^p$ بردار متغیرهای تبیینی، $\beta = (\beta_1, \dots, \beta_p)'$ بردار ضرایب نامعلوم رگرسیون و σ پارامتر مقیاس است. همچنین فرض کنید متغیرهای تصادفی مستقل و هم توزیع $\epsilon_1, \dots, \epsilon_n$ از $SHS_1(\circ, 1, \nu)$ یا $SHS_2(\circ, 1, \alpha)$ پیروی می کنند. در ادامه هر مدل را به طور جداگانه بررسی می کنیم.

مدل رگرسیونی با توزیع خطای SHS_1

اگر عبارت خطا در مدل (۳.۲) از توزیع $SHS_1(\circ, 1, \nu)$ پیروی کند، تابع لگ درستنمایی مدل (۳.۲) به صورت زیر است:

$$\begin{aligned} \ell(\beta, \sigma, \nu) = & n \log \left(2 \Gamma \left(\frac{\nu+1}{2} \right) \right) - \frac{n}{2} \log(\nu\pi) - n \log \left(\Gamma \left(\frac{\nu}{2} \right) \right) - n \log(\sigma) \\ & + \sum_{i=1}^n \left(\frac{y_i - \beta' x_i}{\sigma} \right) - \frac{\nu+1}{2} \sum_{i=1}^n \log \left(1 + \frac{1}{\nu} \exp \left(2 \frac{y_i - \beta' x_i}{\sigma} \right) \right). \end{aligned}$$

ماکسیمم سازی تابع درستنمایی برای یافتن برآوردهای ML، صورت بسته ندارد و این کار با استفاده از روش های عددی مانند روش بهینه سازی برایدن-فلچر-گلدفارب-شانو^{۱۳} که توسط برویدن (۱۹۷۰)، فلچر (۱۹۷۰)، گلدفارب (۱۹۷۰) و شانو (۱۹۷۰) معرفی شده است، انجام می شود. برای استفاده از روش مذکور باید بردار امتیاز پارامترهای مدل را محاسبه کرد. فرض کنید $\theta = (\beta', \sigma, \nu)'$ بردار پارامترهای مدل باشد. مولفه های بردار امتیاز^{۱۴} $S_n =$

¹²Identifiability

¹³Broyden-Fletcher-Goldfarb-Shanno (BFGS)

¹⁴Score vector

به صورت زیر هستند: $\left(\left\{ \frac{\partial \ell(\theta)}{\partial \beta_j} \right\}_{j=1}^k, \frac{\partial \ell(\theta)}{\partial \sigma}, \frac{\partial \ell(\theta)}{\partial \nu} \right)'$

$$\begin{aligned} \frac{\partial \ell(\theta)}{\partial \beta_j} &= \frac{\nu+1}{\sigma} \sum_{i=1}^n x_{ij} \frac{\frac{1}{\nu} \exp\left(\frac{y_i - \beta' x_i}{\sigma}\right)}{1 + \frac{1}{\nu} \exp\left(\frac{y_i - \beta' x_i}{\sigma}\right)} - \frac{1}{\sigma} \sum_{i=1}^n x_{ij}, \quad j = 1, \dots, k \\ \frac{\partial \ell(\theta)}{\partial \sigma} &= \frac{\nu+1}{\sigma} \sum_{i=1}^n \left(\frac{y_i - \beta' x_i}{\sigma} \right) \frac{\frac{1}{\nu} \exp\left(\frac{y_i - \beta' x_i}{\sigma}\right)}{1 + \frac{1}{\nu} \exp\left(\frac{y_i - \beta' x_i}{\sigma}\right)} - \frac{1}{\sigma} \sum_{i=1}^n \left(\frac{y_i - \beta' x_i}{\sigma} \right) - \frac{n}{\sigma} \\ \frac{\partial \ell(\theta)}{\partial \nu} &= \frac{\nu+1}{2\nu} \sum_{i=1}^n \left(\frac{y_i - \beta' x_i}{\sigma} \right) \frac{\frac{1}{\nu} \exp\left(\frac{y_i - \beta' x_i}{\sigma}\right)}{1 + \frac{1}{\nu} \exp\left(\frac{y_i - \beta' x_i}{\sigma}\right)} \\ &\quad - \frac{1}{2} \sum_{i=1}^n \log \left(1 + \frac{1}{\nu} \exp\left(\frac{y_i - \beta' x_i}{\sigma}\right) \right) \\ &\quad - \frac{n}{2} \psi\left(\frac{\nu}{2}\right) + \frac{n}{2} \psi\left(\frac{\nu+1}{2}\right) - \frac{n}{2\nu}. \end{aligned}$$

با درنظر گرفتن تابع درستنمایی و بردار امتیاز، می‌توان برآوردهای ML پارامترها را با روش BFGS به‌دست آورد.

مدل رگرسیون خطی با توزیع خطای SHS₂

با پذیرفتن توزیع SHS₂($\sigma, 1, \alpha$) برای خطاها در مدل (۳.۲)، تابع لگ درستنمایی مدل، با فرض اینکه $\zeta = (\beta', \sigma, \alpha)'$ بردار پارامترها باشد، به صورت زیر به‌دست می‌آید:

$$\begin{aligned} \ell(\zeta) &= n \log \left(\cos \left(\frac{\pi}{2} \alpha \right) \right) - n \log(\sigma) - n \log(\pi) \\ &\quad - \alpha \sum_{i=1}^n \frac{y_i - \beta' x_i}{\sigma} - \sum_{i=1}^n \log \cosh \left(\frac{y_i - \beta' x_i}{\sigma} \right). \end{aligned}$$

مولفه‌های بردار امتیاز $\tilde{S}_n = \left(\left\{ \frac{\partial \ell(\zeta)}{\partial \beta_j} \right\}_{j=1}^k, \frac{\partial \ell(\zeta)}{\partial \sigma}, \frac{\partial \ell(\zeta)}{\partial \alpha} \right)'$ برای تابع لگ درستنمایی $\ell(\zeta)$ به صورت زیر به‌دست می‌آیند:

$$\begin{aligned} \frac{\partial \ell(\zeta)}{\partial \beta_j} &= \alpha \sum_{i=1}^n \frac{x_{ij}}{\sigma} + \sum_{i=1}^n \frac{x_{ij}}{\sigma} \tanh \left(\frac{y_i - \beta' x_i}{\sigma} \right), \quad j = 1, \dots, k \\ \frac{\partial \ell(\zeta)}{\partial \sigma} &= -\frac{n}{\sigma} + \alpha \sum_{i=1}^n \frac{y_i - \beta' x_i}{\sigma^2} + \sum_{i=1}^n \frac{y_i - \beta' x_i}{\sigma^2} \tanh \left(\frac{y_i - \beta' x_i}{\sigma} \right) \\ \frac{\partial \ell(\zeta)}{\partial \alpha} &= -n \frac{\pi}{2} \tan \left(\frac{\pi}{2} \alpha \right) - \sum_{i=1}^n \frac{y_i - \beta' x_i}{\sigma}. \end{aligned}$$

برای محاسبه برآوردهای ML پارامترهای این مدل نیز از روش BFGS استفاده شود.

رفتار مجانبی برآوردهای ML

فرض کنید $X_n = (x_1, \dots, x_n)'$ یک ماتریس طرح $n \times p$ با رتبه p باشد. همچنین فرض کنید Θ فضای پارامتری مدل SHS₁، و Ψ فضای پارامتری مدل SHS₂، زیرمجموعه های بازی از $\mathbb{R}^{p+2}$ باشند. ماتریس هسی $\ell(\theta)$ که مشتق بردار امتیاز نسبت به بردار پارامترها است، به صورت زیر محاسبه می شود:

$$\ell_n^{(\gamma)} = \begin{bmatrix} \frac{\partial^2 \ell(\theta)}{\partial \beta' \partial \beta} & \frac{\partial^2 \ell(\theta)}{\partial \beta' \partial \sigma} & \frac{\partial^2 \ell(\theta)}{\partial \beta' \partial \nu} \\ \frac{\partial^2 \ell(\theta)}{\partial \beta' \partial \sigma} & \frac{\partial^2 \ell(\theta)}{\partial \sigma^2} & \frac{\partial^2 \ell(\theta)}{\partial \sigma \partial \nu} \\ \frac{\partial^2 \ell(\theta)}{\partial \beta' \partial \nu} & \frac{\partial^2 \ell(\theta)}{\partial \sigma \partial \nu} & \frac{\partial^2 \ell(\theta)}{\partial \nu^2} \end{bmatrix}$$

که مولفه های آن به صورت زیر به دست می آیند:

$$\begin{aligned} \frac{\partial^2 \ell(\theta)}{\partial \beta_j \partial \beta_l} &= -\gamma \frac{\nu+1}{\sigma^2} \sum_{i=1}^n x_{ij} x_{il} \left(\frac{\frac{1}{\nu} \exp(\gamma z_i)}{\left(1 + \frac{1}{\nu} \exp(\gamma z_i)\right)^2} \right) \\ \frac{\partial^2 \ell(\theta)}{\partial \beta_j \partial \sigma} &= \frac{1}{\sigma^2} \sum_{i=1}^n x_{ij} - \frac{\nu+1}{\sigma^2} \sum_{i=1}^n x_{ij} \left(\frac{(\gamma z_i) \frac{1}{\nu} \exp(\gamma z_i)}{\left(1 + \frac{1}{\nu} \exp(\gamma z_i)\right)^2} + \frac{\frac{1}{\nu} \exp(\gamma z_i)}{1 + \frac{1}{\nu} \exp(\gamma z_i)} \right) \\ \frac{\partial^2 \ell(\theta)}{\partial \beta_j \partial \nu} &= \frac{1}{\sigma} \sum_{i=1}^n x_{ij} \left(\frac{\frac{1}{\nu} \exp(\gamma z_i)}{1 + \frac{1}{\nu} \exp(\gamma z_i)} \right) - \frac{\nu+1}{\sigma \nu} \sum_{i=1}^n \left(\frac{\frac{1}{\nu} \exp(\gamma z_i)}{\left(1 + \frac{1}{\nu} \exp(\gamma z_i)\right)^2} \right) \\ \frac{\partial^2 \ell(\theta)}{\partial \sigma^2} &= -\frac{\nu+1}{\sigma^2} \sum_{i=1}^n \left(\frac{(\gamma z_i)^2 \frac{1}{\nu} \exp(\gamma z_i)}{\left(1 + \frac{1}{\nu} \exp(\gamma z_i)\right)^2} + \frac{(\gamma z_i) \frac{1}{\nu} \exp(\gamma z_i)}{1 + \frac{1}{\nu} \exp(\gamma z_i)} \right) \\ &\quad + \frac{1}{\sigma^2} \sum_{i=1}^n (\gamma z_i) + \frac{n}{\sigma^2} \\ \frac{\partial^2 \ell(\theta)}{\partial \sigma \partial \nu} &= \frac{1}{\gamma \sigma} \sum_{i=1}^n \left(\frac{(\gamma z_i) \frac{1}{\nu} \exp(\gamma z_i)}{1 + \frac{1}{\nu} \exp(\gamma z_i)} \right) - \frac{\nu+1}{\gamma \sigma \nu} \sum_{i=1}^n \left(\frac{(\gamma z_i) \frac{1}{\nu} \exp(\gamma z_i)}{\left(1 + \frac{1}{\nu} \exp(\gamma z_i)\right)^2} \right) \\ \frac{\partial^2 \ell(\theta)}{\partial \nu^2} &= -\frac{\nu+1}{\gamma \nu^2} \sum_{i=1}^n \left(\frac{\frac{1}{\nu} \exp(\gamma z_i)}{\left(1 + \frac{1}{\nu} \exp(\gamma z_i)\right)^2} \right) - \frac{1}{\gamma \nu^2} \sum_{i=1}^n \left(\frac{\frac{1}{\nu} \exp(\gamma z_i)}{1 + \frac{1}{\nu} \exp(\gamma z_i)} \right) \\ &\quad + \frac{1}{\gamma \nu} \sum_{i=1}^n \left(\frac{\frac{1}{\nu} \exp(\gamma z_i)}{1 + \frac{1}{\nu} \exp(\gamma z_i)} \right) - \frac{n}{\gamma} \left(\psi' \left(\frac{\nu}{\gamma} \right) - \psi' \left(\frac{\nu+1}{\gamma} \right) \right) + \frac{n}{\gamma \nu^2} \end{aligned}$$

که در آن $z_i = \frac{y_i - \beta' x_i}{\sigma}$. به‌طور مشابه، ماتریس هسی $\ell(\zeta)$ را می‌توان به شکل زیر نوشت:

$$\tilde{\ell}_n^{(\gamma)} = \begin{bmatrix} \frac{\partial^2 \ell(\zeta)}{\partial \beta' \partial \beta} & \frac{\partial^2 \ell(\zeta)}{\partial \beta' \partial \sigma} & \frac{\partial^2 \ell(\zeta)}{\partial \beta' \partial \alpha} \\ \frac{\partial^2 \ell(\zeta)}{\partial \beta' \partial \sigma} & \frac{\partial^2 \ell(\zeta)}{\partial \sigma^2} & \frac{\partial^2 \ell(\zeta)}{\partial \sigma \partial \alpha} \\ \frac{\partial^2 \ell(\zeta)}{\partial \beta' \partial \alpha} & \frac{\partial^2 \ell(\zeta)}{\partial \sigma \partial \alpha} & \frac{\partial^2 \ell(\zeta)}{\partial \alpha^2} \end{bmatrix}.$$

که مولفه‌های آن برابر هستند با

$$\begin{aligned} \frac{\partial^2 \ell(\zeta)}{\partial \beta_j \partial \beta_l} &= - \sum_{i=1}^n \frac{x_{ij} x_{il}}{\sigma^2} \frac{1}{\cosh^2 \left(\frac{y_i - \beta' x_i}{\sigma} \right)} \\ \frac{\partial^2 \ell(\zeta)}{\partial \beta_j \partial \sigma} &= -\alpha \sum_{i=1}^n \frac{x_{ij}}{\sigma^2} - \sum_{i=1}^n \frac{x_{ij}}{\sigma^2} \tanh \left(\frac{y_i - \beta' x_i}{\sigma} \right) \\ &\quad - \sum_{i=1}^n \frac{x_{ij}}{\sigma^2} \frac{\frac{y_i - \beta' x_i}{\sigma}}{\cosh^2 \left(\frac{y_i - \beta' x_i}{\sigma} \right)} \\ \frac{\partial^2 \ell(\zeta)}{\partial \beta_j \partial \alpha} &= \sum_{i=1}^n \frac{x_{ij}}{\sigma} \\ \frac{\partial^2 \ell(\zeta)}{\partial \sigma^2} &= -\frac{n}{\sigma^2} - 2 \sum_{i=1}^n \frac{y_i - \beta' x_i}{\sigma^2} \tanh \left(\frac{y_i - \beta' x_i}{\sigma} \right) \\ &\quad - \sum_{i=1}^n \frac{\left(\frac{y_i - \beta' x_i}{\sigma} \right)^2}{\cosh^2 \left(\frac{y_i - \beta' x_i}{\sigma} \right)} - 2\alpha \sum_{i=1}^n \frac{y_i - \beta' x_i}{\sigma^2} \\ \frac{\partial^2 \ell(\zeta)}{\partial \sigma \partial \alpha} &= \sum_{i=1}^n \frac{y_i - \beta' x_i}{\sigma^2} \\ \frac{\partial^2 \ell(\zeta)}{\partial \alpha^2} &= -n \frac{\pi^2}{6} \left(1 + \tanh^2 \left(\frac{\pi}{2} \alpha \right) \right). \end{aligned}$$

با فرض توزیع SHS_1 برای عبارت خطا، ماتریس اطلاع فشر به‌صورت زیر به‌دست می‌آید

$$B_n = -E(\ell_n^{(\gamma)}) = \begin{bmatrix} B_n(\beta) & B_n(\beta, \sigma) & B_n(\beta, \nu) \\ B_n(\beta, \sigma) & B_n(\sigma) & B_n(\sigma, \nu) \\ B_n(\beta, \nu) & B_n(\sigma, \nu) & B_n(\nu) \end{bmatrix}$$

که مولفه های آن عبارتند از

$$\begin{aligned}
 B_n(\beta) &= \frac{c'_\nu}{\sigma^\nu} \text{Beta} \left(\frac{\nu}{\nu}, \frac{\nu+\nu}{\nu} \right) X'_n X_n, \\
 B_n(\beta_j, \nu) &= n \frac{c'_\nu}{\nu \sigma^\nu} \text{Beta} \left(\frac{\nu}{\nu}, \frac{\nu+\nu}{\nu} \right) - \frac{c'_\nu}{\nu \sigma(\nu+1)} \text{Beta} \left(\frac{\nu}{\nu}, \frac{\nu+1}{\nu} \right) \sum_{i=1}^n x_{ij}, \\
 B_n(\beta_j, \sigma) &= \frac{c'_\nu}{\sigma^\nu} \left(\text{Beta} \left(\frac{\nu}{\nu}, \frac{\nu+\nu}{\nu} \right) (\log(\nu) + \psi_{\nu\nu}) + \frac{1}{\nu} \text{Beta} \left(\frac{\nu}{\nu}, \frac{\nu+1}{\nu} \right) \right) \sum_{i=1}^n x_{ij} \\
 &\quad - \frac{1}{\sigma^\nu} \sum_{i=1}^n x_{ij}, \\
 B_n(\sigma) &= n \frac{c'_\nu}{\sigma^\nu} \text{Beta} \left(\frac{\nu}{\nu}, \frac{\nu+1}{\nu} \right) (\log(\nu) + (\psi_{\nu 1})) - n \frac{1}{\sigma^\nu} (\log(\nu) + \psi_{\nu 5}) - \frac{n}{\sigma^\nu} \\
 &\quad + \text{Beta} \left(\frac{\nu}{\nu}, \frac{\nu+\nu}{\nu} \right) \left((\log(\nu))^2 + \nu \log(\nu) \psi_{\nu\nu} + \psi_{\nu\nu}^2 + \psi' \left(\frac{\nu}{\nu} \right) + \psi_{\nu\nu}^2 \right) \\
 &\quad + \text{Beta} \left(\frac{\nu}{\nu}, \frac{\nu+\nu}{\nu} \right) \left(\psi' \left(\frac{\nu+\nu}{\nu} \right) - \nu \psi_{\nu\nu} \psi_{\nu\nu} \right), \\
 B_n(\sigma, \nu) &= \frac{\sigma c'_\nu}{\nu^\nu} \text{Beta} \left(\frac{\nu}{\nu}, \frac{\nu+\nu}{\nu} \right) (\log(\nu) + \psi_{\nu\nu}) \\
 &\quad - \frac{\sigma c'_\nu}{\nu^\nu (\nu+1)} \text{Beta} \left(\frac{\nu}{\nu}, \frac{\nu+1}{\nu} \right) (\log(\nu) + \psi_{\nu 1}), \\
 B_n(\nu) &= \frac{n}{\nu} \left(\frac{c'_\nu}{\nu^\nu} \text{Beta} \left(\frac{\nu}{\nu}, \frac{\nu+\nu}{\nu} \right) + \frac{1-\nu}{\nu^\nu (\nu+1)} \text{Beta} \left(\frac{\nu}{\nu}, \frac{\nu+1}{\nu} \right) + \psi_{\nu\nu} - \frac{\nu}{\nu^\nu} \right),
 \end{aligned}$$

که در آن $c'_\nu = \frac{(\nu+1)c_\nu\sqrt{\nu}}{\nu}$ ، $\psi_{\nu\nu} = \psi \left(\frac{\nu}{\nu} \right) - \psi \left(\frac{\nu+\nu}{\nu} \right)$ ، $\psi_{\nu 1} = \psi \left(\frac{\nu}{\nu} \right) - \psi \left(\frac{\nu+1}{\nu} \right)$ ، $\psi_{\nu 5} = \psi \left(\frac{1}{\nu} \right) - \psi \left(\frac{\nu}{\nu} \right)$ ، $\psi_{\nu\nu}^2 = \psi \left(\frac{\nu+\nu}{\nu} \right) - \psi \left(\frac{\nu+6}{\nu} \right)$ ، $\psi \left(\frac{\nu}{\nu} \right) - \psi \left(\frac{\nu+6}{\nu} \right)$ و $B_n(\beta_j, \nu)$ و $B_n(\beta_j, \sigma)$ به ترتیب j -امین مولفه $B_n(\beta, \nu)$ و $B_n(\beta, \sigma)$ هستند.

با فرض توزیع SHS₂ برای عبارت خطا داریم

$$\tilde{B}_n = -E(\tilde{\ell}_n^{(\nu)}) = \begin{bmatrix} \tilde{B}_n(\beta) & \tilde{B}_n(\beta, \sigma) & \tilde{B}_n(\beta, \alpha) \\ \tilde{B}_n(\beta, \sigma) & \tilde{B}_n(\sigma) & \tilde{B}_n(\sigma, \alpha) \\ \tilde{B}_n(\beta, \alpha) & \tilde{B}_n(\sigma, \alpha) & \tilde{B}_n(\alpha) \end{bmatrix}$$

که مولفه‌های آن عبارتند از

$$\begin{aligned}\tilde{B}_n(\beta)_{(j,l)} &= \frac{\cos\left(\frac{\pi}{4}\alpha\right)}{\pi} \text{Beta}\left(\frac{\alpha+3}{2}, \frac{3-\alpha}{2}\right) \sum_{i=1}^n \frac{x_{ij}x_{il}}{\sigma^2}, & \tilde{B}_n(\beta_j, \alpha) &= -\sum_{i=1}^n \frac{x_{ij}}{\sigma}, \\ \tilde{B}_n(\beta_j, \sigma) &= \frac{\cos\left(\frac{\pi}{4}\alpha\right)}{\pi} \left(2 \text{Beta}\left(\frac{\alpha+3}{2}, \frac{1-\alpha}{2}\right) - \text{Beta}\left(\frac{\alpha+1}{2}, \frac{1-\alpha}{2}\right)\right) \sum_{i=1}^n \frac{x_{ij}}{\sigma^2} \\ &\quad + 2 \frac{\cos\left(\frac{\pi}{4}\alpha\right)}{\pi} \text{Beta}\left(\frac{\alpha+3}{2}, \frac{3-\alpha}{2}\right) \left(\psi\left(\frac{\alpha+3}{2}\right) - \psi\left(\frac{3-\alpha}{2}\right)\right) \sum_{i=1}^n \frac{x_{ij}}{\sigma^2} \\ &\quad + \alpha \sum_{i=1}^n \frac{x_{ij}}{\sigma^2}, \\ \tilde{B}_n(\sigma) &= +\frac{n}{\sigma^2} + \frac{\cos\left(\frac{\pi}{4}\alpha\right)}{\pi} \left(\text{Beta}\left(\frac{\alpha+3}{2}, \frac{1-\alpha}{2}\right) \left(\psi\left(\frac{\alpha+3}{2}\right) - \psi\left(\frac{1-\alpha}{2}\right)\right) \right. \\ &\quad \left. - \text{Beta}\left(\frac{\alpha+1}{2}, \frac{1-\alpha}{2}\right) \left(\psi\left(\frac{\alpha+1}{2}\right) - \psi\left(\frac{1-\alpha}{2}\right)\right) \right) + n \frac{\alpha\pi}{\sigma^2} \tan\left(\frac{\pi}{4}\alpha\right) \\ &\quad + 2 \frac{\cos\left(\frac{\pi}{4}\alpha\right)}{\pi} \text{Beta}\left(\frac{\alpha+3}{2}, \frac{3-\alpha}{2}\right) \left(\psi\left(\frac{\alpha+3}{2}\right) - \psi\left(\frac{3-\alpha}{2}\right)\right), \\ \tilde{B}_n(\sigma, \alpha) &= n \frac{\pi}{\sigma} \tan\left(\frac{\pi}{4}\alpha\right), & \tilde{B}_n(\alpha) &= n \frac{\pi^2}{4} \left(1 + \tan^2\left(\frac{\pi}{4}\alpha\right)\right)\end{aligned}$$

که در آن $\tilde{B}_n(\beta)_{(j,l)}$ -امین مولفه $\tilde{B}_n(\beta)$ است.

شرایط مورد نیاز برای برهان نتایج جانبی عبارتند از

۱. بردار واقعی پارامترها در مدل SHS_1 ، یک نقطه درونی از Θ است.
۲. بردار واقعی پارامترها در مدل SHS_2 ، یک نقطه درونی از Ψ است.
۳. برای $j = 1, \dots, k$ ، $n^{-1} \sum_{i=1}^n x_{ij} \rightarrow c_j$ ، وقتی که $n \rightarrow \infty$ ، که در آن c_j یک عدد حقیقی متناهی است.

۴. برای $n^{-1} X_n' X_n$ ، $n \rightarrow \infty$ به یک ماتریس متناهی و مثبت معین C همگرا است.

۵. برای $j = 1, \dots, k$ ، وقتی $n \rightarrow \infty$ ، $n^{-2} \sum_{i=1}^n x_{ij}^2 x_{ik}^2 \rightarrow 0$.

دو قضیه زیر سازگاری و نرمال بودن جانبی برآوردهای ML را در دو مدل پیشنهادی بیان می‌کنند.

قضیه ۲.۱.۲. تحت شرایط ۱، ۳، ۴ و ۵، برآوردهای ML مدل رگرسیون خطی (۳.۲) با در نظر گرفتن توزیع SHS_1 برای عبارت خطا، $\hat{\theta}$ ، دارای سازگاری ضعیف است و

$$\sqrt{n}(\hat{\theta} - \theta_0) \sim N(0, B_n^{-1}).$$

□

برهان. به پیوست ۱.۰ مراجعه کنید.

قضیه ۳.۱.۲. تحت شرایط ۲، ۳، ۴ و ۵، برآوردگرهای ML مدل رگرسیون خطی (۳.۲) با در نظر گرفتن توزیع SHS₂ برای عبارت خطا، ξ ، دارای سازگاری ضعیف است و

$$\sqrt{n}(\hat{\xi} - \xi_0) \sim N(0, \tilde{B}_n^{-1}).$$

به پیوست ۱.آ مراجعه کنید.

۳.۱.۲ مطالعات شبیه‌سازی

در این بخش برای ارزیابی و مقایسه عملکرد دو توزیع پیشنهادی نسبت به هم و نسبت به توزیع‌های رقیب SSt به عنوان توزیع چوله دم‌سنگین و SN به عنوان توزیع چوله دم‌سبک، یک مطالعه شبیه‌سازی انجام دادیم. متغیر پاسخ را از مدل خطی زیر شبیه‌سازی کردیم:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \sigma \epsilon_i, \quad i = 1, \dots, n.$$

روند مطالعه شبیه‌سازی به صورت زیر است:

- اندازه نمونه را برابر ۱۰۰، ۲۰۰ و ۵۰۰ در نظر گرفتیم.
 - متغیرهای تبیینی x_1 و x_2 از توزیع نرمال استاندارد تولید شدند.
 - مقادیر واقعی ضرایب رگرسیون $(\beta_1, \beta_2) = (1, 2)$ در نظر گرفته شدند. مقدار واقعی پارامتر مقیاس نیز برابر $\sigma = 1$ انتخاب شد.
 - هر دو متغیر تبیینی و متغیر پاسخ مرکزی شدند.
- سناریوهای تولید خطا در این مطالعه به صورت زیر تنظیم شدند:

$$1. \quad \epsilon_i \sim \text{SHS}_1(0, 1, \nu = 4).$$

$$2. \quad \epsilon_i \sim \text{SHS}_2(0, 1, \alpha = 0.5).$$

$$3. \quad \epsilon_i \sim \text{SSt}(0, 1, \alpha = 2, \nu = 4).$$

در دو مورد اول، توزیع خطا مربوط به توزیع‌های SHS در نظر گرفته شده‌است. بنابراین مشاهدات در داده‌های شبیه‌سازی دارای دم نیمه‌سنگین هستند. انتخاب $\nu = 4$ و $\alpha = 0.5$ برای SHS₁ و SHS₂ برای داشتن چولگی مثبت و منفی بوده است. هم‌چنین برای ارزیابی مدل زمانی که خطاها از توزیع دم‌سنگین پیروی کند، سومین حالت نمونه را از توزیع SSt تولید کردیم. در هر سناریو و برای هر کدام از حجم نمونه‌ها ۱۰۰۰ تکرار انجام شده است. برای بررسی و تحلیل نتایج در این مطالعه اریبی^{۱۵} و میانگین توان‌های دوم خطا^{۱۶} (MSE) برآوردهای ML به دست آمده را محاسبه کردیم.

¹⁵bias

¹⁶Mean square error

برای نمایش و خواندن بهتر نتایج از جدول، نسبت اریبی ها و MSE ها را محاسبه و گزارش کردیم. بدین صورت که در هر سناریو اریبی و MSE مدل صحیح برآورد شده را عیناً و برای سایر مدل ها نسبت $\frac{\text{اریبی مدل رقیب}}{\text{اریبی مدل صحیح}}$ و $\frac{\text{MSE مدل رقیب}}{\text{MSE مدل صحیح}}$ را گزارش کردیم. نتایج سناریوی اول در جدول ۱.۲، سناریوی دوم در جدول ۲.۲ و سناریوی سوم در جدول ۳.۲ آمده‌اند.

جدول ۱.۲: اریبی و MSE برآورد پارامترها برای مدل صحیح و نسبت اریبی و MSE برای مدل های رقیب در سناریوی اول

		۵۰۰		۲۰۰		۱۰۰		n	مدل
		$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$		
SHS ₁	Bias	-۰/۰۰۰۰۰	-۰/۰۰۰۳۹	-۰/۰۰۰۳۸	-۰/۰۰۰۷۶	-۰/۰۰۰۱۲	۰/۰۰۰۰۲	۷۰	
	MSE	۰/۰۰۰۲۳	۰/۰۰۰۲۵	۰/۰۰۰۵۸	۰/۰۰۰۵۲	۰/۰۰۱۱۷	۰/۰۰۱۱۷	۷۰	
SHS ₂	Bias	۴/۸۵۷۳	۱/۰۷۶۹	۱/۰۲۹۸	۱/۱۰۴۱	۱/۱۹۲۶	۰/۲۰۳۱	۷۰	
	MSE	۱/۰۹۱۴	۱/۰۸۰۳	۱/۰۸۴۳	۱/۰۷۹۸	۱/۰۶۵۹	۱/۰۶۶۰	۷۰	
SSt	Bias	۸/۹۰۱۸	۰/۹۵۲۷	۱/۱۲۴۰	۰/۹۸۶۲	۱/۲۲۰۸	۱/۸۹۳۰	۷۰	
	MSE	۱/۰۲۵۰	۱/۰۰۳۵	۱/۰۳۴۸	۱/۰۲۰۲	۱/۰۲۳۶	۱/۰۳۳۴	۷۰	
SN	Bias	۴۴/۹۹۵۴	۰/۸۵۳۳	۱/۹۲۱۹	۱/۲۳۸۴	۰/۶۲۹۰	-۰/۳۱۸۵	۷۰	
	MSE	۱/۲۵۵۹	۱/۱۷۰۹	۱/۲۲۱۵	۱/۲۱۴۶	۱/۲۰۷۳	۱/۲۰۷۶	۷۰	

جدول ۲.۲: اریبی و MSE برآورد پارامترها برای مدل صحیح و نسبت اریبی و MSE برای مدل های رقیب در سناریوی دوم

		۵۰۰		۲۰۰		۱۰۰		n	مدل
		$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$		
SHS ₁	Bias	۱/۰۱۶۶	۰/۸۳۶۳	۱/۰۸۴۰	۰/۹۹۲۵	۰/۹۴۴۲	۰/۹۴۱۹	۷۰	
	MSE	۰/۹۶۱۲	۰/۹۶۱۷	۰/۹۶۶۴	۰/۹۷۳۸	۰/۹۶۵۵	۰/۹۵۹۶	۷۰	
SHS ₂	Bias	-۰/۰۰۰۲۲	-۰/۰۰۰۱۰	۰/۰۰۰۲۳	۰/۰۰۰۵۸	۰/۰۰۰۴۴	۰/۰۰۰۳۸	۷۰	
	MSE	۰/۰۰۰۸۳	۰/۰۰۰۷۹	۰/۰۰۰۲۹	۰/۰۰۰۲۴	۰/۰۰۰۴۱	۰/۰۰۰۴۲۷	۷۰	
SSt	Bias	۱/۰۳۲۶	۰/۲۷۵۱	۱/۳۵۸۰	۱/۰۲۱۲	۰/۸۵۸۴	۱/۶۲۰۲	۷۰	
	MSE	۰/۹۲۸۲	۰/۹۳۳۵	۰/۹۴۳۰	۰/۹۳۷۹	۰/۹۵۰۴	۰/۹۵۱۰	۷۰	
SN	Bias	۲/۰۶۲۰	۰/۸۳۴۰	۰/۹۴۴۸	۰/۵۰۸۵	۰/۹۲۹۷	۱/۴۴۴۴	۷۰	
	MSE	۱/۱۶۶۴	۱/۲۷۳۳	۱/۲۰۰۰	۱/۱۸۰۸	۱/۱۷۳۸	۱/۲۰۲۰	۷۰	

نتایج به دست آمده نشان می دهند که تقریباً برای همه سناریوها، اریبی و MSE برای مدل صحیح با افزایش نمونه به صفر میل می کنند. در واقع سازگاری و نااریبی برآوردگرها را می توان

جدول ۳.۲: ارزیابی و MSE برآورد پارامترها برای مدل صحیح و نسبت ارزیابی و MSE برای مدل های رقیب در سناریوی سوم

مدل	n	۱۰۰		۲۰۰		۵۰۰	
		$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$
Bias SHS ₁		۱/۱۴۸۹	۱/۱۰۹۲	۱/۰۵۵۹	۱/۰۱۷۷	-۲/۴۴۰۰	۱/۲۴۶۱
MSE		۱/۰۱۱۲	۱/۰۲۶۳	۱/۰۱۷۲	۱/۰۱۶۰	۱/۰۴۵۴	۱/۰۱۱۹
Bias SHS ₂		۱/۱۶۷۰	۱/۱۶۷۸	۰/۹۷۶۶	۱/۰۱۴۷	-۱/۷۷۳۳	۱/۴۵۳۹
MSE		۱/۰۴۹۶	۱/۰۶۸۰	۱/۰۵۲۴	۱/۰۵۱۵	۱/۰۷۲۲	۱/۰۵۰۲
Bias SSt		-۰/۰۰۴۸	۰/۰۰۲۷	-۰/۰۰۳۳	-۰/۰۰۱۴	۰/۰۰۰۱	-۰/۰۰۰۷
MSE		۰/۰۰۸۴	۰/۰۰۷۴	۰/۰۰۳۹	۰/۰۰۴۱	۰/۰۰۱۶	۰/۰۰۱۶
Bias SN		۰/۹۱۵۹	۰/۲۱۴۷	۱/۴۱۸۴	۱/۴۴۱۲	-۲۴/۲۲۷۴	۱/۹۱۰۴
MSE		۱/۵۱۲۳	۱/۵۶۵۸	۱/۴۱۳۲	۱/۴۰۸۹	۱/۵۶۲۹	۱/۴۷۷۱

نتیجه گرفت. هم چنین برای همه حجم نمونه ها و سناریوها، برآوردهای به دست آمده مدل SN دارای کارایی کمتری هستند. برای سناریوی سوم که خطاها از توزیع SSt تولید شده اند، برآوردهای تولید شده از مدل صحیح، به طور یکنواختی کاراتر از بقیه هستند ولی با مدل های SHS₁ و SHS₂ تفاوت اندکی دارد. همین نتیجه برای سناریوی اول نیز برقرار است. بنابراین می توان نتیجه گرفت اگر داده های دم سنگین را به اشتباه نیمه دم سنگین در نظر بگیریم و به آنها توزیع های SHS را برازش کنیم، برآوردهای داده های رگرسیونی به خوبی برآورد خواهند شد.

شکل ۴.۲ میانگین ارزیابی به اضافه و منهای میانگین خطای معیار^{۱۷} (SE) برای ضرایب رگرسیون برای هر سه حجم نمونه در هر سناریو را نمایش می دهد. نتایج شکل ۴.۲ تحلیل به دست آمده از جداول ۱.۲ تا ۳.۲ را تایید می کنند. یعنی هم ارزیابی و هم MSE با افزایش اندازه نمونه به صفر میل می کند. طول بازه ها در سناریوهای مختلف برای اندازه نمونه های یکسان تقریباً برابر است. برای مدل SN، بازه ها بیشترین طول را دارند.

برای برآوردهای ضرایب، نرخ های پوشش تجربی را نیز تحت اندازه های نمونه و سناریوهای مختلف محاسبه و در شکل ۵.۲ نمایش داده ایم. با توجه به شکل ۵.۲ نرخ های پوشش برآورده شده نزدیک به سطح ۹۵ درصد برای تمام حالت ها است.

بررسی و مقایسه نیکویی برازش نیز از ابزارهای مقایسه مدل های مختلف می باشد که برای این هدف معیارهای AIC و BIC را در حالت های مختلف تولید نمونه محاسبه کردیم. برای

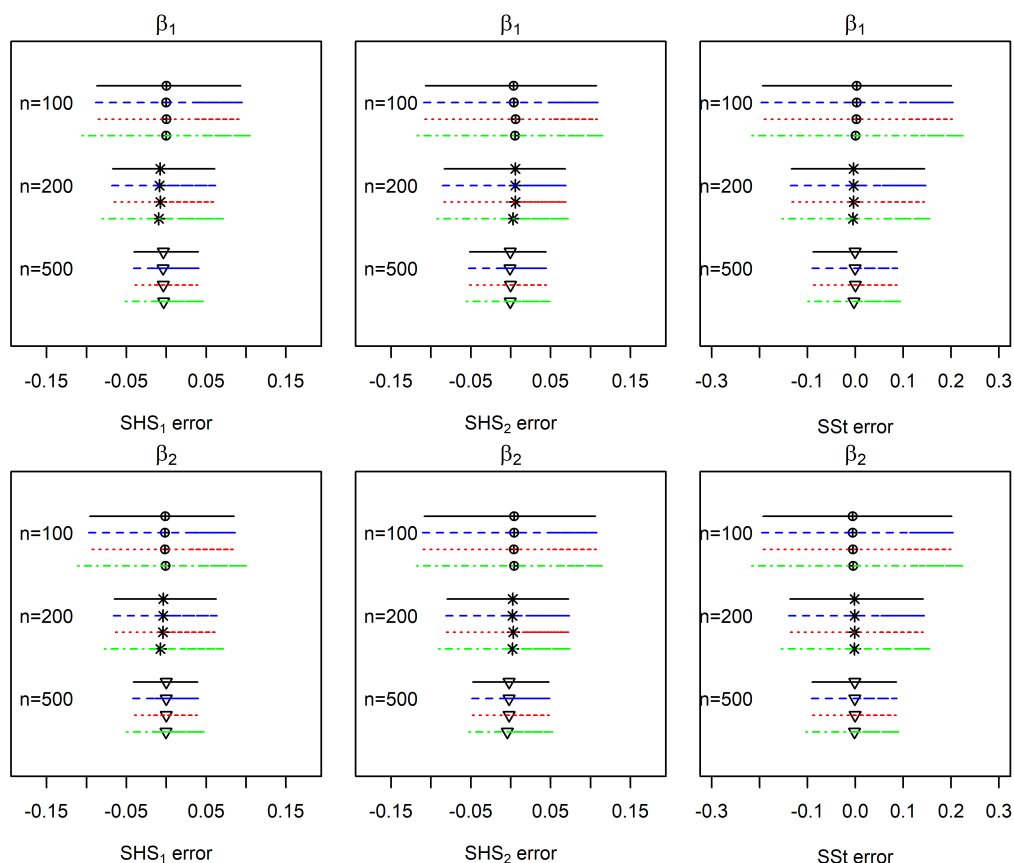
¹⁷Standard error

مقایسه بهتر و راحت‌تر به جای بررسی مستقیم AIC و BIC اختلاف‌های آنها را به صورت

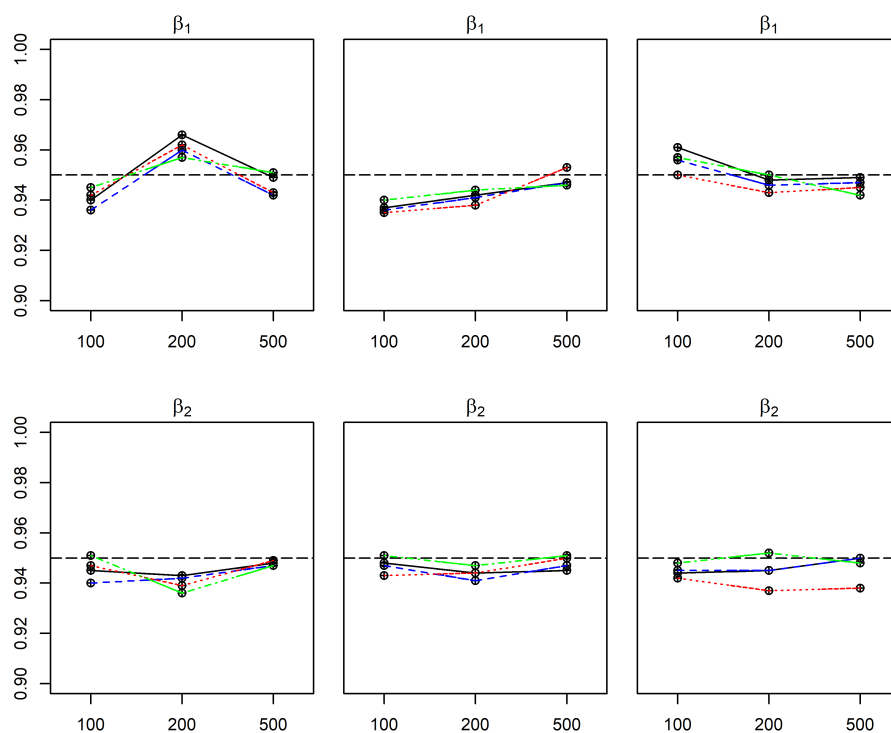
$$DAIC(\mathcal{M}) = AIC(\mathcal{M}_0) - AIC(\mathcal{M})$$

$$DBIC(\mathcal{M}) = BIC(\mathcal{M}_0) - BIC(\mathcal{M})$$

بررسی کردیم که در آن $\mathcal{M}_0$ مدل صحیح و $\mathcal{M}$ مدل رقیب هستند. مقادیر منفی DAIC و DBIC به نفع مدل صحیح است. نمودار جعبه‌ای این نتایج در شکل‌های ۶.۲ و ۷.۲ رسم شده‌اند. برای سناریوی ۱ و ۳، مدل صحیح با افزایش حجم نمونه بهترین عملکرد را نشان می‌دهند. برای سناریوی ۲ مدل ناصحیح SSt یا SHS₁ عملکرد بهتری از نظر معیارهای AIC و BIC دارند. هم‌چنین می‌توان گفت که برای مدل صحیح SSt، مدل‌های SHS نیز از نظر معیار AIC و BIC تقریباً به خوبی SSt است. برآش این معیارها نیز استواری SHS ها را در مقابله با داده‌های دم‌سنگین تایید می‌کنند.



شکل ۴.۲: میانگین آریبی به اضافه و منهای میانگین sd برای سناریوها و حجم نمونه‌های مختلف: SHS₁ (چپ)، SHS₂ (وسط) و SSt (راست). SHS₁ (خط ممتد)، SHS₂ (خط چین)، SSt (نقطه چین) و SN (نقطه-خط).



شکل ۵.۲: نرخ های پوشش برآورد شده برای هر یک از پارامترهای رگرسیون برای سناریوها و حجم نمونه های مختلف: SHS_1 (چپ)، SHS_2 (وسط) و SSt (راست). SHS_1 (خط ممند)، SHS_2 (خط چین)، SSt (نقطه چین) و SN (نقطه-خط).

۲.۲ مثال های کاربردی

در این بخش، سه مجموعه واقعی برای مقایسه مدل های پیشنهادی با مدل های رقیب، مورد بررسی قرار می گیرند. مجموعه داده های واقعی عبارتند از

۱. داده های صوت شناسی

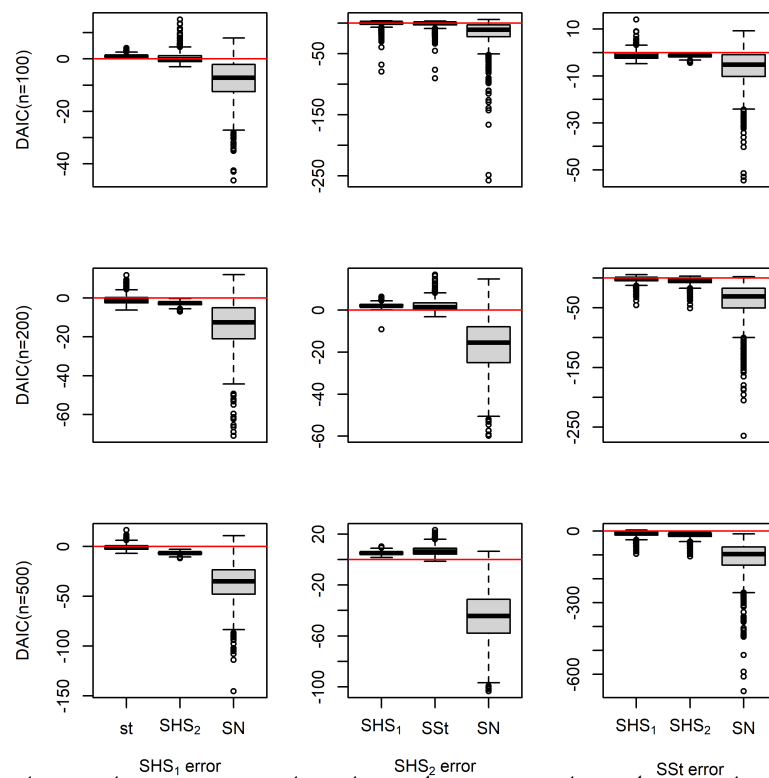
۲. داده های قیمت گذاری سنگ الماس

۳. داده های شرکت مارتین ماریتا

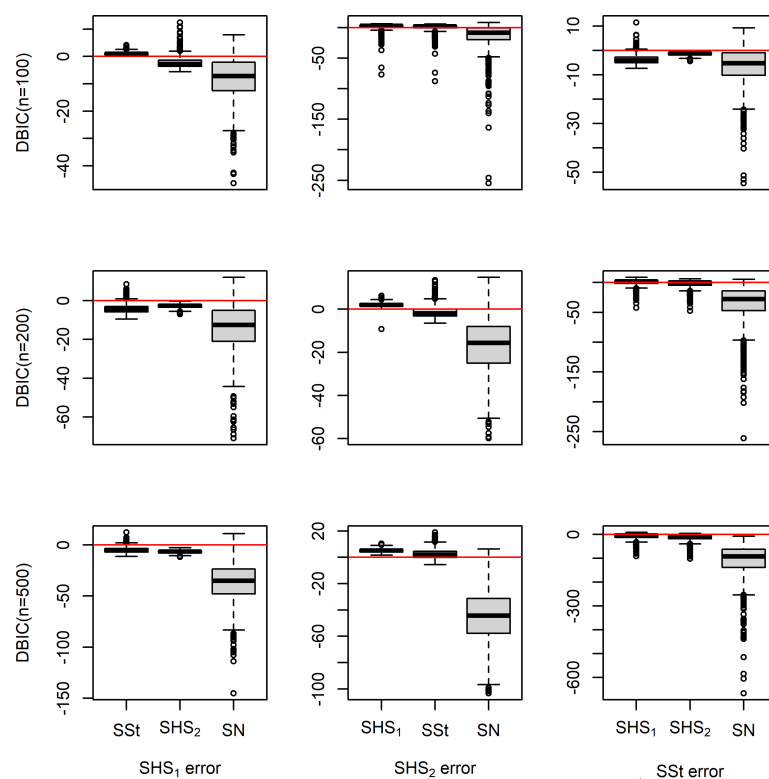
در این مطالعه توزیع عبارت خطا را SHS_1 ، SHS_2 ، SN ، SSt ، GSH ، $SGHS$ ، BHS و نرمال با پارامتر مکان $\mu = 0$ در نظر گرفته ایم. برای نیکویی برازش مدل های رگرسیون خطی روی داده ها با توزیع های مختلف روی عبارت خطاها از آزمون کولموگروف-اسمیرنوف^{۱۸} (KS) بر پایه توزیع تجربی مانده ها بهره گرفتیم. مدل رگرسیون خطی که در نظر گرفته ایم برابر است با

$$Y = \beta'x + \sigma\epsilon$$

¹⁸Kolmogorov-Smirnov



شکل ۶.۲: نمودارهای جعبه‌ای مقادیر DAIC برای سناریوها و حجم نمونه‌های مختلف: SHS_1 (چپ)، SHS_2 (وسط) و SSt (راست).



شکل ۷.۲: نمودارهای جعبه‌ای مقادیر DBIC برای سناریوها و حجم نمونه‌های مختلف: SHS_1 (چپ)، SHS_2 (وسط) و SSt (راست).

که در آن تابع توزیع جمله خطا عبارتست از $F_{\epsilon}(y) = P(\frac{Y - \beta'x}{\sigma} \leq y)$. اگر $\hat{\theta}$ برآوردگر ML ضرایب رگرسیونی و پارامترهای توزیع خطا باشد، آن گاه توزیع خطا تحت این مدل به صورت زیر ساخته می شود:

$$F_{\epsilon_0}(y) = P(\frac{Y - \hat{\beta}'x}{\hat{\sigma}} \leq y).$$

بنابراین در مدل رگرسیون خطی با توزیع خطای مشخص، زمانی که درست برازش شده باشد، هر دو توزیع F_{ϵ} و F_{ϵ_0} مشابه هم هستند. با استفاده از این واقعیت آزمون KS برای توزیع خطا با مقایسه تابع توزیع تجربی خطا، $\hat{F}_{\epsilon}$ ، و تابع توزیع برآورد شده تحت توزیع فرض شده، $\hat{F}_{\epsilon_0}$ ، به صورت زیر ساخته می شود:

$$T = \sqrt{n} \sup_{y \in \mathcal{R}} |\hat{F}_{\epsilon}(y) - \hat{F}_{\epsilon_0}(y)|.$$

۱.۲.۲ داده های صوت شناسی

ژانگ^{۱۹} و همکاران (۲۰۱۸) یک مدل رگرسیون خطی با پذیره نرمال بودن خطاها را برای ارزیابی آسایش صوتی در مسائل مربوط به وسایل نقلیه با سر و صدای محیطی بالا، بررسی کردند. آلودگی صوتی بالا، سلامتی جسمانی و ذهنی رانندگان را تهدید می کند. ژانگ و همکاران (۲۰۱۸) اشاره کردند که، در آینده، آلودگی صوتی وسایل به دلیل ارتباط مستقیم با زندگی روزمره از موضوعات اضطراری تحقیقات خواهد بود. بنابراین پرداختن به داده های صوت شناسی در وسائل مهندسی مهم و حیاتی است.

اکثر مطالعات آسایش صوتی روی ارزیابی عینی و ذهنی و روابط ریاضی بین آنها متمرکز است (دی^{۲۰} و همکاران (۲۰۱۶)، ما و همکاران (۲۰۱۶)، ژو^{۲۱} و همکاران (۲۰۱۴)). ارزیابی عینی آسایش صوتی بر پایه پارامترهای صداشناختی مانند بلندی و تیزی صدا انجام می شود. هم چنین رنجش^{۲۲}، می تواند شاخص مناسبی برای ارزیابی ذهنی باشد. رگرسیون خطی، یک مدل استاندارد برای یافتن روابط بین پارامترهای ذهنی صداشناختی و سنجش عینی صدا هست (ژانگ و همکاران (۲۰۱۸)).

داده های این قسمت شامل ۵۰ نمونه از صدای جرثقیل از یک مطالعه کاربردی موردی است. شاخص ارزیابی عینی (رنجش) به ۱۰ بخش مختلف تقسیم شده است. به طور دقیق تر شاخص رنجش به ۵ دسته از کم به زیاد و هر دسته به دو طبقه تقسیم شده است که در جدول ۴.۲ ارائه شده است. همانند ژانگ و همکاران (۲۰۱۸)، ۹ پارامتر عینی را شامل:

¹⁹Zhang

²⁰Di

²¹Xu

²²Annoyance

جدول ۴.۲: بخش‌های شاخص ارزیابی عینی (رنجش)

رنجش	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
بخش	۱ ۲	۳ ۴	۵ ۶	۷ ۸	۹ ۱۰

سطح فشار صوت خطی^{۲۳} (LSPL)، سطح فشار صوت وزنی^{۲۴} (ASPL)، بلندی صدا^{۲۵}، تیزی صدا^{۲۶}، زبری صدا^{۲۷}، نوسان صدا^{۲۸}، تناژ صدا^{۲۹}، شاخص بیان^{۳۰} (AI) و تکانشگری^{۳۱} به عنوان شاخص غیرارزیابی انتخاب کردیم. جزئیات داده‌ها در جدول ۲ ژانگ و همکاران (۲۰۱۸) ارائه شده‌است. این مجموعه داده شامل یک یا حتی دو داده پرت (جدول ۲ ژانگ و همکاران (۲۰۱۸)) است، و برای توزیع نرمال برای خطای مدل رگرسیون خطی مناسب نیست. با این مقایسه، برای توزیع خطای این مدل، مدل‌های پیشنهادی خودمان را فرض می‌کنیم. کاری که در ابتدا انجام دادیم این است که با روش ML، برآوردهای به دست آمده از مدل‌های مختلف را بررسی و با حذف برخی متغیرهای تبیینی بهترین مدل را انتخاب کردیم.

بهترین مدل انتخاب‌شده شامل متغیرهای تبیینی بلندی صدا، تیزی صدا و تکانشگری هستند. برای مدل انتخاب‌شده برآوردهای ML را به همراه خطای استاندارد^{۳۲} (sd) و مقادیر AIC و BIC تحت توزیع‌های رقیب محاسبه کردیم. نتایج در جدول ۵.۲ گزارش شده‌اند.

نتایج نشان می‌دهند که برآوردهای مدل نرمال به صورت قابل توجهی متفاوت از سایر مدل‌ها است. در واقع مدل مناسب برای این مجموعه داده نرمال نیست. تحت SHS‌ها هر سه متغیر تبیینی تاثیر معنی‌دار مثبتی روی متغیر پاسخ دارند. مقادیر کوچک AIC و BIC نشان‌دهنده برتری مدل هستند. بنابراین در بین مدل‌های با پیچیدگی یکسان برازش‌شده، مدل‌های SHS بر طبق معیارهای AIC و BIC برازش بهتری دارد. همچنین SHS₁ نسبت به توزیع نرمال نیز برتری قابل توجهی دارد. توزیع SHS₂ نیز عملکرد نزدیکی به SHS₁ دارد. جدول ۶.۲ نتایج آزمون KS را برای تمام توزیع‌های فرض‌شده برای خطاها نشان می‌دهد. P-مقدارهای جدول ۶.۲ نیکویی برازش تمام مدل‌ها تایید می‌کند. بافت نگار مانده‌ها و تابع چگالی برآورده‌شده آنها نیز در شکل ۸.۲ رسم شده‌اند که بیانگر این حقیقت است که همه مدل‌ها به درستی برازش شده‌اند.

²³Linear sound pressure level

²⁴A-weighted Sound Pressure Level

²⁵Loudness

²⁶Sharpness

²⁷Roughness

²⁸Fluctuation

²⁹Tonality

³⁰Articulation Index

³¹Impulsiveness

³²Standard Deviation

جدول ۵.۲: نتایج MLE تحت مدل های مختلف برای داده های صوت شناسی

Normal	SN	SSt	SGHS	GSH	BHS	SHS _۲	SHS _۱	
								پارامترهای رگرسیون:
-۰/۰۰۶۹	-۱/۳۰۹۱	-۱/۱۸۵۵	-۰/۱۵۹۴	-۰/۷۹۷۷	-۱/۰۲۲۲	-۱/۴۲۸۰	-۱/۳۸۶۳	ضریب ثابت
۰/۶۹۰۱	۰/۶۱۷۴	۰/۴۴۱۶	۱/۵۴۵۹	۰/۴۶۲۷	۰/۹۳۰۸	۰/۵۲۲۹	۰/۴۹۲۶	se
۰/۰۴۷۴	۰/۰۴۷۲	۰/۰۴۱۵	۰/۰۴۷۷	۰/۰۵۳۴	۰/۰۵۳۴	۰/۰۴۷۹	۰/۰۴۷۳	loudness
۰/۰۰۸۴	۰/۰۰۶۵	۰/۰۰۴۰	۰/۰۱۲۶	۰/۰۰۵۴	۰/۰۰۹۷	۰/۰۰۶۲	۰/۰۰۶۵	se
۰/۵۵۸۵	۰/۷۰۷۰	۰/۸۴۷۱	۰/۵۲۸۲	۰/۵۰۶۲	۰/۴۹۰۲	۰/۷۰۴۶	۰/۷۳۰۰	sharpness
۰/۳۰۲۶	۰/۲۴۴۱	۰/۱۲۶۲	۰/۵۴۶۶	۰/۱۷۹۹	۰/۳۷۳۴	۰/۲۲۲۵	۰/۲۲۷۹	se
۱/۲۱۹۰	۱/۷۰۴۰	۱/۵۸۵۲	۱/۳۵۱۵	۱/۹۰۱۳	۲/۰۶۶۹	۲/۰۴۲۰	۲/۰۵۷۷	impulsiveness
۰/۴۶۸۲	۰/۴۰۰۵	۰/۳۷۸۱	۰/۶۲۳۳	۰/۲۸۵۴	۰/۲۸۶۸	۰/۲۸۲۶	۰/۲۹۱۶	se
								پارامترهای توزیعی:
۰/۵۰۸۲	۰/۶۸۶۹	۰/۲۹۸۵	۰/۳۸۸۰	۰/۴۹۰۱	۰/۱۳۰۱	۰/۱۹۰۶	۰/۱۳۵۴	مقیاس
۰/۰۵۰۸	۰/۰۸۳۱	۰/۰۹۱۴	۰/۰۸۲۰	۰/۱۱۳۱	۰/۱۱۷۶	۰/۰۵۱۰	۰/۰۴۶۸	se
	۲/۶۰۲۹	۲/۰۵۱۱	۰/۸۰۵۸	-۲/۷۲۹۱	۰/۶۷۵۱	۰/۵۱۳۱		شکل ۱
	۰/۸۸۳۲	۱/۱۷۸۷	۰/۳۶۷۵	۰/۲۷۷۷	۰/۸۷۲۹	۰/۱۹۰۳		
			۱/۰۹۷۲		۰/۳۳۷۶			شکل ۲
			۱/۳۱۶۱		۰/۳۳۲۷			se
		۰/۹۵۶۳					۰/۳۳۲۵	درجه آزادی
		۰/۲۱۸۹					۰/۱۵۵۲	se
۸۴/۱۹۹۰	۷۴/۶۹۵۲	۷۱/۷۴۲۳	۶۸/۵۸۹۸	۵۹/۰۱۶۱	۶۰/۶۶۶۵	۵۸/۲۶۸۵	۵۷/۹۱۴۳	AIC
۹۳/۷۵۹۱	۸۴/۱۶۷۳	۸۵/۱۲۶۵	۸۱/۹۷۳۹	۷۰/۴۸۸۲	۷۴/۰۵۰۷	۶۹/۷۴۰۶	۶۹/۳۸۶۵	BIC

جدول ۶.۲: آزمون KS برای داده های داده های صوت شناسی

Normal	SN	SSt	SGHS	GSH	BHS	SHS _۲	SHS _۱	
۰/۱۸۱۳	۰/۱۴۲۷	۰/۱۵۹۴	۰/۰۹۷۸	۰/۰۵۸۴	۰/۰۶۲۸	۰/۰۵۴۷	۰/۰۵۷۷	آماره آزمون
۰/۰۶۵۶	۰/۲۳۶۵	۰/۱۴۱۱	۰/۶۸۸۱	۰/۹۹۱۷	۰/۹۸۲۲	۰/۹۹۶۴	۰/۹۹۲۸	P-مقدار

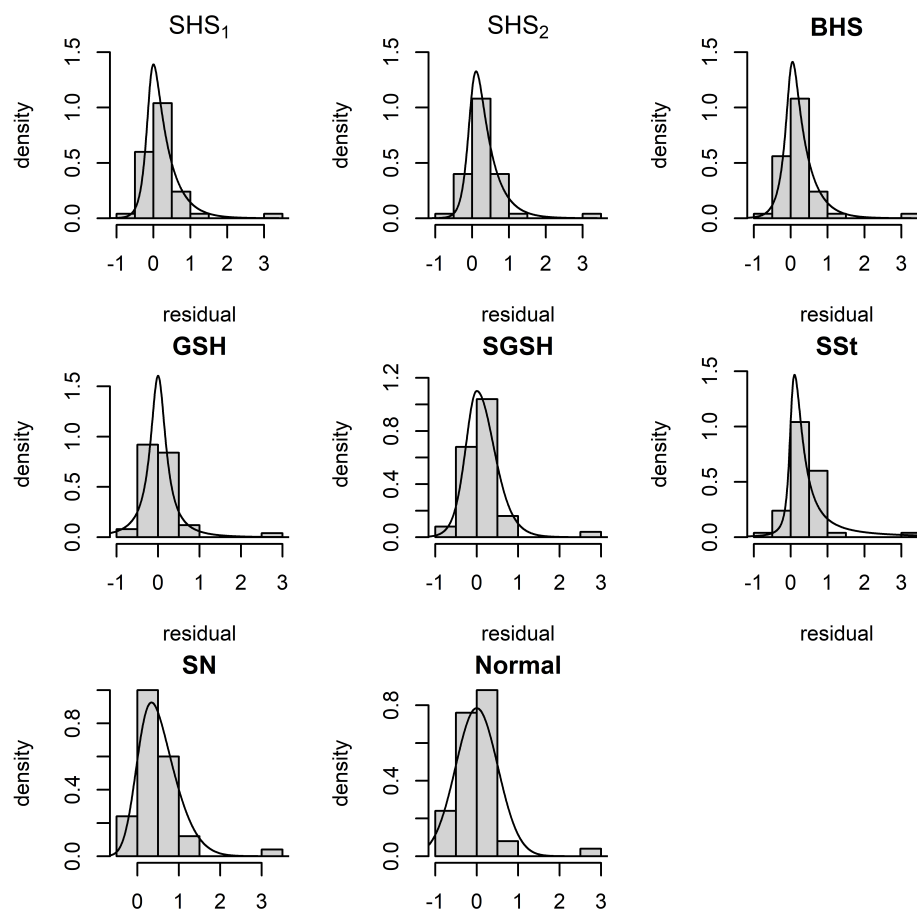
۲.۲.۲ داده های الماس

چو^{۳۳} (۲۰۰۱) استفاده از یک مدل قیمت گذاری برای سنگ های الماس را با کمک داده های چاپ شده در مجله زمان تجارت سنگاپور^{۳۴} تشریح کرد. در مجله مذکور، اطلاعات ۳۰۸ سنگ الماس ثبت و مورد بررسی قرار گرفته بودند. این مجموعه داده در بسته Ecdat نرم افزار R قرار دارد. در مدل خطی که ما در نظر گرفتیم، لگاریتم قیمت سنگ الماس را به عنوان متغیر پاسخ و وزن (قیراط)، شفافیت، رنگ و متن تاییدیه را به عنوان متغیرهای تبیینی در نظر گرفتیم. واحد وزن سنگ های الماس نیز قیراط (برابر ۰/۲ گرم) است. سه متغیر تبیینی باقی مانده، متغیرهای طبقه ای هستند. طبقات شفافیت الماس به ترتیب نزولی از معیوب ذاتی^{۳۵} (IF)،

^{۳۳}Chu

^{۳۴}Singapore's Business Times

^{۳۵}Internally Flawless



شکل ۸.۲: بافت نگار مانده‌ها و تابع چگالی برآوردشده داده‌های اصوت شناسی برای توزیع‌های مختلف برای عبارت خطا

معیوب خیلی خیلی کم^{۳۶} (VVS1 یا VVS2)، و معیوب خیلی کم^{۳۷} (VS1 یا VS2) هستند. باارزش‌ترین الماس‌ها رنگ خالصی دارند. بالاترین خلوص رنگ را درجه D می‌نامند. درجات بعدی آن به ترتیب درجه‌های E، F، G و غیره را به خود اختصاص می‌دهند. برای مجموعه داده‌های حاضر، ۶ درجه مختلف از خلوص رنگ را داریم. همچنین، تاییدیه‌های مختلف نیز با بررسی سنگ‌های الماس، شفافیت و خلوص رنگ آنها را مشخص می‌کنند. سه نوع متن تاییدیه مختلف IGI، GIA و HRD وجود دارند.

در تحلیل این داده‌ها، متغیرهای طبقه‌ای را به‌عنوان متغیرهای نشانگر وارد مدل کردیم. رنگ I را به‌عنوان طبقه مبنا برای خلوص رنگ انتخاب و آن را با پنج رنگ دیگر مقایسه کردیم. برای شفافیت، VVS2 را به‌عنوان طبقه مبنا انتخاب کردیم. همچنین، برای پیام تاییدیه، IGI به‌عنوان طبقه مبنا انتخاب شد. بنا بر نتایج چو (۲۰۰۱)، بین قیراط و پیام‌های تاییدیه نیز اثرات متقابل را در نظر گرفتیم.

در جدول ۷.۲ نتایج برآورد ML و خطاهای استاندارد آنها برای ضرایب رگرسیون و پارامترهای

^{۳۶}Very very slightly imperfect

^{۳۷}very slightly imperfect

توزیع خطا تحت مدل‌های مختلف برای عبارت خطا، گزارش شده‌اند. با توجه به جدول ۷.۲، خطای استاندارد حاصل از مدل نرمال تفاوت زیادی با بقیه مدل‌ها دارد. در جدول ۷.۲ مقادیر AIC و BIC را نیز گزارش کردیم. هر دو معیار مقایسه، برتری قوی را برای مدل SHS₂ نشان می‌دهند که بر اساس آن مدل پیشنهادی بهترین توزیع برای عبارات خطا در این مجموعه داده است. برای این مثال، عملکرد مدل SHS₂ تقریباً معادل مدل SHS₁ است. جدول ۸.۲ نتایج آزمون KS را برای مدل‌های مختلف نشان می‌دهد که نیکویی برازش همه مدل‌ها را تایید می‌کند.

۳.۲.۲ داده‌های شرکت مارتین ماریتا

باتلر^{۳۸} و همکاران (۱۹۹۰) داده‌های نرخ بازده شرکت مارتین ماریتا را تحلیل کرده‌اند. آنها مدل رگرسیون خطی با متغیر پاسخ نرخ بازده شرکت (Y) و متغیر تبیینی^{۳۹} CRSP را به صورت زیر در نظر گرفتند:

$$Y = \beta_0 + \beta_1 \text{CRSP} + \epsilon. \quad (4.2)$$

این داده‌ها شامل ۶۰ مشاهده ماهانه است که از ژانویه ۱۹۸۲ تا دسامبر ۱۹۸۶ به مدت ۵ سال جمع‌آوری شده‌اند. این مجموعه داده توسط آزالینی و کاپیتانیو (۲۰۰۳)، تیلور و وربیلا^{۴۰} (۲۰۰۴)، دیکیشیو و مانتی^{۴۱} (۲۰۰۴) و سالازار^{۴۲} و همکاران (۲۰۱۲) تحلیل شده‌اند. این داده‌ها شامل یک مشاهده دورافتاده است (شکل ۹.۲)؛ افزون بر این بنا بر دیکیشیو و مانتی (۲۰۰۴) بر اساس بعضی معیارهای سنجش مشاهده دورافتاده نشان می‌دهند که در این داده‌ها ممکن است دو مشاهده دورافتاده باشند. بنابراین، ممکن است نرمال در نظر گرفتن توزیع خطاها در مدل (۴.۲) مناسب نباشد. بنابراین ما خانواده‌های مختلف ارائه‌شده را برای توزیع‌های خطا در مدل (۴.۲) در نظر گرفتیم. نتایج برآوردهای ML در جدول ۹.۲ گزارش شده‌اند. این نتایج نشان می‌دهند که برآوردهای توزیع نرمال اساساً برآوردهای ضریب رگرسیون متفاوتی در مقایسه با دیگر مدل‌ها ارائه می‌دهد. همانطور که باتلر و همکاران (۱۹۹۰) نیز اشاره کرده‌اند، ضرایب رگرسیون برآورده‌شده که با روش نرمال به دست آمده‌اند، شرکت مارتین ماریتا را در رده پریسک قرار می‌دهد. برآورد $\hat{\beta}_1 = 1/8024$ در حالت نرمال به این معنی است که نرخ بازده شرکت مارتین ماریتا نسبت به بازار سرمایه‌گزاری خیلی بیشتر است. با این حال برآوردهای متناظر برای مدل‌های SSt و SN به نحوی متفاوت از خانواده‌های توزیع دم‌نیمه‌سنگین هستند. مقادیر کوچک‌تر برای $\hat{\beta}_1$ در مدل‌های با دم نیمه‌سنگین پیشنهادی نسبت به مدل‌های SSt و SN، تاکید بیشتری بر طبقه‌بندی شرکت به عنوان نیمه قدرتمند دارد.

³⁸Butler

³⁹index of the excess rate of return for the New York stock exchange

⁴⁰Taylor and Verbyla

⁴¹DiCiccio and Monti

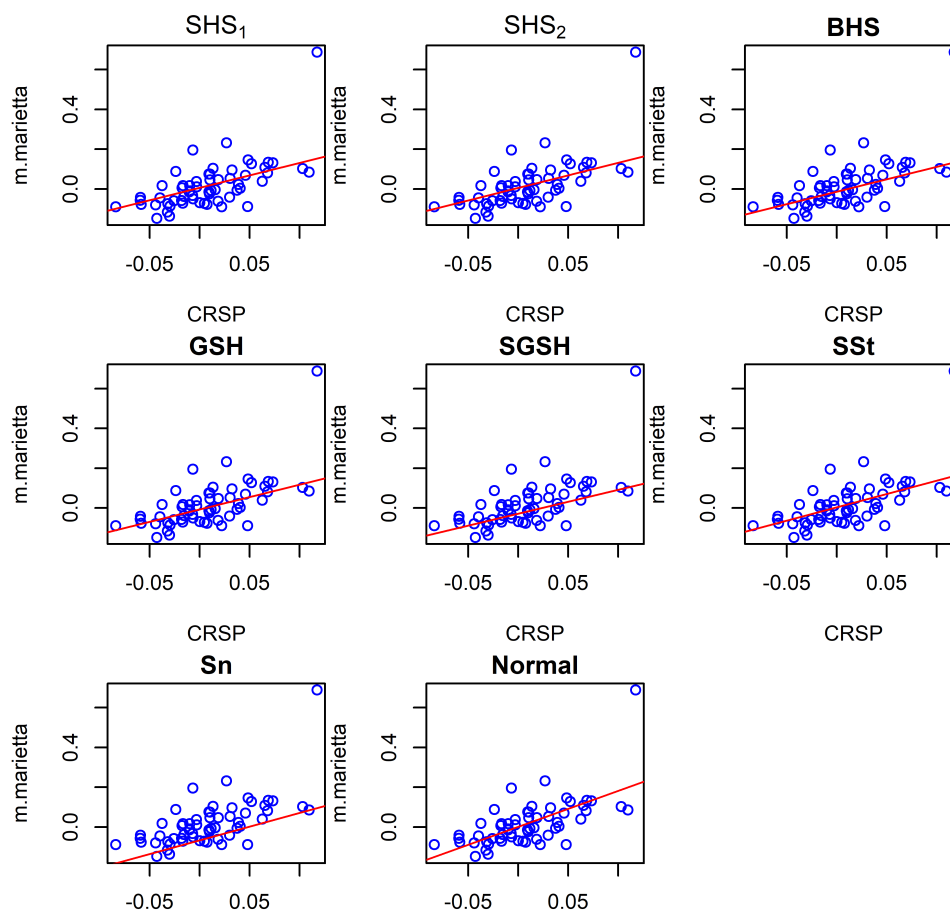
⁴²Salazar

جدول ۷.۲: نتایج MLE داده الماس برای توزیع‌های مختلف برای عبارت خطا

Normal	SN	SSt	SGHS	GSH	BHS	SHS _۲	SHS _۱	پارامترهای رگرسیون:
۵/۵۴۷۷	۵/۹۷۲۵	۵/۹۲۲۴	۵/۷۹۵۹	۵/۸۰۶۳	۵/۹۰۳۹	۵/۷۹۰۰	۵/۹۵۰۰	ضریب ثابت
۰/۰۹۳۴	۰/۰۴۱۵	۰/۰۴۰۹	۰/۰۴۵۳	۰/۰۴۰۴	۰/۱۱۵۳	۰/۰۳۲۰	۰/۰۴۳۷	se
۳/۷۳۰۱	۳/۴۴۶۲	۳/۴۷۸۲	۳/۶۵۵۵	۳/۵۴۶۲	۳/۴۳۴۴	۳/۸۵۴۰	۳/۴۶۳۸	carat
۰/۱۰۸۲	۰/۰۷۹۸	۰/۰۹۴۲	۰/۰۹۴۵	۰/۱۰۴۳	۰/۱۶۸۰	۰/۰۷۳۶	۰/۰۹۲۰	se
۰/۵۸۵۲	۰/۴۲۶۸	۰/۴۵۸۸	۰/۴۱۲۳	۰/۴۴۲۲	۰/۴۲۸۴	۰/۴۶۰۹	۰/۴۵۶۷	D
۰/۰۶۰۳	۰/۰۳۲۳	۰/۰۲۷۲	۰/۰۳۲۷	۰/۰۳۲۴	۰/۰۳۹۱	۰/۰۱۹۴	۰/۰۲۷۶	se
۰/۴۷۹۰	۰/۴۰۰۶	۰/۴۳۵۳	۰/۴۳۲۷	۰/۳۹۸۲	۰/۳۵۱۳	۰/۳۸۸۷	۰/۳۶۷۶	E
۰/۰۴۲۰	۰/۰۲۴۶	۰/۰۲۳۴	۰/۰۲۶۹	۰/۰۲۴۲	۰/۰۴۳۰	۰/۰۱۴۵	۰/۰۱۹۴	se
۰/۴۰۶۱	۰/۳۲۲۳	۰/۳۴۹۰	۰/۳۳۱۹	۰/۳۲۷۷	۰/۲۶۲۴	۰/۳۴۳۶	۰/۲۸۶۸	F
۰/۰۴۰۱	۰/۰۲۲۰	۰/۰۱۹۹	۰/۰۲۲۵	۰/۰۲۱۸	۰/۰۳۵۵	۰/۰۱۵۴	۰/۰۱۷۴	se
۰/۳۱۱۵	۰/۲۱۶۹	۰/۲۵۰۰	۰/۱۹۶۹	۰/۲۳۱۷	۰/۱۷۷۶	۰/۲۳۹۹	۰/۲۰۱۴	G
۰/۰۴۳۸	۰/۰۲۲۸	۰/۰۲۱۱	۰/۰۲۳۲	۰/۰۲۲۹	۰/۰۳۶۱	۰/۰۱۴۶	۰/۰۱۸۲	se
۰/۲۲۹۶	۰/۱۲۲۱	۰/۱۵۲۱	۰/۱۰۷۴	۰/۱۳۲۹	۰/۰۷۲۲	۰/۱۶۶۲	۰/۱۲۷۷	H
۰/۰۴۵۰	۰/۰۲۲۶	۰/۰۱۹۷	۰/۰۲۲۸	۰/۰۲۲۷	۰/۰۳۱۱	۰/۰۱۴۷	۰/۰۱۸۳	se
۰/۳۱۲۰	۰/۱۸۵۷	۰/۲۱۲۲	۰/۱۸۸۸	۰/۱۶۸۰	۰/۱۸۹۳	۰/۲۴۱۸	۰/۱۷۰۷	IF
۰/۰۴۹۹	۰/۰۲۴۳	۰/۰۲۳۴	۰/۰۲۵۴	۰/۰۲۳۰	۰/۰۲۷۸	۰/۰۱۷۴	۰/۰۱۹۴	se
-۰/۰۰۸۶	-۰/۱۱۹۱	-۰/۰۸۲۵	-۰/۱۰۵۰	-۰/۰۸۹۲	-۰/۰۸۸۶	-۰/۰۵۹۳	-۰/۰۸۷۹	VS1
۰/۰۳۷۹	۰/۰۱۹۱	۰/۰۱۶۰	۰/۰۲۰۱	۰/۰۱۷۸	۰/۰۲۳۷	۰/۰۱۱۱	۰/۰۱۴۲	se
-۰/۱۱۰۲	-۰/۱۹۶۸	-۰/۱۸۰۰	-۰/۲۴۳۷	-۰/۱۸۴۵	-۰/۱۸۲۰	-۰/۱۴۵۵	-۰/۱۵۴۹	VS2
۰/۰۳۷۵	۰/۰۲۱۰	۰/۰۱۹۲	۰/۰۲۵۶	۰/۰۱۹۸	۰/۰۲۳۴	۰/۰۱۳۷	۰/۰۱۶۸	se
۰/۱۷۱۵	۰/۰۹۷۶	۰/۱۰۰۱	۰/۰۵۰۶	۰/۰۸۸۴	۰/۰۹۷۵	۰/۰۸۲۴	۰/۰۶۹۷	VVS1
۰/۰۳۵۷	۰/۰۲۰۸	۰/۰۱۹۱	۰/۰۱۹۵	۰/۰۲۰۱	۰/۰۲۷۹	۰/۰۱۱۶	۰/۰۱۴۷	se
۰/۶۲۴۲	۰/۴۹۷۱	۰/۵۲۲۹	۰/۵۷۰۸	۰/۴۹۱۹	۰/۴۴۶۶	۰/۷۳۸۱	۰/۶۲۵۱	GIA
۰/۰۶۶۹	۰/۰۵۲۸	۰/۰۶۴۴	۰/۰۵۷۲	۰/۰۴۵۴	۰/۱۰۲۶	۰/۰۴۳۳	۰/۰۵۶۲	se
۱/۲۲۲۵	۰/۷۹۹۰	۰/۸۵۸۷	۰/۸۷۵۰	۰/۸۸۲۸	۰/۸۳۵۷	۰/۸۴۴۴	۰/۷۵۵۵	HRD
۰/۱۲۹۸	۰/۰۷۲۰	۰/۰۷۵۱	۰/۰۷۲۹	۰/۰۶۹۰	۰/۰۷۵۳	۰/۰۵۵۲	۰/۰۶۲۰	se
-۰/۹۱۲۷	-۰/۶۶۱۰	-۰/۷۱۵۲	-۰/۸۲۹۷	-۰/۷۳۸۲	-۰/۶۲۶۱	-۱/۲۴۹۴	-۰/۹۲۲۸	GIA × carat
۰/۱۱۶۰	۰/۰۹۹۳	۰/۱۲۶۵	۰/۱۲۰۹	۰/۱۰۸۳	۰/۲۲۷۴	۰/۰۸۹۶	۰/۱۰۸۱	se
-۱/۶۱۰۴	-۱/۰۷۰۱	-۱/۱۵۲۸	-۱/۲۲۷۱	-۱/۲۰۵۲	-۱/۱۰۷۲	-۱/۳۹۲۶	-۱/۱۱۷۵	HRD × carat
۰/۱۷۵۲	۰/۱۰۷۹	۰/۱۱۹۲	۰/۱۱۴۵	۰/۱۲۱۸	۰/۱۵۹۵	۰/۰۹۵۸	۰/۱۰۸۰	se
۰/۱۵۶۴	۰/۱۷۲۲	۰/۱۶۳۱	۰/۱۲۱۲	۰/۱۱۱۶	۰/۱۳۰۱	۰/۰۳۳۸	۰/۱۰۵۲	پارامترهای توزیعی:
۰/۰۱۹۸	۰/۰۱۱۲	۰/۰۱۲۳	۰/۰۱۱۷	۰/۰۰۵۹	۰/۰۷۳۴	۰/۰۰۷۳	۰/۰۰۶۰	مقیاس
	-۲/۲۷۱۲	-۳/۰۲۳۴	۱/۲۳۳۶	۳/۰۶۷۰	۱/۰۸۱۶	-۰/۷۴۳۹		شکل ۱
	۰/۵۳۹۳	۰/۵۶۵۷	۰/۲۱۲۰	۱/۲۶۹۲	۲/۲۷۳۰	۰/۰۵۸۲	-	شکل ۲
			-۰/۶۱۱۶		۱/۸۵۸۶			se
			۰/۵۸۲۹		۰/۳۹۲۰	-	-	درجه آزادی
		۵/۸۹۵۷					۹/۳۵۶۵	se
		۱/۲۹۵۹			-	-	(۳/۴۰۹۸)	AIC
-۳۲۲/۴۲۸۵	-۴۵۹/۴۳۶۴	-۴۴۸/۶۸۵۵	-۴۳۰/۲۰۶۷	-۴۶۲/۴۳۸۶	-۴۲۱/۹۱۱۳	-۴۹۳/۳۹۳۲	-۴۷۱/۶۵۵۴	BIC
-۲۶۲/۷۴۶۹	-۳۹۸/۰۲۴۷	-۳۸۱/۵۴۳۷	-۳۶۳/۰۶۴۹	-۳۹۹/۰۲۶۹	-۳۵۴/۷۶۹۵	-۴۲۹/۹۸۱۵	-۴۰۸/۲۴۳۷	

جدول ۸.۲: آزمون KS برای داده الماس

Normal	SN	SSt	SGHS	GSH	BHS	SHS _۲	SHS _۱	آماره آزمون
۰/۰۶۸۸	۰/۰۷۲۶	۰/۰۶۷۲	۰/۰۶۶۹	۰/۰۷۴۴	۰/۰۶۹۲	۰/۰۶۵۶	۰/۰۷۲۵	-P مقدار
۰/۱۰۸۲	۰/۰۷۷۹	۰/۱۲۳۹	۰/۱۲۶۴	۰/۰۶۶۳	۰/۱۰۴۴	۰/۱۴۱۱	۰/۰۷۸۸	



شکل ۹.۲: نمودار پراکندگی داده و خطوط رگرسیون منطبق با توزیع های مختلف برای عبارت خطا برای مجموعه داده مارتین ماریتا

برای مدل هایی که در این جا در نظر گرفتیم، مدل SHS_2 بنا بر معیارهای AIC و BIC برازش بهتری دارد. مشابه دو مثال قبلی، عملکرد مدل SHS_1 تقریباً معادل مدل SHS_2 است. شکل ۹.۲ خط رگرسیون برآورد شده روی نمودار پراکندگی داده ها برای توزیع های مختلف برای عبارت خطا را نشان می دهد. خط برازش شده نرمال به دلیل وجود داده پرت به بالا مایل شده است. ولی مدل های دمنیمه سنگین تحت تاثیر مشاهدات دورافتاده قرار نگرفته اند. از طرفی، مدل های SN و SSt نیز زیاد تحت تاثیر مشاهده دور افتاده نشده اند. برای این مجموعه داده، آزمون KS (جدول ۱۰.۲) نیکویی برازش توزیع های برازش شده را تایید می کند.

۳.۲ مدل رگرسیون نیمه پارامتری مکان-مقیاس برای داده های پرت

تعمیم های رو به گسترش مدل های آماری و روش های برازش آنها، در هر دو دیدگاه کلاسیک و بیزی، بستری برای مدل بندی واقعی تر داده های پیچیده فراهم کرده است. مدل های

مدل رگرسیون نیمه پارامتری مکان-مقیاس برای داده‌های پرت ۴۳

جدول ۹.۲: نتایج MLE داده شرکت مارتین ماریتا برای توزیع‌های مختلف برای عبارت خطا

Normal	SSt	SN	SGHS	GSH	BHS	SHS _۲	SHS _۱	
۰/۰۰۱۱	-۰/۰۷۵۵	-۰/۰۹۳۳	-۰/۰۰۷۵	-۰/۰۰۷۳	-۰/۰۴۹۲	-۰/۰۴۶۰	-۰/۰۲۷۵	پارامترهای رگرسیون: عرض از مبدأ
(۰/۰۰۸۷)	(۰/۰۲۰۷)	(۰/۰۱۲۵)	(۰/۰۰۸۹)	(۰/۰۰۸۷)	(۰/۰۳۳۲)	(۰/۰۱۶۷)	(۰/۰۱۲۵)	
۱/۸۰۲۴	۱/۳۳۹۱	۱/۳۷۹۰	۱/۲۴۵۹	۱/۲۴۵۶	۱/۲۵۱۵	۱/۲۶۴۳	۱/۲۴۸۲	شیب
(۰/۲۰۱۵)	(۰/۲۰۸۷)	(۰/۲۴۱۵)	(۰/۱۹۴۲)	(۰/۲۰۲۹)	(۰/۲۰۳۵)	(۰/۱۹۷۹)	(۰/۱۹۳۷)	
۰/۰۹۲۸	۰/۰۹۳۷	۰/۱۳۶۲	۰/۰۸۴۹	۰/۰۹۱۱	۰/۰۳۶۸	۰/۰۴۰۰	۰/۰۳۱۳	پارامترهای توزیعی: مقیاس
(۰/۰۰۸۵)	(۰/۰۰۴۲)	(۰/۰۲۹۳)	(۰/۰۱۵۵)	(۰/۰۱۸۰)	(۰/۰۲۳۲)	(۰/۰۰۷۲)	(۰/۰۰۷۶)	
	۲/۴۵۸۷	۳/۹۱۵۴	۰/۹۰۱۹	-۲/۵۶۰۷	۱/۲۶۵۷	۰/۴۳۸۶		شکل ۱
	(۱/۴۵۷۵)	(۱/۴۰۴۹)	(۰/۰۷۵۲)	(۰/۳۵۲۰)	(۱/۳۳۷۰)	(۰/۱۵۵۵)		
			-۲/۴۰۹۰		۰/۵۱۰۸			شکل ۲
			(۰/۴۴۷۴)		(۰/۳۷۳۱)			
	۶/۰۶۶۰						۰/۴۳۲۰	d.f.
	(۵/۴۵۰۸)						(۰/۱۴۸۸)	
-۱۰۸/۹۸۹۸	-۱۲۴/۰۳۴۱	-۱۳۴/۰۷۲۹	-۱۲۴/۰۳۴۱	-۱۳۲/۱۲۳۲	-۱۳۳/۱۷۶	-۱۳۵/۳۰۸۹	-۱۳۵/۱۷۶۵	AIC
-۱۰۲/۷۰۶۸	-۱۱۵/۶۵۶۷	-۱۲۳/۶۰۱۲	-۱۱۵/۶۵۶۷	-۱۲۳/۷۴۵۸	-۱۲۲/۷۰۴۳	-۱۲۶/۹۳۱۶	-۱۲۶/۷۹۹۱	BIC

جدول ۱۰.۲: آزمون KS برای داده شرکت مارتین ماریتا

Normal	SN	SSt	SGHS	GSH	BHS	SHS _۲	SHS _۱	
۰/۱۴۷۴	۰/۱۳۸۶	۰/۰۷۶۸	۰/۰۷۵۹	۰/۰۷۶۱	۰/۰۸۹۷	۰/۰۸۵۲	۰/۰۸۶۴	آماره آزمون
۰/۱۳۳۴	۰/۱۸۱۶	۰/۸۴۴۵	۰/۸۵۳۴	۰/۸۵۱۳	۰/۶۸۶۱	۰/۷۴۳۶	۰/۷۲۸۹	-P مقدار

رگرسیونی یکی از مهم‌ترین رده‌های مدل‌های آماری هستند که در اغلب زمینه‌های علمی، مانند اقتصاد، پزشکی، علوم اجتماعی، هواشناسی و صنعت، مورد استفاده قرار می‌گیرند. در چارچوب مدل‌بندی رگرسیونی، مدل‌های خطی تعمیم‌یافته و مدل‌های جمعی تعمیم‌یافته که در فصل اول به آنها اشاره شد جایگاه ویژه‌ای دارند. در هر دوی این زیررده‌ها، توزیع متغیر پاسخ، y ، عضوی از خانواده توزیع‌های نمایی در نظر گرفته می‌شود، که در آن میانگین پاسخ، μ ، به‌عنوان تابعی از متغیرهای تبیینی رگرسیونی مدل‌بندی می‌شود و واریانس پاسخ به پارامتر ثابت پراکندگی^{۴۳} φ وابسته است. همچنین میانگین μ از طریق تابع واریانس $\tau(\mu)$ به صورت $V(y) = \varphi\tau(\mu)$ وابسته است. بنابراین، در مدل‌های GLM و GAM واریانس توزیع پاسخ به صورتی واضح و سراسر بر حسب متغیرهای تبیینی مدل‌بندی نمی‌شود و تنها از طریق میانگین به‌صورت تلویحی به آنها وابسته است. توجه داشته باشید که مدل‌های رگرسیون خطی نیز عضوی از GLM هستند، به‌طوری که برای آنها $\tau(\mu) = 1$.

اگرچه مدل‌های GAM با در نظر گرفتن اثرات غیرخطی (ناپارامتری) متغیرهای تبیینی به‌صورت جمعی، نسبت به مدل‌های GLM، انعطاف بیشتری دارند، اما به غیر از پارامتر میانگین (مکان) اجازه مدل‌بندی سایر پارامترها، مثل واریانس توزیع پاسخ را نمی‌دهند. برای رفع این محدودیت، ریگی و استاسینوپولوس (۲۰۰۵) رده مدل‌های جمعی تعمیم‌یافته برای مکان، مقیاس و شکل (GAMLSS) را معرفی کردند. دو ویژگی مهم این رده از مدل‌ها عبارتند

⁴³Dispersion

از:

۱. یک خانواده کلی از توزیع های پارامتری، جایگزین خانواده توزیع های نمایی برای متغیر پاسخ می شود.

۲. نه فقط میانگین (یا مکان) بلکه همه پارامترهای توزیع شرطی پاسخ را می توان به عنوان تابعی پارامتری یا ناپارامتری جمعی از متغیرهای تبیینی مدل بندی کرد.

لازم به ذکر است که برای تحلیل پاسخ های وابسته، معمولاً از نسخه های آمیخته^{۴۴} مدل های GLM (بریسلو و کلیتون^{۴۵}، ۱۹۹۳، لی و نلدرد^{۴۶}، ۲۰۰۱) و GAM (لین و ژانگ^{۴۷}، ۱۹۹۹، فاهرمیر و لانگ^{۴۸}، ۲۰۰۱) استفاده می شود، که در آنها با افزودن اثرات تصادفی به صورت جمعی به پیشگوی مدل، ساختار وابستگی داده ها وارد مدل می شود. رده مدل های GAMLSS نیز این ویژگی را داراست که اثرات تصادفی را به هر کدام از پارامترهای توزیع اضافه کند. در دنیای واقعی، همیشه داده ها رفتاری ایده آل برای مدل بندی ندارند و معمولاً یک یا چند داده مشکوک (پرت) در جمع داده ها حضور دارند که می توانند بر الگوی حاکم بر داده ها تاثیر زیادی داشته باشند. در رده مدل های GAMLSS، توزیع های پارامتری مختلفی وجود دارند که به طور ذاتی دم سنگین و قادر به مدل بندی داده های پرت توسط پارامترهای مقیاس و شکل هستند (استاسینوپولوس و همکاران، ۲۰۱۷). به عنوان مثال، می توان به توزیع های SN و SSt اشاره کرد. علی رغم انعطاف قابل توجه مدل های موجود در رده GAMLSS که در دو بسته نرم افزار gamlss (استاسینوپولوس و ریگی (۲۰۰۷)) و gamlss.dist^{۴۹} در محیط نرم افزار R وجود دارند، امکان رخداد دو مشکل بالقوه وجود دارد:

۱. در ساختار زیررده مدل های رگرسیون خطی، معمولاً مقارن بودن جمله خطا با میانگین صفر پذیرفته می شود. این پذیره برای توزیع های چوله برقرار نیست و بنابراین استفاده از برخی از اعضای رده GAMLSS برای حالت خاص مدل های خطی منطقی نیست.

۲. زمانی که تعداد داده های پرت نسبت به کل حجم نمونه کوچک است، توزیع های دم سنگین مقارن تاکید بیش از حد نیاز را به دم ها اختصاص می دهند و در مقابل چگالی کمتری را برای مقادیر مرکزی توزیع (که اغلب داده های موجود را شامل می شوند) در نظر می گیرند. این پدیده می تواند به مدلی نامناسب منتهی شود.

با توجه به این دو مشکل، معرفی و امکان به کارگیری یک توزیع مقارن با دم های نیمه سنگین در رده GAMLSS می تواند کارساز باشد.

⁴⁴Mixed models

⁴⁵Breslow and Clayton

⁴⁶Lee and Nelder

⁴⁷Lin and Zhang

⁴⁸Fahrmeir and Lang

⁴⁹<https://cran.r-project.org/package=gamlss.dist>

۱.۳.۲ خانواده توزیع‌های هایپربولیک سکانت

توزیع HS اولین بار توسط فیشر^{۵۰} (۱۹۲۱) معرفی شد. این توزیع حول پارامتر مکان خود که میانگین توزیع هم هست، متقارن است. برخی از ویژگی‌های این خانواده توسط دود^{۵۱} (۱۹۲۵) و پرک^{۵۲} (۱۹۳۲) مورد بررسی قرار گرفتند. وگان (۲۰۰۲) این توزیع را به حالت چوله تعمیم داد و فیشچر و وگان (۲۰۰۲) توزیع هایپربولیک سکانت تعمیم‌یافته چوله را معرفی کردند. کتاب فیشچر (۲۰۱۳) با عنوان توزیع‌های HS تعمیم‌یافته و کاربردهای آن در داده‌های مالی، منبع جامع و کاملی برای مبانی نظری و کاربردهای این توزیع است. استفاده از این توزیع در مدل‌بندی و تحلیل داده‌های مالی، معمول است.

توزیع HS نسبت به توزیع نرمال دم‌های سنگین‌تری دارد ولی به سنگینی توزیع‌هایی مثل کوشی، لاپلاس یا t استیودنت نیست. بنابراین در متون مربوطه از آن به‌عنوان یک توزیع با دم نیمه‌سنگین یاد می‌شود. انتظار داریم این توزیع در مواقعی که دم توزیع جمله خطا بین دم‌های یک توزیع دم‌سنگین و نرمال باشد، عملکرد مناسبی داشته باشد. در واقع نیمه‌سنگین بودن دم توزیع خطا به معنی امکان مدل‌بندی تعداد نه‌چندان زیاد از داده‌های پرت است. نیمه‌سنگینی دم یک توزیع البته به‌طور رسمی تعریف و برای توزیع HS برهان می‌شود. تعریف رسمی از دم‌سنگینی را می‌توانید در ساتو^{۵۳} (۱۹۹۹) ببینید.

قضیه ۱.۳.۲. توزیع HS یک توزیع با دم نیمه‌سنگین است.

برهان. برای مشاهده برهان به ساتو (۱۹۹۹) مراجعه کنید. □

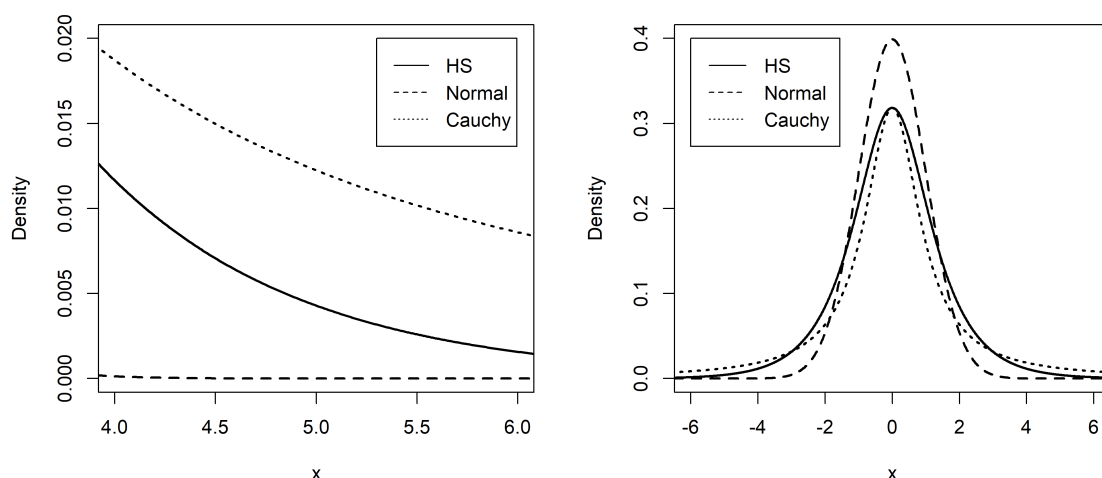
برای درک نیمه‌سنگینی دم توزیع HS، شکل ۱۰.۲ منحنی توابع چگالی و دم سمت راست این توزیع را در کنار دو توزیع نرمال و کوشی نشان می‌دهد. با توجه به شکل، کاملاً واضح است که دم توزیع HS بین دو توزیع نرمال (با دمی سبک) و کوشی (با دمی سنگین) قرار دارد. در این فصل، ما با تعبیه کردن توزیع HS به چارچوب رده مدل‌های GAMLSS امکان مدل‌بندی منعطف داده‌های پیوسته‌ای را فراهم کرده‌ایم که یک یا چند داده پرت را شامل می‌شوند. مدل مورد نظر ما یک مدل نیمه‌پارامتری مکان-مقیاس است که در ساختار مدل (۵.۱) با دو پارامتر $\mu = \theta_1$ و $\sigma = \theta_2$ صدق می‌کند. با توجه به این که خانواده توزیع‌های HS دارای دو پارامتر مکان و مقیاس است، مدل رگرسیون معرفی‌شده یک مدل مکان-مقیاس محسوب می‌شود. از اولین کارهایی که در آن مدل‌بندی همزمان مکان و مقیاس پیشنهاد شده است، می‌توان به ریگی و استاسینوپولوس (۱۹۹۶) اشاره کرد.

⁵⁰Fisher

⁵¹Dodd

⁵²Perks

⁵³Sato



شکل ۱۰.۲: نمودار توابع چگالی توزیع های HS، نرمال و کوشی (سمت راست) و دم های سمت راست آنها (سمت چپ)

استنباط مبتنی بر درست نمایی

رهیافت استنباطی مد نظر برای مدل پیشنهادی، مبتنی بر درست نمایی است. با توجه به این که در این مدل از مولفه های ناپارامتری جمعی مختلف می توان استفاده کرد، برای جلوگیری از پیچیدگی بیش از حد، از یک چارچوب درست نمایی توانانیده استفاده می شود. مدل GAMLSS پیشنهادی به صورت

$$g_1(\mu) = X_1\beta_1 + \sum_{j=1}^{J_1} h_{j1}(x_{j1})$$

9

$$g_2(\sigma) = X_2\beta_2 + \sum_{j=1}^{J_2} h_{j2}(x_{j2})$$

است که در آن، $h(x)$ ها را می توان به صورت یک اثر تصادفی، مشابه $h(x) = Z\gamma$ ، نمایش داد که Z یک ماتریس پایه است و بر اساس مقادیر x ساخته می شود و γ برداری از ضرایب (اثرات) تصادفی است که فرض می شود دارای توزیع نرمال چندمتغیره است. تحت فرض مستقل بودن مشاهدات پاسخ، تابع لگ درست نمایی مدل به صورت

$$\ell(\mu, \sigma) = \sum_{i=1}^n \log f(y_i; \mu_i, \sigma_i)$$

است که در آن

$$\log f(y_i; \mu_i, \sigma_i) = -\log \pi - \log \sigma_i + \log \Psi + \frac{y_i - \mu_i}{\sigma_i} - \log(e^{\Psi \frac{y_i - \mu_i}{\sigma_i}} + 1)$$

است. پارامترهای مکان و مقیاس توزیع HS با تابع چگالی $f(\cdot)$ توسط مدل (۵.۱) به متغیرهای تبیینی مرتبط می‌شوند. تابع درست‌نمایی تاوانیده برای این مدل به صورت زیر است:

$$\ell_p = \ell(\mu, \sigma) - \frac{1}{2} \sum_{k=1}^2 \sum_{j=1}^{J_k} \gamma_{kj}^T G_{kj}(\lambda_{kj}) \gamma_{kj}. \quad (5.2)$$

مجموعه پارامترهایی که باید برآورد شوند، عبارتند از بردار پارامترهای مولفه خطی مدل، یعنی $\beta = (\beta_1, \beta_2)'$ ، بردار پارامترهای اثرات تصادفی (توابع ناپارامتری)، یعنی

$$\gamma = (\gamma_{11}, \dots, \gamma_{J_1 1}, \gamma_{12}, \dots, \gamma_{J_2 2})'$$

و ابرپارامترهای مدل، یعنی $\lambda = (\lambda_{11}, \dots, \lambda_{J_1 1}, \lambda_{12}, \dots, \lambda_{J_2 2})'$. در چارچوب مدل‌های GAMLSS، پارامترهای خطی β و پارامترهای اثرات تصادفی γ ، به ازای مقادیر ثابتی برای ابرپارامترهای هموارساز λ ، با ماکسیم کردن تابع درست‌نمایی تاوانیده ℓ_p در (۵.۲) برآورد می‌شوند. برای این کار دو روش پایه‌ای با نام‌های الگوریتم‌های RS^{۵۴} و CG^{۵۵} (ریگبی و استاسینوپولوس (۲۰۰۵)) وجود دارند. هر دوی این روش‌ها از یک الگوریتم کمترین توان‌های دوم (تاوانیده) موزون تکراری^{۵۶} برای محاسبه برآوردهای ML پارامترهای مذکور، با مفروض بودن ابرپارامترهای هموارساز، استفاده می‌کنند. ریگبی و استاسینوپولوس (۲۰۰۵) نشان دادند برآوردهای حاصل از این روش‌ها، با در نظر گرفتن توزیع‌های پیشین ناآگاهی بخش تخت برای ضرایب رگرسیونی و نرمال برای ضرایب تصادفی γ ، برآوردهای مد پسینی^{۵۷} (MAP) در چارچوب یک رهیافت بیز تجربی نیز هستند.

روش RS تعمیم الگوریتم به کار گرفته شده توسط ریگبی و استاسینوپولوس (۱۹۹۶) است. روش CG نیز تعمیم الگوریتم کول و گرین^{۵۸} (۱۹۹۲) است، که در آن مشتقات مرتبه اول تابع درست‌نمایی و مقادیر امید ریاضی مشتقات مرتبه دوم تابع درست‌نمایی نسبت به پارامترهای توزیع یعنی μ و σ مورد نیاز هستند. البته برای مشتقات مرتبه دوم، مشابه روش بهینه‌سازی BFGS، می‌توان از تقریب‌های عددی نیز استفاده کرد. مشتقات اول توزیع HS نسبت به μ و σ عبارتند از

$$\frac{\partial \ell(\mu, \sigma)}{\partial \mu} = -\frac{1}{\sigma} + \frac{2}{\sigma} \frac{\exp(\frac{2(y-\mu)}{\sigma})}{\exp(\frac{2(y-\mu)}{\sigma}) + 1}$$

$$\frac{\partial \ell(\mu, \sigma)}{\partial \sigma} = -\frac{1}{\sigma} - \frac{y-\mu}{\sigma^2} + \frac{2}{\sigma^2} \frac{y-\mu}{\exp(\frac{2(y-\mu)}{\sigma}) + 1}.$$

برای مشتقات دوم نیز می‌توان از تقریب‌های عددی استفاده کرد. استاسینوپولوس و ریگبی (۲۰۰۷) تاکید کردند که الگوریتم CG نسبت به الگوریتم RS کلی‌تر ولی در همگرایی کندتر

⁵⁴Rigby and Stasinopoulos

⁵⁵Cole and Green

⁵⁶Penalty iterative weighted least squares (PIWLS)

⁵⁷Maximum a posteriori

⁵⁸Cole and Green

است. آنها نشان دادند در مواردی که پارامترهای توزیع از نظر اطلاع متعامد هستند (به این معنی که امید ریاضی مشتقات حاصل ضرب تابع درستنمایی نسبت به پارامترها صفر هستند)، استفاده از روش RS به دلیل سرعت بالاتر آن توصیه می شود؛ و در مواردی که پارامترها متعامد نباشند، الگوریتم CG از کارایی بیشتری برخوردار است. البته یک روش سازوار^{۵۹} توسط استاسینوپولوس و ریگی (۲۰۰۷) پیشنهاد شده است که از ترکیب هر دو الگوریتم در بهینه سازی تابع درستنمایی توانیده استفاده می کند.

برای برآورد ابرپارامترهای λ دو روش موضعی^{۶۰} (ریگی و همکاران، ۲۰۱۳) و فراموضعی^{۶۱} (ریگی و استاسینوپولوس، ۲۰۰۵) پیشنهاد شده اند. روش موضعی سریع تر و اجرای آن ساده تر است. در واقع در روش موضعی برآورد ابرپارامترها در هر مرحله الگوریتم های RS یا CG انجام می شود، در حالی که در روش فراموضعی، حداقل سه معیار برای برآورد ابرپارامترها وجود دارند:

۱. می نیم کردن معیار اعتبارسنجی متقابل تعمیم یافته^{۶۲} (GCV)؛ (وود^{۶۳}، ۲۰۰۶).

۲. می نیم کردن معیار AIC (آکائیک، ۱۹۷۴) یا معیار BIC (شوارتز، ۱۹۷۸).

۳. ماکسیم ساختن تابع درستنمایی.

ما در این فصل، از معیار سوم و به صورت موضعی استفاده کرده ایم. برای مقایسه و انتخاب مدل برتر در بین مدل های نامزد می توان از معیارهای AIC و BIC استفاده کرد. بررسی نیکویی برازش مدل ها نیز توسط تحلیل مانده ها انجام می شود. در ادامه برای ارزیابی مدل پیشنهادی ما در این فصل از یک مثال شبیه سازی استفاده کرده ایم.

۲.۳.۲ مطالعه شبیه سازی

برای ارزیابی عملکرد مدل پیشنهادی، در این بخش از یک مثال شبیه سازی استفاده کردیم. هدف از ارزیابی، مقایسه عملکرد مدل پیشنهادی با دو مدل نرمال (به عنوان یک مدل با دم سبک) و مدل رگرسیونی t استیودنت (به عنوان یک مدل با دم سنگین) بود. دلایل مقایسه با این دو مدل عبارتند از

- مدل مکان-مقیاس جدید HS را به عنوان جانشینی استوار^{۶۴} برای مدل مکان-مقیاس نرمال معرفی کردیم. برای همین برازش هر دو مدل HS و نرمال مورد نظر بودند.

⁵⁹ Adaptive

⁶⁰ Locally

⁶¹ Globally

⁶² Generalized Cross Validation (GCV)

⁶³ Wood

⁶⁴ Robust

• مدل مکان-مقیاس مبتنی بر توزیع t استیودنت نیز جانشینی استوار برای مدل نرمال محسوب می‌شود. اما با توجه به دم‌سنگین بودن این توزیع مایل بودیم نتیجه استفاده از آن و مدل با دم نیمه‌سنگین HS را بدانیم.

برای شبیه‌سازی داده‌ها از هیچ‌کدام از مدل‌های رقیب استفاده نشد. داده‌ها را از یک مدل رگرسیونی مکان-مقیاس بر پایه توزیع لجستیک تولید کردیم. توزیع لجستیک دمی سنگین‌تر از نرمال دارد ولی به اندازه توزیع t استیودنت به‌عنوان یک توزیع دم‌سنگین معروف و معمول نیست. برای تولید داده‌ها فرض کردیم مشاهده i ام متغیر پاسخ دارای توزیع لجستیک با پارامتر مکان μ_i و پارامتر مقیاس σ_i باشد، به طوری که

$$\mu_i = x_{1i} + \sin\left(\frac{\pi}{5}x_{2i}\right)$$

و

$$\sigma_i = \exp\left(x_{3i} + \sin\left(\frac{\pi}{5}x_{4i}\right)\right) \quad (6.2)$$

که در آن، برای $i = 1, \dots, n$ ، همه متغیرهای تبیینی x_1, x_2, x_3, x_4 از توزیع نرمال با میانگین صفر تولید شدند. انحراف معیار دو متغیر x_1, x_3 برابر 0.2 و دو متغیر تبیینی دیگر برابر 2 انتخاب شدند. همان‌طور که معلوم است در هر دو پارامتر مکان و مقیاس توزیع، هم اثر خطی وجود دارد و هم مولفه غیرخطی. حجم نمونه را برابر 100 ، 200 و 500 در نظر گرفتیم و برای هر کدام از آنها 200 مجموعه داده را از مدل (۶.۲) تولید و برای برازش مدل از روش‌های تشریح‌شده در بخش ۱.۳.۲ استفاده کردیم. برای مدل‌بندی مولفه ناپارامتری جمعی در هر دو پارامتر نیز از اسپلاین‌های توانیده استفاده کردیم.

برای برازش مدل‌های نرمال و t استیودنت از بسته gamlss در نرم‌افزار R استفاده کردیم که هر دو توزیع را در خود دارد. ولی مدل HS در این بسته وجود ندارد. پدیدآوران بسته gamlss این امکان را فراهم کرده‌اند که بتوان با کدنویسی در محیط نرم‌افزار R از امکانات مدل‌بندی و برازش موجود در این بسته برای سایر توزیع‌های جدید توسط سایرین استفاده کرد. با این امکان، از قابلیت موجود برای معرفی یک مدل نیمه‌پارامتری بر پایه توزیع HS بهره برده و کدهای مورد اشاره همراه با مثال ارائه‌شده در فصل را در لینک زیر^{۶۵} در دسترس خوانندگان قرار داده‌ایم.

برای ارزیابی برآوردهای پارامترهای خطی از معیار میانگین توان‌های دوم خطا^{۶۶} (MSE) استفاده کردیم که به صورت زیر تعریف می‌شود:

$$MSE(\hat{\beta}) = \frac{1}{200} \sum_{j=1}^{200} (\hat{\beta}_j - \beta)^2$$

^{۶۵}<https://github.com/JamilOwnuk/GAMLSS-HS/blob/master/HS%20distribution>

^{۶۶}Mean Squared Errors

که در آن $\hat{\beta}_j$ ضریب رگرسیونی برآوردشده در مجموعه داده تولیدشده j ام و β مقدار واقعی آن است. برای ارزیابی مدل بندی ناپارامتری و برآورد مولفه غیرخطی در هر دو پارامتر مکان و مقیاس از معیار MSE تجمیع شده^{۶۷} (IMSE، ساکس^{۶۸} و همکاران (۱۹۸۹)) با تعریف زیر استفاده کردیم:

$$IMSE\hat{f} = \frac{1}{200} \sum_{j=1}^{200} \frac{1}{n} \sum_{i=1}^n \left(\hat{f}_j(x_i) - f(x_i) \right)^2$$

که در آن $\hat{f}_j$ تابع برآوردشده برای مجموعه داده j ام و x_i ها مقادیری معلوم هستند که برای محاسبه اختلاف تابع برآوردشده و تابع واقعی در نظر گرفته می شوند؛ در این مثال مقادیر تولیدشده متغیرهای تبیینی را برای مقداردهی به کار گرفتیم.

نتایج دو معیار برای هر سه مدل رقیب در جدول ۱۱.۲ گزارش شده اند. با توجه به نتایج جدول، با افزایش حجم نمونه مقادیر MSE و $IMSE$ اثرات برآوردشده کاهش پیدا می کنند. بنابراین می توان برقراری نتایج نظری سازگاری مجانبی در این مدل ها را نتیجه گرفت. از طرفی، عملکرد دو مدل HS و t استیودنت در برآورد پارامترهای رگرسیونی و همین طور تابع جمعی هم برای مکان و هم مقیاس برتر از مدل نرمال است. اما برتری مشخصی برای دو مدل HS نسبت به t استیودنت یا برعکس در پارامتر مکان دیده نمی شود. در مقابل، برای پارامتر مقیاس مدل رگرسیونی، هم در برآورد ضرایب خطی و هم تابع جمعی، مدل HS کاملاً نسبت به مدل t استیودنت در برآزش داده ها برتری دارد. برای یک مجموعه از داده های

جدول ۱۱.۲: مقادیر MSE پارامترهای خطی و $IMSE$ اثر جمعی برای داده های شبیه سازی شده

حجم نمونه	مدل	پارامتر مکان			پارامتر مقیاس		
		جمعی	عرض از مبدأ	شیب	جمعی	عرض از مبدأ	شیب
۱۰۰	نرمال	۰/۰۵۶۱	۰/۰۳۰۴	۰/۴۶۵۳	۰/۳۴۴۹	۰/۳۰۳۶	۰/۲۸۴۷
	St	۰/۰۵۷۴	۰/۰۳۱۱	۰/۴۴۳۳	۰/۰۵۶۵	۰/۱۸۹۵	۰/۲۷۷۲
	HS	۰/۰۵۴۱	۰/۰۲۹۵	۰/۴۶۲۵	۰/۰۵۳۹	۰/۰۲۸۷	۰/۲۷۹۲
۲۰۰	نرمال	۰/۰۳۴۷	۰/۰۱۸۵	۰/۱۹۶۵	۰/۳۴۶۲	۰/۳۳۴۰	۰/۱۱۰۹
	St	۰/۰۳۴۶	۰/۰۱۹۷	۰/۱۸۱۴	۰/۰۴۹۹	۰/۲۰۲۷	۰/۱۰۷۵
	HS	۰/۰۳۲۳	۰/۰۱۶۸	۰/۱۸۴۹	۰/۰۴۴۰	۰/۰۳۷۲	۰/۱۰۳۷
۵۰۰	نرمال	۰/۰۲۱۸	۰/۰۱۱۳	۰/۰۷۰۵	۰/۳۵۹۴	۰/۳۷۳۳	۰/۰۴۹۶
	St	۰/۰۲۴۴	۰/۰۱۰۹	۰/۰۶۳۴	۰/۰۶۰۰	۰/۲۲۲۱	۰/۰۴۵۲
	HS	۰/۰۲۸۵	۰/۰۱۲۱	۰/۰۷۳۲	۰/۰۴۳۶	۰/۰۵۵۰	۰/۰۴۱۷

شبیه سازی شده با حجم نمونه ۵۰۰ نمودارهای چندک-چندک^{۶۹} باقی مانده های حاصل از هر سه مدل در شکل ۱۱.۲ نمایش داده شده اند. نمودارهای سمت راست در شکل ۱۱.۲ به نمودار کرمی^{۷۰} معروف هستند که اولین بار توسط ون بورن و فردریکس^{۷۱} (۲۰۰۱) معرفی

^{۶۷}Integrated Mean Squared Errors

^{۶۸}Sacks

^{۶۹}Quantile-Quantile plot

^{۷۰}Worm plot

^{۷۱}Buuren and Fredriks

شدند. نمودار کرمی نوعی نمودار چندک-چندک باقی‌مانده‌های روندزدایی شده^{۷۲} است که نشان‌دهنده فاصله توزیع متغیر پاسخ مشاهده شده از توزیع مفروض، به عنوان توزیع واقعی پاسخ، است. در صورت نیکویی برازش توزیع مفروض به داده‌ها، نقاط روندزدایی شده می‌توانند در یک کران محدود شده پراکنش داشته باشند. انحنای پراکنش نقاط به صورت درجه دو یا سه و همچنین خارج از کران قرار گرفتن نقاط، نشان از عدم کفایت مدل برازش شده دارد. برای توضیح بیشتر در مورد تفسیر این نوع نمودار می‌توانید ون بورن و فردریکس (۲۰۰۱) را ببینید.

با توجه به شکل ۱۱.۲ واضح است که مدل مکان-مقیاس نرمال برای این داده‌ها برازش نامناسبی دارند و در مقابل هر دو مدل HS و t استیودنت برازشی کاملاً مناسب و قابل قبول دارند.

شکل‌های ۱۲.۲ تا ۱۴.۲ نیز اثرات برآورد شده خطی و غیرخطی را برای هر دو پارامتر مکان و مقیاس در سه مدل نرمال، t استیودنت و HS نشان می‌دهد. به همراه برآورد اثرات، کران‌های اطمینان ۹۵ درصد نیز رسم شده‌اند که بر اساس توزیع مجانبی نرمال برآوردگرهای ML محاسبه شده‌اند (ریگی و استاسینوپولوس (۲۰۰۵)). در برآورد ضرایب خطی، هر سه مدل تقریباً یکسان عمل کرده‌اند. اما با توجه به کران‌های اطمینان محاسبه شده، مدل HS نسبت به دو مدل دیگر در برآورد اثرات غیرخطی کارتر عمل کرده است.

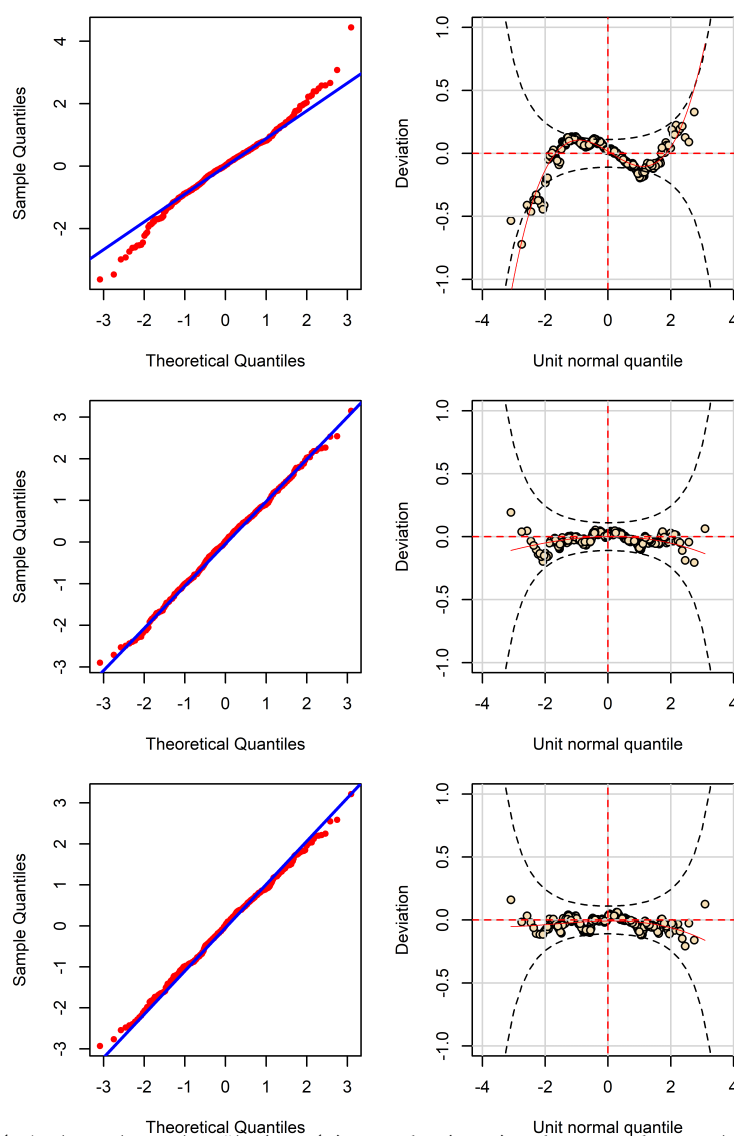
۳.۳.۲ مثال واقعی

برای نمایش کاربرست و برتری ممکن مدل پیشنهادی جدید، تحلیل رگرسیونی یک مجموعه داده واقعی را که در داگلاس و دلامپادی (۱۹۹۰) گزارش شده است، مد نظر قرار دادیم. هدف از تحلیل این داده‌ها بررسی اسیدی بودن دریاچه‌های واقع در رشته‌کوه تیغه آبی در کشور آمریکا است. این مجموعه داده از ۱۱۲ مشاهده برای ۴ متغیر تشکیل شده است. متغیر پاسخ مقدار pH دریاچه‌ها (y) و متغیرهای تبیینی رگرسیونی مختصات مرکزی هر دریاچه، شامل طول جغرافیایی (x_1) و عرض جغرافیایی (x_2)، و لگاریتم غلظت کلسیم با واحد میلی گرم در هر لیتر است (x_3)، که لگاریتم آن در مبنای ۱۰ محاسبه شده است. این مجموعه داده در بسته‌افزار assist در نرم‌افزار R در دسترس است.

شکل ۱۵.۲ نمودارهای پراکنش کناری متغیر پاسخ در مقابل متغیرهای تبیینی را نشان می‌دهد. از روی این نمودارها چند نکته قابل بیان هستند:

۱. نمودار پراکنش پاسخ در مقابل لگاریتم غلظت کلسیم (نمودار سمت راست) یک رابطه مثبت را به همراه نوسان بیشتر برای مقادیر بزرگتر لگاریتم غلظت نشان می‌دهد. این شهود نشانی از برقرار نبودن پذیره هم‌واریانسی در مدل رگرسیون مکانی است.

⁷²Quantile-Quantile plot for detrended residuals

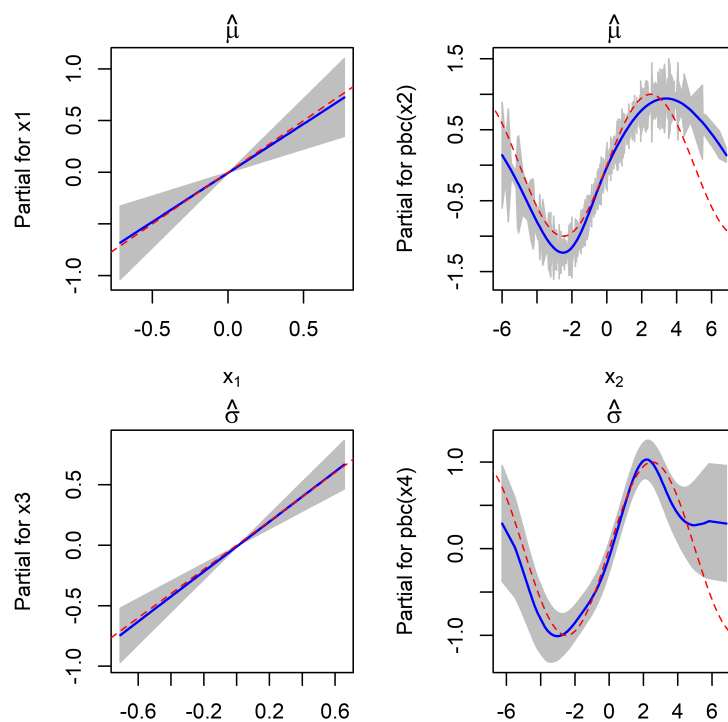


شکل ۱۱.۲: نمودارهای چندک-چندک (چپ) و کرمی (راست) باقی مانده های حاصل از سه مدل نرمال (بالا)، t استیودنت (وسط) و HS (پایین) برای یک مجموعه داده شبیه سازی شده با حجم نمونه ۵۰۰

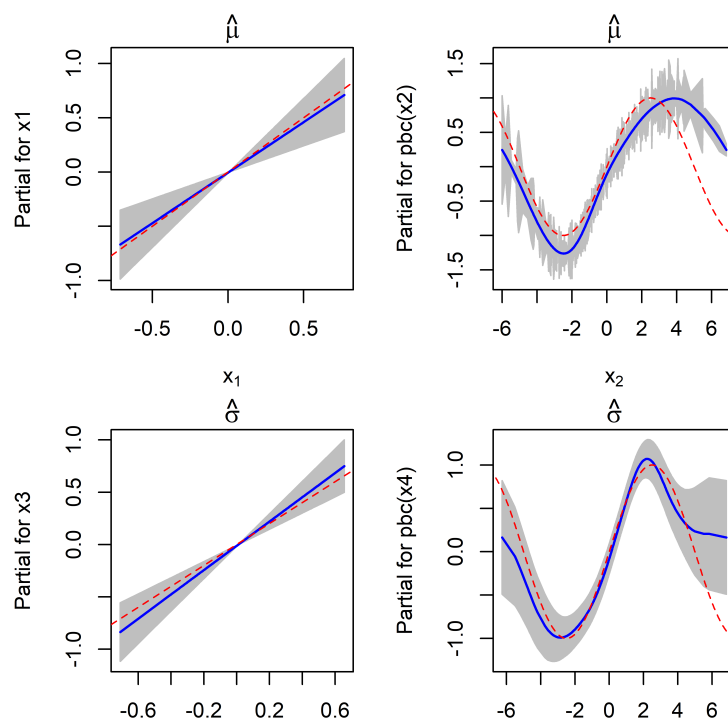
۲. نمودار پراکنش پاسخ در مقابل طول جغرافیایی، یک روند غیرخطی کاهشی و سپس افزایشی را به همراه نوسان ناهمگن در طی این روند نمایش می دهد. در مقابل روند مشخص و قابل درکی در مقابل متغیر عرض جغرافیایی ملموس نیست.

۳. جدا از ناهم واریانسی که در مشاهدات مشهود است، دو داده پرت در هر سه نمودار پراکنش که کاملاً از سایر مشاهدات جدا هستند، دیده می شوند. این مساله بر لزوم لحاظ کردن آنها توسط مدلی که توانایی برخورد با داده های پرت را (که تعداد اندکی از کل مشاهدات را شامل می شوند) دارند، تاکید دارد.

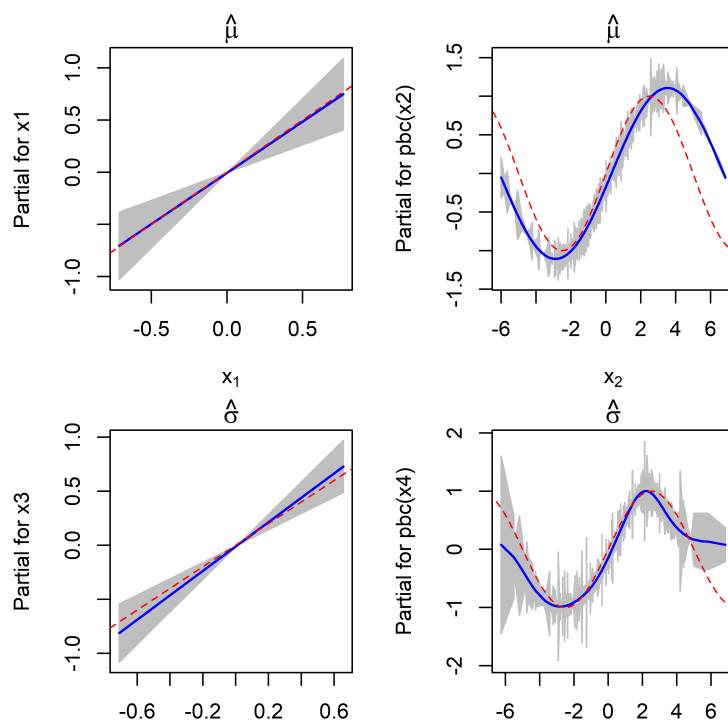
این سه نکته، حاکی از پیچیدگی مدل بندی این داده ها و کافی نبودن مدل میانگین است. برای مدل بندی داده ها دو مدل میانگین (فقط مکان) (مدل (۷.۲)) و مکان-مقیاس (مدل



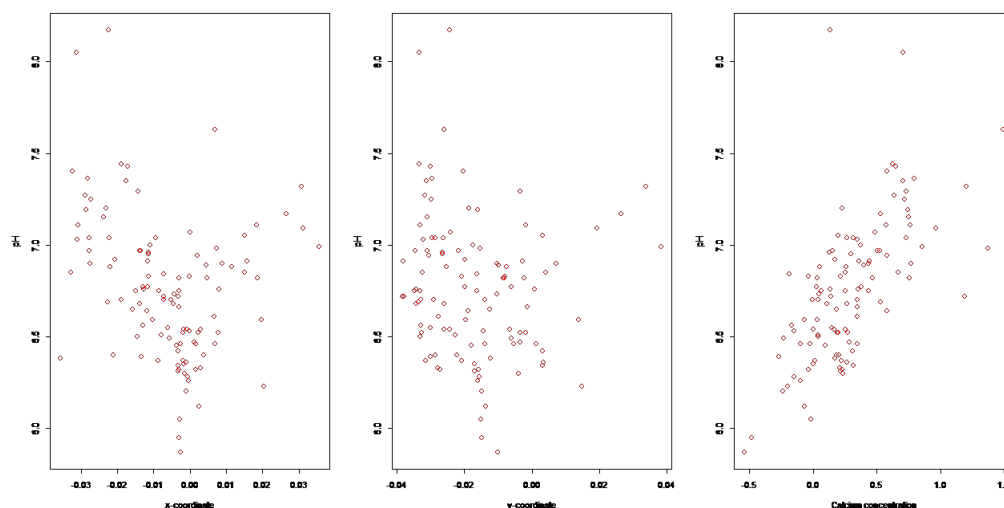
شکل ۱۲.۲: اثرات خطی و غیرخطی برآوردشده (ممتد) و واقعی (خط چین) پارامترهای مکان و مقیاس مدل نرمال برای حجم نمونه ۵۰۰



شکل ۱۳.۲: اثرات خطی و غیرخطی برآوردشده (ممتد) و واقعی (خط چین) پارامترهای مکان و مقیاس مدل، t استیودنت برای حجم نمونه ۵۰۰



شکل ۱۴.۲: اثرات خطی و غیرخطی برآوردشده (ممتد) و واقعی (خط چین) پارامترهای مکان و مقیاس مدل HS برای حجم نمونه ۵۰۰



شکل ۱۵.۲: نمودارهای پراکنش کناری پاسخ و متغیرهای تبیینی برای داده های دریاچه های اسیدی

(۸.۲) را با تنها دو متغیر تبیینی x_1 و x_3 به صورت زیر در نظر گرفتیم:

$$\begin{cases} \mu_i = \beta_0 + h_1(x_1) + h_2(x_3) \\ \log \sigma_i = \beta_1 \end{cases} \quad (۷.۲)$$

$$\begin{cases} \mu_i = \beta_2 + h_1(x_1) + h_2(x_3) \\ \log \sigma_i = \gamma_0 + \gamma_1 x_1 + h_3(x_3) \end{cases} \quad (۸.۲)$$

در این جا نیز، با توجه به حضور دو داده پرت، برای مقایسه مدل پیشنهادی ما با مدلی که قادر به مدل بندی داده های پرت باشد، از مدل مکان و مکان-مقیاس t استیودنت در کنار مدل نرمال استفاده کردیم. نتایج برازش هر سه مدل در جدول ۱۲.۲ گزارش شده اند. از نظر معیارهای AIC و BIC می توان گفت مدل مکان (۷.۲) نسبت به مدل مکان-مقیاس (۸.۲) ضعیف تر است. در هر دو نوع مدل بندی نیز مدل پیشنهادی حتی نسبت به مدل استوار t استیودنت برتر است و می توان نتیجه گرفت که در حضور داده های با دم نیمه سنگین مدل های دم سنگینی مثل t استیودنت به خوبی مدل پیشنهادی نخواهند بود.

نمودارهای نیکویی برازش مبتنی بر باقی مانده ها برای هر سه مدل نیز در شکل ۱۶.۲ رسم شده اند که نشان می دهند مدل نرمال از دو مدل دیگر از کارایی کمتری برخوردار است. نیکویی برازش مدل های HS و t استیودنت به ویژه برای مدل (۸.۲) قابل تایید است. منحنی های

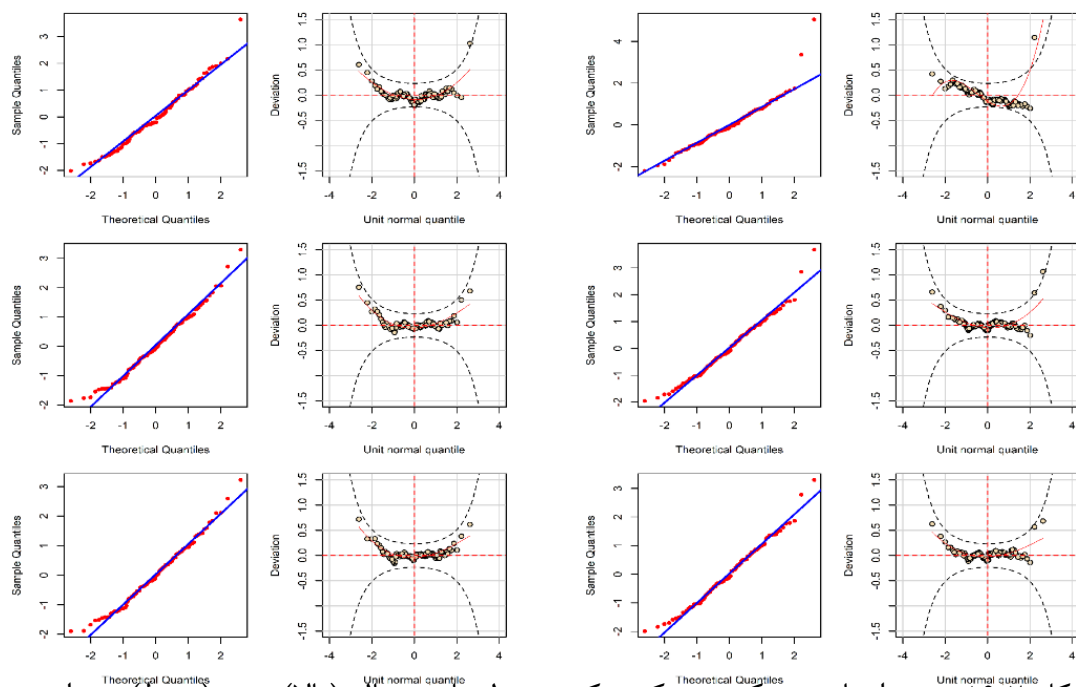
جدول ۱۲.۲: نتایج برازش مدل های (۷.۲) و (۸.۲) بر روی داده های دریاچه های اسیدی

مدل	توزیع	β_0	β_1	β_2	γ_0	γ_1	ν	AIC	BIC
نرمال		۶/۵۰۵	-۱/۴۰۴	-	-	-	-	۲۳/۲۷	۵۰/۴۶
مدل (۷.۲)	St	۶/۴۹۴	-۱/۶۹۵	-	-	-	۱/۵۸۱	۸/۳۰	۳۸/۲۱
	HS	۶/۵۵۲	-۱/۸۹۱	-	-	-	-	۷/۰۴	۳۴/۲۳
نرمال		-	-	۶/۴۹۹	-۱/۶۳۳	-۲۴/۶۵۰	-	۹/۱۳	۴۹/۹۱
مدل (۸.۲)	St	-	-	۶/۴۷۸	-۱/۸۹۸	-۱۹/۵۳۸	۱/۶۶۱	۱۱/۳۵	۵۴/۸۵
	HS	-	-	۶/۳۸۶	-۱/۸۰۵	۳/۳۷۴	-	۶/۹۴	۴۷/۷۲

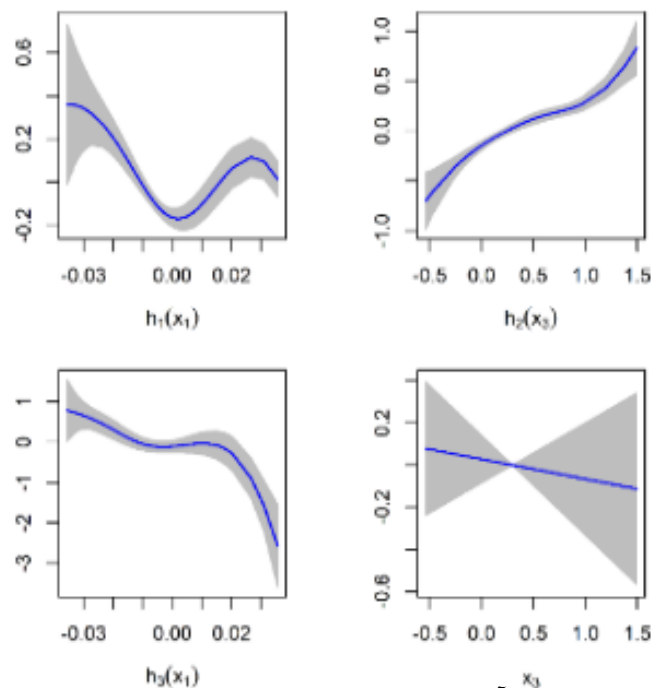
اثرات برآورد شده خطی و غیرخطی برای هر دو پارامتر مکان و مقیاس تحت سه توزیع نرمال، t استیودنت و HS برای مدل (۸.۲) در شکل های ۱۷.۲ تا ۱۹.۲ رسم شده اند. اثر برآورد شده برای متغیر طول جغرافیایی در مقیاس مدل نرمال کاملاً در تناقض با دو مدل دیگر است؛ در واقع شیب اثر در مدل نرمال منفی ولی در دو مدل دیگر مثبت است. این اتفاق منجر به یک نتیجه گیری نادرست خواهد شد. سایر اثرها در هر سه مدل برای هر دو پارامتر مکان و مقیاس تقریباً شبیه به هم برآورد شده اند، با این تفاوت که عدم قطعیت برآورد شده برای متغیر تبیینی لگاریتم غلظت کلسیم در کران های مشاهدات در پارامتر مکان مدل HS از دو مدل رقیب بیشتر است.

۴.۲ توزیع های پربولیک سکانت چندمتغیره

از ابزارهای مهم در بسیاری از عرصه های مطالعاتی مانند محیط زیست، فیزیک، زمین شناسی، اقتصاد و معدن، توزیع های چندمتغیره است. به همین دلیل، توزیع های آماری یک متغیره توسط آماردانان به نسخه های چندمتغیره گسترش داده می شوند. برای مثال، نسخه های



شکل ۱۶.۲: نمودارهای چند-چندک و کرمی مدل های نرمال (بالا)، HS (وسط) و t استیودنت (پایین) برای مدل های (۷.۲) (الف) و (۸.۲) (ب)

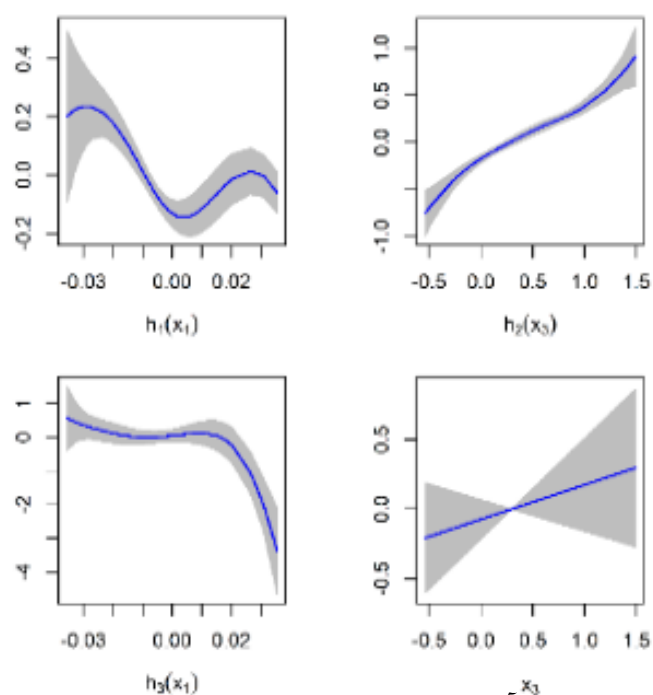


شکل ۱۷.۲: منحنی های اثرات برآورد شده پارامترهای مکان (بالا) و مقیاس (پایین) در مدل نرمال برای مدل (۷.۲)

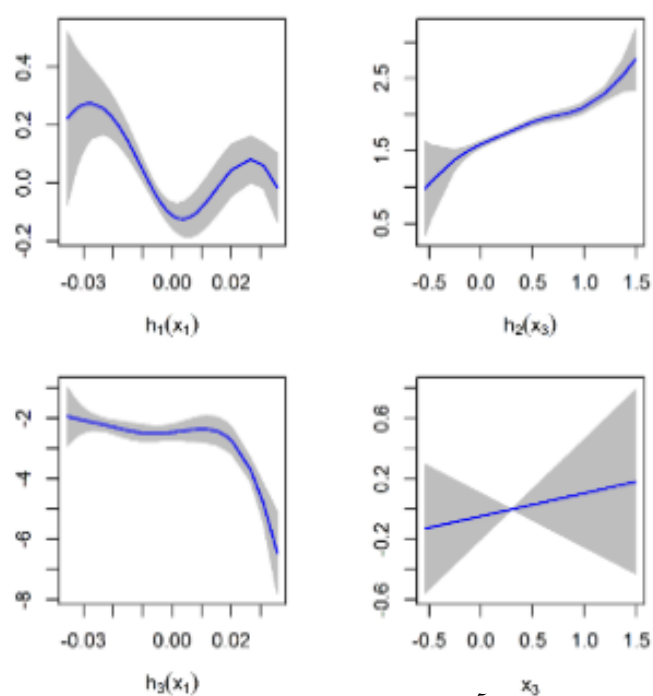
چندمتغیره مختلفی از توزیع SN توسط آزالینی و وال^{۷۳} (۱۹۹۶)، آزالینی (۲۰۰۵) و جو^{۷۴} و

⁷³Valle

⁷⁴Joe



شکل ۱۸.۲: منحنی‌های اثرات برآوردشده پارامترهای مکان (بالا) و مقیاس (پایین) در مدل HS برای مدل (۷.۲)



شکل ۱۹.۲: منحنی‌های اثرات برآوردشده پارامترهای مکان (بالا) و مقیاس (پایین) در مدل t استیودنت برای مدل (۷.۲)

همکاران (۲۰۱۲) گسترش داده شده‌اند. اندرسون^{۷۵} (۱۹۹۲) توزیع لاپلاس چندمتغیره را

⁷⁵Anderson

به عنوان حالت خاص از توزیع لنینیک چندمتغیره^{۷۶} توسعه داده است. التفت^{۷۷} و همکاران (۲۰۰۶) توزیع لاپلاس چندمتغیره را به عنوان آمیخته مقیاس توزیع نرمال بحث کرده اند. توزیع لجستیک چندمتغیره توسط مالیک و آبراهام^{۷۸} (۱۹۷۳) معرفی شده است. یه^{۷۹} (۲۰۱۰) توزیع شبه لجستیک چندمتغیره را مطالعه کرده است. توزیع هایپربولیک تعمیم یافته چوله نرمال^{۸۰} چندمتغیره توسط ویلکا^{۸۱} و همکاران (۲۰۱۴) بررسی شده است. خانواده تولید شده گاما^{۸۲} چندمتغیره نیز توسط بالاکریشنان و دیستیک (۲۰۱۶) مورد مطالعه قرار گرفته است. در این بخش توزیع یک متغیره HS را به حالت چندمتغیره گسترش می دهیم.

همان طور که در بخش ۱.۱.۲ گفته شد توزیع HS ساختارش به این صورت است که اگر V توزیع نیمه کوشی باشد، $U = \log(V)$ از توزیع HS پیروی خواهد کرد. با استفاده از این حقیقت می توان، توزیع HS را به حالت چندمتغیره آن تعمیم داد. فرض کنید $Y = (Y_1, \dots, Y_p)'$ از توزیع نیمه کوشی $-p$ متغیره تبعیت می کند. تابع چگالی بردار تصادفی Y به صورت زیر است:

$$g_Y(y) = \frac{2^p \Gamma\left(\frac{p+1}{2}\right)}{\pi^{\frac{p+1}{2}} (1 + y'y)^{\frac{p+1}{2}}}, \quad y_i > 0, \quad i = 1, \dots, p,$$

تابع چگالی بردار تصادفی $Z = \log(Y)$ به صورت زیر به دست می آید:

$$f_Z(z) = \frac{2^p \Gamma\left(\frac{p+1}{2}\right) \exp(V'z)}{\pi^{\frac{p+1}{2}} (1 + \exp(2z))^{p+1}}, \quad z \in \mathbb{R}^p, \quad (9.2)$$

که در آن $\log(Y) = (\log(Y_1), \dots, \log(Y_p))$ و $\exp(2z) = \sum_{k=1}^p \exp(2z_k)$. اگر بردار تصادفی Z دارای تابع چگالی (۹.۲) باشد، گوئیم Z دارای توزیع هایپربولیک سکانت چندمتغیره^{۸۳} (MHS) است و با نماد $Z \sim \text{MHS}(0, I)$ نمایش می دهیم. با اضافه کردن پارامترهای بردار مکان μ و ماتریس مقیاس Σ ، به راحتی می توان توزیع MHS را با افزودن دو پارامتر بردار مکان μ و ماتریس مقیاس Σ بسیار منعطف تر ساخت. در این صورت توزیع با نماد $\text{MHS}(\mu, \Sigma)$ نمایش داده می شود. می توان نوشت

$$X = \frac{2}{\pi} \Sigma_z^{-\frac{1}{2}} \Sigma_z^{-\frac{1}{2}} Z + \mu \quad (10.2)$$

که در آن Σ_z یک ماتریس $p \times p$ با اعضای قطر اصلی برابر ۱ و اعضای غیر قطر اصلی ۰ است.

⁷⁶Multivariate Linnik

⁷⁷Eltoft

⁷⁸Malik and Abraham

⁷⁹Yeh

⁸⁰Skew Normal Generalized Hyperbolic

⁸¹Vaika

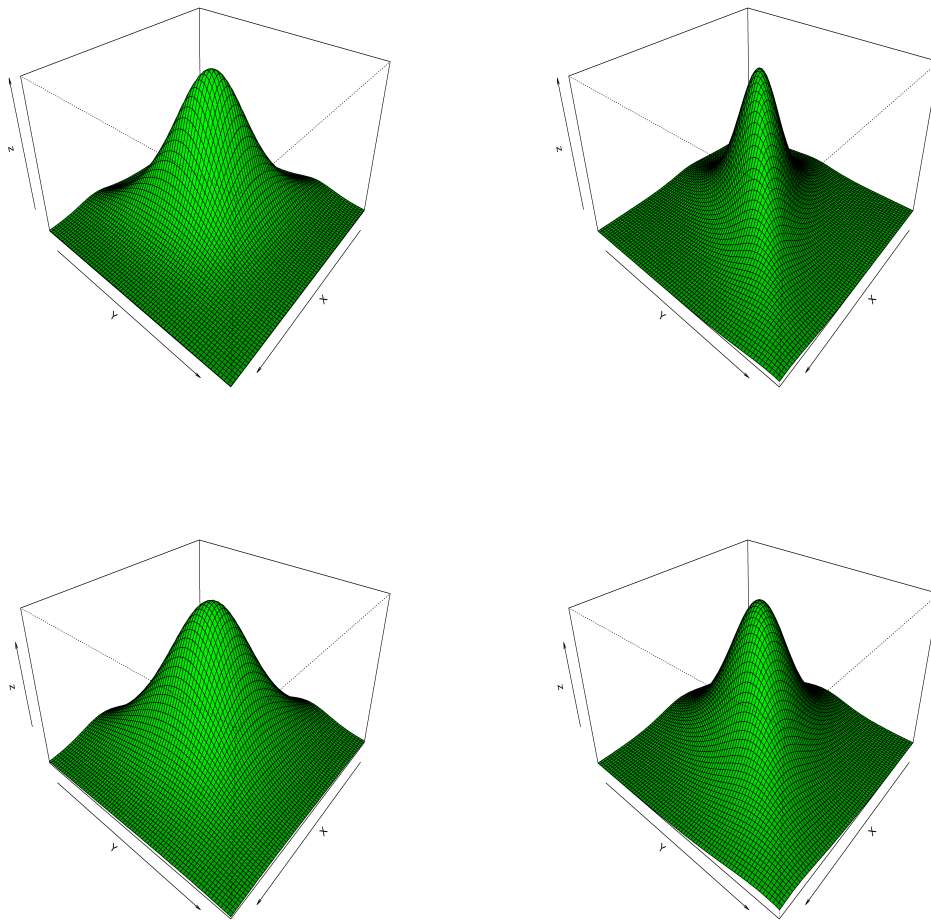
⁸²Gamma-generated

⁸³Multivariate Hyperbolic Secant

تابع چگالی بردار تصادفی X به صورت زیر داده می شود:

$$f_X(\mathbf{x}) = \frac{|\Sigma_z|^{\frac{1}{2}} |\Sigma|^{-\frac{1}{2}} \pi^{\frac{p-1}{2}} \Gamma\left(\frac{p+1}{2}\right) \exp\left(\frac{\pi}{2} \mathbf{1}' \Sigma_z^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} (\mathbf{x} - \mu)\right)}{\left(1 + \sum_{k=1}^p \exp\left(\frac{\pi}{2} e_k' \Sigma_z^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} (\mathbf{x} - \mu)\right)\right)^{\frac{p+1}{2}}}, \quad \mathbf{x} \in \mathbb{R}^p, \quad (11.2)$$

که در آن e_i برای $i = 1, \dots, p$ برداری با اندازه p از صفرها است که فقط i -امین مؤلفه آن برابر ۱ است. منحنی های MHS دومتغیره برای برخی مقادیر پارامترهای μ و Σ در شکل ۲۰.۲ نمایش داده شده اند.



شکل ۲۰.۲: منحنی های توزیع دومتغیره برای مقادیر مختلف پارامترها. $\text{vech}(\Sigma) = (1, 0.5, 1)'$ (بالا سمت راست)، $\text{vech}(\Sigma) = (1, -0.5, 1)'$ (بالا سمت چپ)، $\text{vech}(\Sigma) = (1/5, 0.5, 1/5)'$ (پایین سمت راست)، $\text{vech}(\Sigma) = (1/5, -0.5, 1/5)'$ (پایین سمت چپ) و در همه نمودارها بردار پارامتر $\mu = (0, 0)'$ است.

۱.۴.۲ تابع مولد گشتاور توزیع MHS

تابع مولدگشتاور $Z \sim \text{MHS}(\circ, I)$ به صورت زیر به دست می آید:

$$M_Z(t) = \pi^{-\frac{p+1}{2}} \prod_{k=1}^p \Gamma\left(\frac{1+t_k}{2}\right) \Gamma\left(\frac{1-\sum_{k=1}^p t_k}{2}\right). \quad (12.2)$$

بنابراین با استفاده از آن می توان نتیجه گرفت

$$\begin{aligned} E(Z) &= \circ \\ \text{Cov}(Z) &= \frac{\pi^2}{4} \Sigma_Z. \end{aligned}$$

اکنون با تبدیل (۱۰.۲) داریم

$$\begin{aligned} E(\mathbf{X}) &= \mu \\ \text{Cov}(\mathbf{X}) &= \Sigma \end{aligned}$$

این یعنی پارامترهای μ و Σ همان بردار میانگین و ماتریس کوواریانس توزیع هستند.

۲.۴.۲ برآورد پارامترهای توزیع MHS

فرض کنید $\mathbf{x}_i$ برای $i = 1, 2, \dots, n$ مشاهده مستقل از توزیع MHS با پارامترهای نامعلوم μ و Σ باشند. در این بخش دو روش برآورد گشتاوری (MME) و برآورد ماکسیمم درستنمایی (MLE) را مورد بررسی قرار می دهیم.

روش برآورد گشتاوری

برآوردگرهای MM پارامترهای μ و Σ برابر هستند با:

$$\begin{aligned} \tilde{\mu} &= \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \\ \tilde{\Sigma} &= \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \tilde{\mu})(\mathbf{x}_i - \tilde{\mu})'. \end{aligned}$$

برآوردگرهای ماکسیمم درستنمایی

تابع لگ درستنمایی توزیع MHS متناسب است با

$$\begin{aligned} \ell(\mu, \Sigma) &\propto -\frac{n}{2} \log(|\Sigma|) + \frac{\pi}{2} \sum_{i=1}^n \left(\mathbf{1}' \Sigma_Z^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} (\mathbf{x}_i - \mu) \right) \\ &\quad - \frac{p+1}{2} \sum_{i=1}^n \log \left(1 + \sum_{k=1}^p \exp \left(2 e_k' \frac{\pi}{2} \Sigma_Z^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} (\mathbf{x}_i - \mu) \right) \right). \end{aligned}$$

بنابراین اعضای بردار امتیاز $S_n = \left(\frac{\partial \ell(\mu, \Sigma)}{\partial \mu}, \frac{\partial \ell(\mu, \Sigma)}{\partial \Sigma} \right)'$ به صورت زیر نتیجه می شوند:

$$\begin{aligned} \frac{\partial \ell(\mu, \Sigma)}{\partial \mu} &= \pi \frac{p+1}{2} \sum_{i=1}^n \frac{\sum_{k=1}^p \left(\Sigma^{-\frac{1}{2}} \Sigma_z^{\frac{1}{2}} e_k \right) \exp \left(\pi e_k' \Sigma_z^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} (x_i - \mu) \right)}{1 + \sum_{k=1}^p \exp \left(\pi e_k' \Sigma_z^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} (x_i - \mu) \right)} - \frac{n\pi}{2} \left(\Sigma^{-\frac{1}{2}} \Sigma_z^{\frac{1}{2}} \mathbf{1} \right) \\ \frac{\partial \ell(\mu, \Sigma)}{\partial \Sigma} &= \pi \frac{p+1}{2} \sum_{i=1}^n \frac{\sum_{k=1}^p \left(\Sigma^{-\frac{1}{2}} (x_i - \mu) e_k' \Sigma^{-\frac{1}{2}} \Sigma_z^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right) \exp \left(\pi e_k' \Sigma_z^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} (x_i - \mu) \right)}{1 + \sum_{k=1}^p \exp \left(\pi e_k' \Sigma_z^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} (x_i - \mu) \right)} \\ &\quad - \frac{n}{2} \Sigma^{-1} - \frac{\pi}{2} \sum_{i=1}^n \left(\Sigma^{-\frac{1}{2}} (x_i - \mu) \mathbf{1}' \Sigma^{-\frac{1}{2}} \Sigma_z^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right). \end{aligned}$$

برای محاسبه MLE نیاز داریم تا دستگاه معادلات $S_n = 0$ را حل کنیم. حل این دستگاه جواب بسته ندارد و باید از روش های عددی مانند نیوتن رافسون استفاده کنیم.

۳.۴.۲ کاربرد توزیع MHS

مطالعه شبیه سازی

برای ارزیابی عملکرد MMEs و MLEs پارامترهای توزیع MHS، مطالعه شبیه سازی شده ای را برای نسخه دومتغیره MHS، یعنی MHS₂، طراحی کردیم. حجم نمونه هایی که برای این مطالعه در نظر گرفتیم عبارتند از $n = 50, 100, 200, 500, 1000$. مقادیر واقعی $\mu_1 = (0.5, 1)'$ و Σ را در نظر گرفتیم. برای هر حالت شبیه سازی نیز 1000 تکرار اجرا کردیم. ارزیابی و MSE برآوردهای ML و MM در جدول ۱۳.۲ تا ۱۶.۲ گزارش شده اند. با توجه به جدول های ۱۳.۲ تا ۱۶.۲ نتیجه می گیریم که ارزیابی و MSE هر دو MME و MLE با افزایش حجم نمونه کاهش پیدا می کنند. و همچنین برای حجم نمونه های مختلف، MLE کارایی بالاتری نسبت به MME دارد.

مطالعه مجموعه داده های واقعی

در این قسمت کاربرد توزیع MHS را برای برازش داده های واقعی بررسی می کنیم. دو مجموعه داده واقعی را مورد مطالعه قرار می دهیم و توزیع های MHS و نرمال دومتغیره^{۸۴} (BN) را بر روی این داده ها برازش می دهیم. معیار مقایسه مورد استفاده در این قسمت AIC است.

- **مجموعه داده اول:** این مجموعه داده شامل ۱۱۰ مشاهده از بازده ماهانه تاکستان های شرکت نوشیدنی شیلیایی^{۸۵} است که از مارس ۱۹۹۰ تا آوریل ۱۹۹۹ گردآوری شده اند و

⁸⁴Bivariate Normal

⁸⁵Conchha y Toro

جدول ۱۳.۲: آریبی و MSE برآوردهای ML و MM تحت مدل $MHS_2(\mu_1, \Sigma_1)$

n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{12}$	$\hat{\sigma}_{22}$	
۵۰	MLE	آریبی	-۰/۰۰۰۱۲	۰/۰۰۰۱۲	-۰/۰۰۰۴۰	
	MSE		۰/۰۰۱۷۰	۰/۰۲۳۵	۰/۱۲۳۴	
	MME	آریبی	۰/۰۰۰۱۷	۰/۰۰۰۵۲	۰/۰۱۱۲	
	MSE		۰/۰۲۰۱	۰/۰۲۹۴	۰/۱۸۳۴	
۱۰۰	MLE	آریبی	-۰/۰۰۰۱۳	-۰/۰۰۰۴۹	۰/۰۰۰۷۷	
	MSE		۰/۰۰۰۸۱	۰/۰۱۲۰	۰/۰۶۲۸	
	MME	آریبی	-۰/۰۰۰۲۵	-۰/۰۰۰۴۰	۰/۰۰۰۷۱	
	MSE		۰/۰۱۰۰	۰/۰۱۳۸	۰/۰۹۳۰	
۲۰۰	MLE	آریبی	-۰/۰۰۰۱۵	-۰/۰۰۰۳۸	-۰/۰۰۰۸۹	
	MSE		۰/۰۰۰۴۰	۰/۰۰۰۶۰	۰/۰۲۹۸	
	MME	آریبی	-۰/۰۰۰۲۶	-۰/۰۰۰۷۱	-۰/۰۰۰۱۸	
	MSE		۰/۰۰۰۴۹	۰/۰۰۰۷۴	۰/۰۴۵۴	
۵۰۰	MLE	آریبی	۰/۰۰۰۰۷	-۰/۰۰۰۰۲	-۰/۰۰۰۴۳	
	MSE		۰/۰۰۰۱۷	۰/۰۰۰۲۶	۰/۰۱۲۹	
	MME	آریبی	۰/۰۰۰۰۷	۰/۰۰۰۰۴	-۰/۰۰۰۵۴	
	MSE		۰/۰۰۰۲۰	۰/۰۰۰۳۱	۰/۰۱۷۶	
۱۰۰۰	MLE	آریبی	-۰/۰۰۰۱۹	۰/۰۰۰۰۴	-۰/۰۰۰۲۰	
	MSE		۰/۰۰۰۰۸	۰/۰۰۰۱۲	۰/۰۰۰۶۰	
	MME	آریبی	-۰/۰۰۰۲۶	-۰/۰۰۰۰۷	-۰/۰۰۰۰۶	
	MSE		۰/۰۰۰۱۰	۰/۰۰۰۱۵	۰/۰۰۰۸۷	

در بسته heay در نرم افزار R موجود هستند. این مجموعه داده شامل متغیرهای بازده شاخص بازار شیلایی (IPSA) و بازده ماهانه شرکت CYT هستند.

- **مجموعه داده دوم:** مجموعه داده واقعی دوم، یعنی داده های شرکت مارتین ماریتا، در بخش ۳.۲.۲ معرفی شده است. در این مثال برای متغیرهای m.marietta و CRSP هر دو توزیع MHS و نرمال دومتغیره را برازش دادیم.

برای هر دو مجموعه داده مقادیر برآورده شده پارامترها و معیار AIC در جدول ۱۷.۲ و ۱۸.۲ گزارش شده اند. براساس جداول ۱۷.۲ و ۱۸.۲ برای هر دو مجموعه داده طبق معیار AIC، توزیع MHS دومتغیره از توزیع نرمال دومتغیره برتر است.

جدول ۱۴.۲: آریبی و MSE برآوردگرهای ML و MM تحت مدل $MHS_2(\mu_1, \Sigma_2)$

n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{12}$	$\hat{\sigma}_{22}$		
۵۰	۰/۰۰۲۱	۰/۰۰۳۹	-۰/۰۱۶۵	۰/۰۰۵۹	-۰/۰۲۰۱	MLE	آریبی
	۰/۰۱۳۸	۰/۰۱۹۸	۰/۰۳۴۷	۰/۰۲۸۴	۰/۰۷۲۳	MSE	
	۰/۰۰۱۶	۰/۰۰۵۵	-۰/۰۰۸۳	۰/۰۰۱۲	-۰/۰۰۷۶	MME	آریبی
	۰/۰۱۶۲	۰/۰۲۲۹	۰/۰۵۰۱	۰/۰۳۹۱	۰/۱۰۹۰	MSE	
۱۰۰	-۰/۰۰۲۴	۰/۰۰۱۳	۰/۰۰۱۰	-۰/۰۰۳۳	-۰/۰۰۶۸	MLE	آریبی
	۰/۰۰۶۵	۰/۰۱۰۰	۰/۰۱۸۵	۰/۰۱۳۸	۰/۰۴۰۳	MSE	
	-۰/۰۰۲۲	۰/۰۰۰۶	۰/۰۰۲۶	-۰/۰۰۲۶	-۰/۰۰۲۹	MME	آریبی
	۰/۰۰۷۹	۰/۰۱۱۶	۰/۰۲۷۰	۰/۰۱۹۲	۰/۰۵۹۰	MSE	
۲۰۰	-۰/۰۰۱۶	۰/۰۰۰۴	-۰/۰۰۱۴	۰/۰۰۱۷	-۰/۰۰۶۴	MLE	آریبی
	۰/۰۰۳۶	۰/۰۰۵۰	۰/۰۰۸۵	۰/۰۰۷۰	۰/۰۱۹۵	MSE	
	-۰/۰۰۱۳	۰/۰۰۲۸	-۰/۰۰۲۳	۰/۰۰۴۰	-۰/۰۰۷۸	MME	آریبی
	۰/۰۰۴۳	۰/۰۰۵۹	۰/۰۱۱۸	۰/۰۰۹۸	۰/۰۲۸۸	MSE	
۵۰۰	۰/۰۰۲۱	-۰/۰۰۰۹	-۰/۰۰۰۱	۰/۰۰۰۹	-۰/۰۰۰۴	MLE	آریبی
	۰/۰۰۱۳	۰/۰۰۲۰	۰/۰۰۳۶	۰/۰۰۲۵	۰/۰۰۷۴	MSE	
	۰/۰۰۲۵	-۰/۰۰۱۰	-۰/۰۰۰۶	۰/۰۰۱۵	-۰/۰۰۲۳	MME	آریبی
	۰/۰۰۱۶	۰/۰۰۲۴	۰/۰۰۵۳	۰/۰۰۳۸	۰/۰۱۰۵	MSE	
۱۰۰۰	-۰/۰۰۰۴	-۰/۰۰۱۱	۰/۰۰۱۴	-۰/۰۰۰۳	۰/۰۰۰۳	MLE	آریبی
	۰/۰۰۰۶	۰/۰۰۰۹	۰/۰۰۱۵	۰/۰۰۱۲	۰/۰۰۳۹	MSE	
	۰/۰۰۰۰	-۰/۰۰۲۱	۰/۰۰۱۷	-۰/۰۰۰۹	۰/۰۰۰۸	MME	آریبی
	۰/۰۰۰۸	۰/۰۰۱۱	۰/۰۰۲۲	۰/۰۰۱۸	۰/۰۰۵۹	MSE	

۵.۲ نتیجه‌گیری

در این فصل، توزیع نیمه‌دم‌سنگین HS را به حالت چوله تعمیم داده و ویژگی‌های آنها مانند تابع توزیع، تابع مولد گشتاور و گشتاورهای آنها را به‌دست آوردیم. تعمیم‌های جدید را در قالب توزیع خطا در مدل رگرسیون خطی بیان، و رفتار مجانبی برآوردگرهای ML را بررسی، و نشان دادیم سازگار ضعیف و دارای توزیع مجانبی نرمال هستند. با مثال‌های شبیه‌سازی و واقعی نشان دادیم در موقعیت‌هایی که داده‌ها نیمه‌دم‌سنگین هستند، مدل پیشنهادی کارایی بهتری نسبت به توزیع‌های دم‌سبک و دم‌سنگین دارند. همچنین در نرم‌افزار R

۶۴ توزیع هایپربولیک سکانت و تعمیم‌های آن

جدول ۱۵.۲: اربیی و MSE برآوردهای ML و MM تحت مدل $MHS_2(\mu_2, \Sigma_1)$

n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{12}$	$\hat{\sigma}_{22}$	
۵۰	۰/۰۰۵۴	۰/۰۰۷۱	-۰/۰۲۱۴	-۰/۰۰۹۹	-۰/۰۲۶۵	MLE اربیی
	۰/۰۱۷۶	۰/۰۲۵۲	۰/۰۵۳۱	۰/۰۴۳۳	۰/۰۱۱۴۲	MSE
	۰/۰۰۴۵	۰/۰۰۷۶	-۰/۰۱۲۶	-۰/۰۰۷۶	-۰/۰۱۲۳	MME اربیی
	۰/۰۲۰۳	۰/۰۲۸۷	۰/۰۷۶۹	۰/۰۵۹۱	۰/۰۱۶۷۲	MSE
۱۰۰	-۰/۰۰۱۹	۰/۰۰۰۹	-۰/۰۰۲۷	-۰/۰۰۷۶	-۰/۰۱۴۱	MLE اربیی
	۰/۰۰۷۹	۰/۰۱۲۲	۰/۰۲۹۴	۰/۰۲۱۴	۰/۰۶۳۳	MSE
	-۰/۰۰۲۴	-۰/۰۰۰۵	۰/۰۰۰۴	-۰/۰۰۴۴	-۰/۰۰۷۲	MME اربیی
	۰/۰۰۹۷	۰/۰۱۴۳	۰/۰۴۴۶	۰/۰۳۲۳	۰/۰۹۴۰	MSE
۲۰۰	-۰/۰۰۱۴	-۰/۰۰۰۱	-۰/۰۰۱۷	-۰/۰۰۰۷	-۰/۰۰۶۷	MLE اربیی
	۰/۰۰۴۴	۰/۰۰۶۰	۰/۰۱۳۵	۰/۰۱۱۲	۰/۰۳۰۲	MSE
	-۰/۰۰۰۴	۰/۰۰۲۶	-۰/۰۰۰۷	۰/۰۰۱۰	-۰/۰۰۷۱	MME اربیی
	۰/۰۰۵۱	۰/۰۰۷۱	۰/۰۱۸۹	۰/۰۱۵۷	۰/۰۴۵۵	MSE
۵۰۰	۰/۰۰۲۲	۰/۰۰۰۰۴	۰/۰۰۰۷	۰/۰۰۱۵	۰/۰۰۰۷	MLE اربیی
	۰/۰۰۱۶	۰/۰۰۲۵	۰/۰۰۵۳	۰/۰۰۳۹	۰/۰۱۱۹	MSE
	۰/۰۰۲۷	۰/۰۰۰۲	۰/۰۰۰۳	۰/۰۰۱۰	-۰/۰۰۱۵	MME اربیی
	۰/۰۰۲۰	۰/۰۰۳۰	۰/۰۰۸۰	۰/۰۰۵۸	۰/۰۱۶۵	MSE
۱۰۰۰	-۰/۰۰۱۰	-۰/۰۰۱۴	۰/۰۰۲۲	۰/۰۰۰۶	-۰/۰۰۰۱	MLE اربیی
	۰/۰۰۰۸	۰/۰۰۱۱	۰/۰۰۲۶	۰/۰۰۲۳	۰/۰۰۶۶	MSE
	-۰/۰۰۰۹	-۰/۰۰۲۶	۰/۰۰۱۷	-۰/۰۰۰۰	۰/۰۰۰۴	MME اربیی
	۰/۰۰۱۰	۰/۰۰۱۴	۰/۰۰۳۸	۰/۰۰۳۲	۰/۰۰۹۹	MSE

بسته‌ای فراهم کردیم که توابع چگالی، توزیع، تولید نمونه و برآورد ML توزیع‌های جدید را محاسبه می‌کند. همچنین یک مدل مکان-مقیاس، بر اساس توزیع هایپربولیک سکانت با دم نیمه‌سنگین، معرفی کردیم که از توانایی بالاتری نسبت به مدل‌های استوار دم‌سنگین در برخورد با داده‌های پرت اندک نسبت به کل داده‌ها برخوردار است. این مدل جدید عضوی از رده مدل‌های GAMLSS محسوب می‌شود و بنابراین از مزایای استنباطی توسعه‌یافته برای این مدل‌ها برخوردار است. برای برآزش مثال‌های شبیه‌سازی و کاربردی، در بخش مدل‌بندی مولفه‌های غیرخطی، از اسپلاین‌های توانیده استفاده کردیم. جانشین‌های مختلف دیگری مانند اسپلاین‌های درجه سوم و چندجمله‌ای‌های موضعی برای این کار وجود دارند که می‌توان

جدول ۱۶.۲: آریبی و MSE برآوردگرهای ML و MM تحت مدل $MHS_2(\mu_2, \Sigma_2)$

n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{12}$	$\hat{\sigma}_{22}$	
۵۰	MLE	آریبی	۰/۰۰۲۱	۰/۰۰۳۹	۰/۰۱۶۵	۰/۰۰۵۹
	MSE	۰/۰۱۳۸	۰/۰۱۹۸	۰/۰۳۴۷	۰/۰۲۸۴	۰/۰۷۲۳
	MME	آریبی	۰/۰۰۱۶	۰/۰۰۵۵	۰/۰۰۸۳	۰/۰۰۷۶
	MSE	۰/۰۱۶۲	۰/۰۲۲۹	۰/۰۵۰۱	۰/۰۳۹۱	۰/۰۱۹۰
۱۰۰	MLE	آریبی	-۰/۰۰۲۴	۰/۰۰۱۳	۰/۰۰۱۰	-۰/۰۰۳۳
	MSE	۰/۰۰۶۵	۰/۰۱۰۰	۰/۰۱۸۵	۰/۰۱۳۸	۰/۰۴۰۳
	MME	آریبی	-۰/۰۰۲۲	۰/۰۰۰۶	۰/۰۰۲۶	-۰/۰۰۲۶
	MSE	۰/۰۰۷۹	۰/۰۱۱۶	۰/۰۲۷۰	۰/۰۱۹۲	۰/۰۵۹۰
۲۰۰	MLE	آریبی	-۰/۰۰۱۶	۰/۰۰۰۴	-۰/۰۰۱۴	۰/۰۰۱۷
	MSE	۰/۰۰۳۶	۰/۰۰۵۰	۰/۰۰۸۵	۰/۰۰۷۰	۰/۰۱۹۵
	MME	آریبی	-۰/۰۰۱۳	۰/۰۰۲۸	-۰/۰۰۲۳	۰/۰۰۴۰
	MSE	۰/۰۰۴۳	۰/۰۰۵۹	۰/۰۱۱۸	۰/۰۰۹۸	۰/۰۲۸۸
۵۰۰	MLE	آریبی	۰/۰۰۲۱	-۰/۰۰۰۹	-۰/۰۰۰۱	۰/۰۰۰۹
	MSE	۰/۰۰۱۳	۰/۰۰۲۰	۰/۰۰۳۶	۰/۰۰۲۵	۰/۰۰۷۴
	MME	آریبی	۰/۰۰۲۵	-۰/۰۰۱۰	-۰/۰۰۰۶	۰/۰۰۱۵
	MSE	۰/۰۰۱۶	۰/۰۰۲۴	۰/۰۰۵۳	۰/۰۰۳۸	۰/۰۱۰۵
۱۰۰۰	MLE	آریبی	-۰/۰۰۰۴	-۰/۰۰۱۱	۰/۰۰۱۴	-۰/۰۰۰۳
	MSE	۰/۰۰۰۶	۰/۰۰۰۹	۰/۰۰۱۵	۰/۰۰۱۲	۰/۰۰۳۹
	MME	آریبی	۰/۰۰۰۰	-۰/۰۰۲۱	۰/۰۰۱۷	-۰/۰۰۰۹
	MSE	۰/۰۰۰۸	۰/۰۰۱۱	۰/۰۰۲۲	۰/۰۰۱۸	۰/۰۰۵۹

جدول ۱۷.۲: نتایج برآورد ML برای مجموعه داده اول

مدل	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{12}$	$\hat{\sigma}_{22}$	AIC
MHS_2	۰/۰۰۷۵	۰/۰۱۱۹	۰/۰۰۷۳	۰/۰۰۳۹	۰/۰۲۰۸	-۳۷۷/۶۲۶۳
BN	۰/۰۱۶۱	۰/۰۲۷۲	۰/۰۰۶۲	۰/۰۰۵۵	۰/۰۲۲۸	-۳۶۷/۰۰۳۰

از قابلیت‌های آنها نیز بهره برد. در هر دو مثال‌های شبیه‌سازی و کاربردی، عملکرد برتر مدل پیشنهادی ما نسبت به مدل‌های رقیب نمایش داده شد. این نتیجه می‌تواند در موقعیت‌های کاربردی حائز اهمیت باشد و مدل پیشنهادی HS را به‌عنوان یک امکان برای استفاده در

جدول ۱۸.۲: نتایج برآورد ML برای مجموعه داده دوم

مدل	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{12}$	$\hat{\sigma}_{22}$	AIC
MHS ₂	-۰/۰۰۳۰	۰/۰۱۳۹	۰/۰۰۲۲	۰/۰۰۳۱	۰/۰۱۳۰	-۳۲۰/۷۳۲۰
BN	۰/۰۰۸۸	۰/۰۱۶۹	۰/۰۰۱۸	۰/۰۰۳۲	۰/۰۱۴۳	-۳۱۵/۰۰۷۸

جعبه ابزار آماردانان قرار دهد. یک نسخه چندمتغیره از توزیع HS را نیز ارایه کردیم و کار بست آن را در مثال های شبیه سازی و واقعی نمایش دادیم.

فصل ۳

توزیع‌های دومدی-یکمدی

در این فصل برای ساختن توزیع‌های چوله و دومدی با انعطاف‌پذیری بالا روش‌هایی را به همراه ویژگی‌های آنها بیان می‌کنیم. همچنین چند مورد از تعمیم‌های ساخته‌شده را مورد بررسی قرار می‌دهیم. برهان تمام نتایج نظری این فصل در پیوست رساله آمده است.

۱.۳ مقدمه

در سال‌های اخیر تعمیم‌های زیادی برای توزیع‌های آماری توسعه داده شده‌اند که دارای انعطاف‌پذیری بالایی هستند. به عنوان مثال، می‌توان خانواده توزیع‌های کوماراسوامی^۱- G ارایه شده توسط کوردیرو و کاسترو^۲ (۲۰۱۱)، خانواده وایبل^۳- G ارایه‌شده توسط الزاتره^۴ و همکاران (۲۰۱۳)، خانواده پواسون نمایی هندسی^۵- G معرفی‌شده توسط ناداراجا^۶ و همکاران (۲۰۱۳)، توزیع نیم-لجستیک نمایی^۷ کوردیرو و همکاران (۲۰۱۴)، توزیع نرمال-چوله

¹Kumaraswamy

²Castro

³Weibull

⁴Alzaatreh

⁵Geometric Exponential Poisson

⁶Nadarajah

⁷Exponentiated Half-Logistic

نمایی بریده‌شده^۸-G ناداراجا و همکاران (۲۰۱۴)، خانواده پواسون نمایی^۹-G رستیک^{۱۰} و ناداراجا (۲۰۱۴)، توزیع ABPN بولفرونی^{۱۱} و همکاران (۲۰۱۸) و توزیع PLD باکوچ و همکاران (۲۰۱۹) را نام برد.

با توجه به پیچیدگی‌های داده‌ها در دنیای امروز، این تعمیم‌ها قابل برآزش به همه داده‌ها نیستند. ممکن است داده‌هایی با ویژگی‌های دومی و چوله وجود داشته باشند که برای تحلیل آنها تعمیم‌های اشاره‌شده مناسب نخواهند بود. یا در صورت دومی بودن، پارامتری ندارند که ویژگی دومی داده‌ها را کنترل و مثلاً با تغییر در پارامتر خاصی فاصله دو مد را تغییر دهد. در این فصل قضایایی را بیان و برهان می‌کنیم که بتوان به وسیله آنها تعمیم‌هایی را از توزیع‌ها ارائه داد که انعطاف‌پذیری بالایی داشته باشند و با افزودن دو پارامتر شکل، چولگی و دومی بودن یا یک‌مدی بودن و فاصله دو مد را کنترل کند. برای استفاده از این قضایا مثال‌هایی را نیز بیان کرده‌ایم. همچنین توزیع‌های نرمال، t استیودنت و لاپلاس را با این روش تعمیم داده و برای داده‌های چوله و دومی مورد بررسی قرار داده‌ایم.

۲.۳ توسعه خانواده‌های دومی و چوله

۱.۲.۳ روش اول

ابتدا ساز و کار این روش را با یک مثال ساده شرح می‌دهیم. یک روش برای تعمیم توزیع‌ها به توزیع‌های منعطف، ضرب دو تابع چگالی یا توابعی از آنها است. برای مثال، فرض کنید $g_1(x)$ تابع چگالی لاپلاس استاندارد و $g_2(x-k)$ تابع چگالی کوشی با پارامتر مکان k و مقیاس ۱ باشد. در این صورت

$$f(x) = \frac{1}{\pi(3+k^2)} \frac{g_1(x)}{g_2(x-k)} \quad (1.3)$$

تعمیمی از توابع چگالی لاپلاس و کوشی است. می‌توان این توزیع را توزیع لاپلاس-کوشی (LC) نام‌گذاری کرد. توزیع LC هنگامی که $k=0$ باشد متقارن و برای $k \neq 0$ نامتقارن است. این تابع چگالی نسبت به توزیع‌های متقارن لاپلاس و کوشی انعطاف‌پذیری بالاتری دارد و پیچیدگی زیادی ایجاد نمی‌کند. منظور از پیچیدگی این است که ویژگی‌های توزیع مانند گشتاورها صورت بسته دارند. مثلاً r -امین گشتاور این توزیع برابر

$$E(X^r) = \frac{2}{3+k^2} \left\{ k^2 E(X_L^r) - 2k E(X_L^{r+1}) + E(X_L^{r+2}) \right\}$$

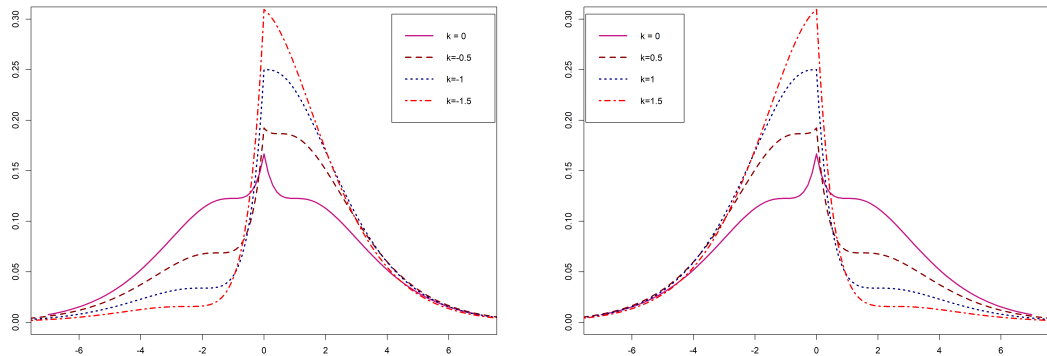
⁸truncated-exponential skew-symmetric

⁹Exponential Poisson

¹⁰Ristic

¹¹ Bolfarine

است، که در آن X_L یک متغیر تصادفی با توزیع لاپلاس استاندارد است. شکل چگالی (۱.۳) برای برخی مقادیر پارامتر k در شکل ۱.۳ نمایش داده شده است. انعطاف‌پذیری بالای این چگالی در شکل ۱.۳ کاملاً مشخص است. قضیه زیر چگونگی ساختن توزیع‌های دومدی



شکل ۱.۳: منحنی چگالی (۱.۳) برای برخی مقادیر k .

مقارن را بیان می‌کند.

قضیه ۱.۲.۳. فرض کنید $g(\cdot)$ تابع چگالی احتمال یک توزیع مقارن حول c با دامنه $\mathbb{R}$ و $w(\cdot)$ تابعی اکیداً مثبت محدب مقارن حول c باشند. تابع

$$f(x) = \frac{w(x)}{E_g[w(X)]} g(x) \quad (2.3)$$

با شرط $E_g[w(X)] < \infty$ یک تابع چگالی احتمال است که $E_g[\cdot]$ امید ریاضی بر حسب تابع چگالی $g(\cdot)$ است و این تابع چگالی

(الف) حول c مقارن است.

(ب) دومدی است.

(پ) مجموع مدهای آن برابر $2c$ است.

□

برهان. برای برهان به پیوست ۲.آ مراجعه کنید.

مثال ۱.۲.۳. فرض کنید $X \sim N(0, 1)$ و $w(x) = \exp(k|x|)$ ، در این صورت

$$f_X(x) = \frac{\exp(k|x|)}{E_\phi[\exp(k|X|)]} \phi(x)$$

حول صفر مقارن و دارای دو مد برابر $\pm k$ است. توجه داشته باشید که در این مثال برای $k > 0$ تابع $w(x)$ محدب است.

در قضیه ۱.۲.۳ اگر تابع محدب $w(\cdot)$ حول $b \neq c$ مقارن باشد، تابع چگالی ایجادشده نامقارن خواهد بود. اگر $w(\cdot)$ مقعر باشد، برای $b = c$ تابع چگالی یک‌مدی مقارن و برای $b \neq c$ تابع چگالی یک‌مدی چوله خواهیم داشت.

به کاربردن قضیه ۱.۲.۳ برای تعمیم توزیع‌ها، نیاز به محاسبه ثابت نرمال‌ساز با محاسبه امیدریاضی $E_g(w(X))$ دارد که ممکن است در برخی موارد سخت و پیچیده باشد. قضیه زیر روشی را برای تعمیم توزیع بیان می‌کند که مشکل محاسبه امیدریاضی و ثابت نرمال‌ساز را ندارد.

قضیه ۲.۲.۳. اگر $g(x)$ یک تابع چگالی اکیداً مثبت متقارن حول صفر با تابع توزیع $G_X(\cdot)$ باشد، آن‌گاه

$$f_X(x) = \frac{k}{2 \left(e^k - e^{\frac{k}{2}} \right)} g(x) e^{kG(|x|)}$$

و

$$f_X(x) = \frac{k+1}{2 \left(1 - \frac{1}{2^{k+1}} \right)} g(x) G(|x|)^k$$

توابع چگالی متقارن دومی برای $k > 0$ هستند.

□

برهان. برای برهان به پیوست ۲.آ مراجعه کنید.

مثال ۲.۲.۳. فرض کنید X دارای توزیع لجستیک استاندارد باشد. تابع چگالی $f_X(x)$ قضیه ۲.۲.۳ به صورت زیر داده می‌شود:

$$f_X(x) = \frac{k}{2 \left(e^k - e^{\frac{k}{2}} \right)} \left\{ \frac{e^{-x}}{(1 + e^{-x})^2} \right\} e^{\left\{ \frac{k}{1 + e^{-|x|}} \right\}} \quad (۳.۳)$$

و

$$f_X(x) = \frac{k+1}{2 \left(1 - \frac{1}{2^{k+1}} \right)} \left\{ \frac{e^{-x}}{(1 + e^{-x})^2} \right\} \left\{ \frac{1}{1 + e^{-|x|}} \right\}^k \quad (۴.۳)$$

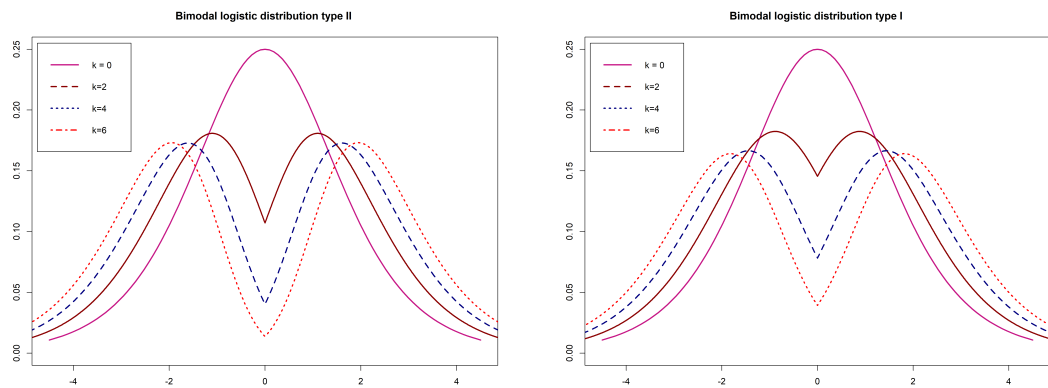
اگر متغیر تصادفی X دارای چگالی‌های (۳.۳) یا (۴.۳) باشد، می‌گوییم دارای توزیع لجستیک دومی نوع اول یا دوم است. شکل توابع چگالی‌های (۳.۳) و (۴.۳) برای برخی مقادیر پارامتر k در شکل ۲.۳ داده شده است.

۲.۲.۳ روش دوم

فرض کنید متغیر تصادفی X دارای تابع چگالی $g(x)$ و $h(\cdot)$ تابع مثبت با شرط $\int_{-\infty}^{\infty} g(h(w)) dw < \infty$ باشد، خانواده جدید توزیع‌ها را به صورت زیر در نظر بگیرید.

$$f_X(x) = \frac{g(h(x))}{\int_{-\infty}^{\infty} g(h(w)) dw}. \quad (۵.۳)$$

قضیه زیر نشان می‌دهد که چگونه توزیع‌های متقارن دومی از توزیع‌های یک‌مدی با دامنه $\mathbb{R}$ ساخته می‌شوند:



شکل ۲.۳: منحنی‌های چگالی‌های (۳.۳) (سمت راست) و (۴.۳) (سمت چپ) برای برخی مقادیر k .

قضیه ۳.۲.۳. فرض کنید تابع چگالی $g(x)$ دارای مد k باشد و $h(\cdot)$ تابعی با شرط $h(\circ) = \circ$ و برای سایر نقاط که محدب و متقارن حول صفر هستند، اکیدا مثبت باشد در این صورت، تابع چگالی (۵.۳):

الف) اگر $k = \circ$ ، یک مدی با مد صفر است.

ب) اگر $k \neq \circ$ ، دو مدی است.

پ) مدهای آن از حل معادله $h(x) = k$ نتیجه می‌شوند.

□

برهان. برای برهان به پیوست ۲.آ مراجعه کنید.

مثال ۳.۲.۳. فرض کنید $X \sim \text{Cauchy}(k, 1)$ و $h(x) = |x|$ آن گاه

$$f_X(x) = \frac{1}{\sqrt{G(k)}} \frac{1}{\pi} \frac{1}{1 + (|x| - k)^2} \quad (۶.۳)$$

یک تابع چگالی متقارن حول صفر با دو مد $\pm k$ است که در آن $G(\cdot)$ تابع توزیع $\text{Cauchy}(\circ, 1)$ است.

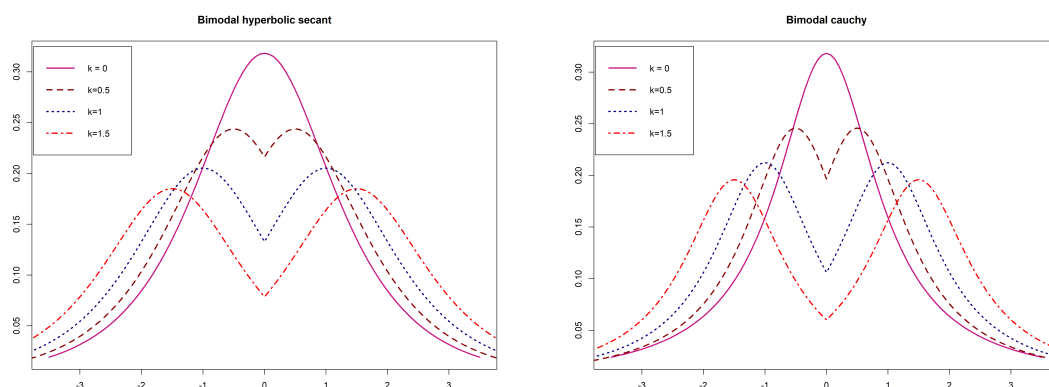
مثال ۴.۲.۳. فرض کنید X از توزیع هایپربولیک سکانت پیروی می‌کند که پارامتر مکان آن k و پارامتر مقیاس آن برابر یک است. اگر $h(x) = |x|$ ، آن گاه

$$f_X(x) = \frac{1}{\sqrt{G(k)}} \frac{2}{\pi} \frac{1}{e^{|x|-k} + e^{-(|x|-k)}} \quad (۷.۳)$$

تابع چگالی متقارن حول صفر با مدهای برابر $\pm k$ است که در آن $G(\cdot)$ تابع توزیع $\text{HS}(\circ, 1)$ می‌باشد.

شکل چگالی‌های (۶.۳) و (۷.۳) در نمودار ۳.۳ برای برخی مقادیر k رسم شده‌اند.

قضیه ۴.۲.۳. برای ساختن توزیع‌های دومی متقارن بدون نیاز به محاسبه ثابت نرمال‌ساز بیان شده است.



شکل ۳.۳: منحنی‌های چگالی‌های (۶.۳) (سمت راست) و (۷.۳) (سمت چپ) برای برخی مقادیر k .

قضیه ۴.۲.۳. اگر $g(x)$ تابع چگالی اکیداً مثبت متقارن حول $0 \neq k$ با تابع توزیع $G_X(\cdot)$ باشد، آن‌گاه $f_X(x) = \frac{1}{\sqrt{2\pi}} \frac{g(|x|)}{G_{(X-k)}(k)}$ تابع چگالی متقارن دومی است.

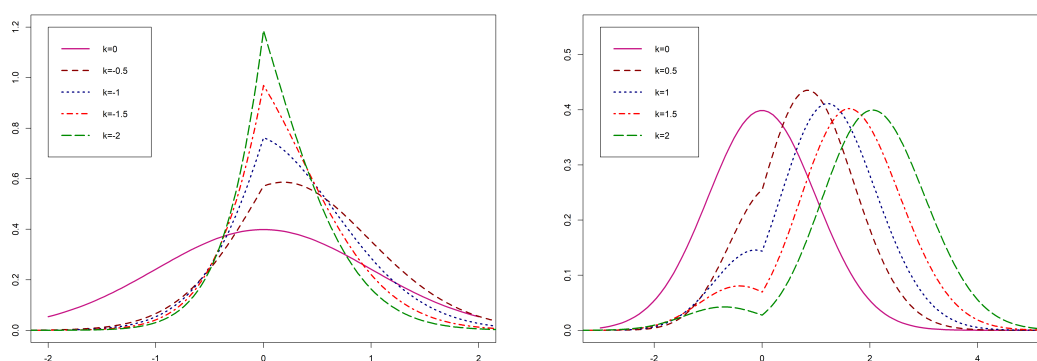
با به کار بردن ایده آزالینی (۱۹۸۵) برای ساختن توزیع‌های چوله از توزیع‌های متقارن، توزیع‌ها متقارن دومی را نیز می‌توان به توزیع‌های چوله تبدیل کرد.

نتیجه ۱.۲.۳. با فرضیات قضیه ۴.۲.۳، $f_X(x) = \frac{g(|x|)}{G_{(X-k)}(k)} F(\lambda x)$ یک تابع چگالی با پارامتر چولگی λ است که در آن $F(\cdot)$ یک تابع توزیع پیوسته با مشتق اول متقارن حول صفر را نشان می‌دهد.

مثال ۵.۲.۳. فرض کنید $X \sim N(k, 1)$ ، $k > 0$ ، آن‌گاه

$$f_X(x) = \frac{\phi(|x| - k)}{\Phi(k)} \Phi(\lambda x) \quad (۸.۳)$$

یک تابع چگالی چوله برای $0 \neq \lambda$ و تابع چگالی متقارن برای $\lambda = 0$ است. شکل تابع چگالی (۸.۳) برای برخی مقادیر k در نمودار ۴.۳ رسم شده است.



شکل ۴.۳: منحنی‌های چگالی (۸.۳) برای برخی مقادیر k .

۳.۳ برخی حالات خاص توزیع‌های دومدی-یکمدی

توزیع‌های شناخته‌شده و کاربردی برای تحلیل داده‌های متقارن، توزیع‌های نرمال، St و لاپلاس هستند که به ترتیب دارای دم سبک، دم سنگین و دم نیمه‌سنگین هستند. تعمیم‌های متعددی از این توزیع‌ها مانند SN، SSt و لاپلاس چوله^{۱۲} (SL) وجود دارند که برای داده‌های دارای چولگی مناسب هستند. تعمیم‌هایی که در این بخش برای این توزیع‌ها انجام خواهیم داد، انعطاف‌پذیری بالایی نسبت به تعمیم‌های ذکرشده دارند. یکی از مزیت‌هایی که آن تعمیم‌ها دارند این است که برای داده‌های یک مدی چوله، متقارن و دومدی چوله و متقارن قابل برازش هستند. مزیت‌های این تعمیم‌ها را می‌توان به صورت زیر فهرست کرد:

- به رده بزرگی از داده‌ها با ویژگی‌هایی مانند متقارن، چوله و دومدی قابل برازش هستند.
- توزیع‌های نرمال، St و لاپلاس را به عنوان حالت خاص در بردارند.
- تابع چگالی پیچیده‌ای ندارند و کار با آنها راحت است.
- برخی از این تعمیم‌ها دارای نمایش تصادفی^{۱۳} هستند.

۱.۳.۳ توزیع نرمال دومدی-یک مدی

در این بخش تعمیم جدیدی از توزیع نرمال را با استفاده از قضیه ۱.۲.۳ ارائه خواهیم داد (مثال ۱.۲.۳ را ببینید). این تابع چگالی با ضرب $\exp(k|\frac{x-\mu}{\sigma}|)$ در تابع چگالی توزیع نرمال به دست آمده است که دارای تابع چگالی به شکل زیر است:

$$f(x) = c_{\sigma,k,a} \exp\left(k\left|\frac{x-\mu}{\sigma}\right|\right) \frac{1}{\sigma} \phi\left(\frac{x-\mu-a}{\sigma}\right) \quad x \in \mathbb{R} \quad (9.3)$$

که در آن

$$c_{\sigma,k,a}^{-1} = \exp\left(\frac{ka}{\sigma} + \frac{k^2}{2}\right) \Phi\left(k + \frac{a}{\sigma}\right) + \exp\left(-\frac{ka}{\sigma} + \frac{k^2}{2}\right) \Phi\left(k - \frac{a}{\sigma}\right).$$

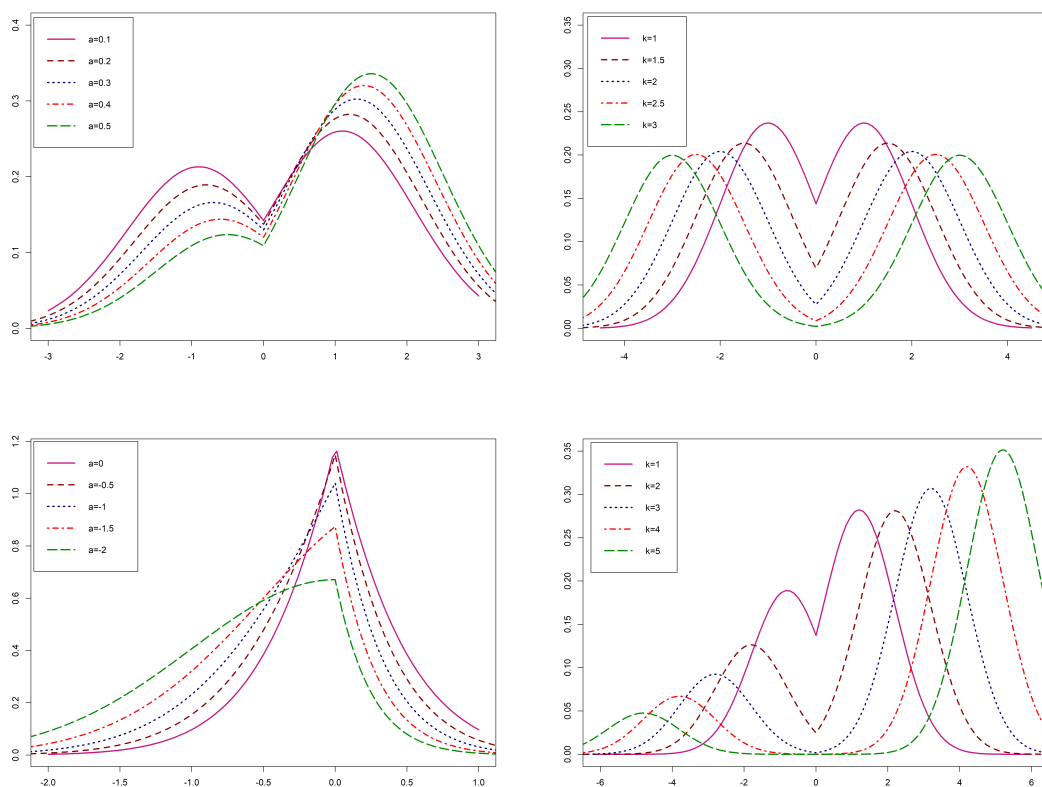
اگر متغیر تصادفی X دارای تابع چگالی (۹.۳) باشد، می‌گوئیم X دارای توزیع نرمال دومدی-یک مدی^{۱۴} (BUN) است و آن را با نماد $X \sim \text{BUN}(\mu, \sigma, k, a)$ نمایش می‌دهیم. در این توزیع $\mu \in \mathbb{R}$ پارامتر مکان، $\sigma > 0$ پارامتر مقیاس و $k \in \mathbb{R}$ و $a \in \mathbb{R}$ پارامترهای شکل هستند. این چگالی برای $a = 0$ متقارن است که در این حالت از نماد $X \sim \text{BUN}(\mu, \sigma, k)$ استفاده می‌کنیم. برای $a \neq 0$ تابع چگالی نامتقارنی خواهیم داشت که برای $k = 0$ توزیع نرمال به دست می‌آید. محدوده‌ی دومدی بودن (۹.۳) برابر $\sigma k > |a|$ می‌باشد. پارامتر k علاوه بر این که در دومدی

¹²Skew Laplace

¹³Stochastic Representation

¹⁴Bimodal-Unimodal Normal

بودن یا نبودن نقش دارد، فاصله دو مد را نیز کنترل می‌کند. در شکل ۵.۳ منحنی‌های تابع چگالی (۹.۳) به ازای مقادیر مختلفی از پارامترهای k و a نمایش داده شده‌اند. همان‌طور که مشاهده می‌کنید، با تغییر پارامترهای توزیع انواع شکل‌های دومی و یک‌مدی برای تابع چگالی توزیع را می‌توان ساخت. در قسمت بالا (سمت راست) از نمودار، منحنی چگالی (۹.۳) را مشاهده می‌کنیم که شکل چگالی متقارن است ($a = 0$) و با افزایش مقدار پارامتر k فاصله دو مد بیشتر شده است. در قسمت بالا (سمت چپ) از نمودار با افزایش مقدار پارامتر a چولگی زیاد شده است. قسمت پایین (سمت راست) نمودار نیز تأکید دارد با افزایش پارامتر k فاصله دو مد زیاد می‌شود. قسمت بالا (سمت چپ) نیز حالتی است که چگالی یک مد دارد. دقت کنید که این انعطاف تنها به وسیله دو پارامتر به توزیع ارث داده می‌شود.



شکل ۵.۳: منحنی‌های چگالی (۹.۳) برای برخی مقادیر پارامترهای شکل

گزاره ۱.۳.۳. اگر X از توزیع $BNU(\mu, \sigma, k, a)$ پیروی کند، داریم

$$\text{Mode} = \begin{cases} \mu + a \pm \sigma k & \sigma k > |a| \\ \mu + a - \sigma k & a < \sigma k < -a \\ \mu + a + \sigma k & -a < \sigma k < a \\ \mu & \sigma k \leq -|a|. \end{cases} \quad (10.3)$$

تابع توزیع تجمعی، تابع مولد گشتاورها و گشتاورهای توزیع BUN

برای راحتی در به‌دست آوردن ویژگی‌های توزیع BUN مانند گشتاورها، حالت استاندارد BUN را مورد بررسی قرار می‌دهیم. اگر $Z = \frac{X-\mu}{\sigma}$ ، آن‌گاه $Z \sim \text{BUN}(\circ, 1, k, \frac{a}{\sigma})$. در این صورت تابع توزیع تجمعی متغیر تصادفی Z برابر است با

$$F(z) = \begin{cases} \frac{e^{-\frac{ka}{\sigma}} \Phi(z+k-\frac{a}{\sigma})}{e^{\frac{ka}{\sigma}} \Phi(k+\frac{a}{\sigma}) + e^{-\frac{ka}{\sigma}} \Phi(k-\frac{a}{\sigma})} & z \leq \circ \\ \frac{e^{-\frac{ka}{\sigma}} \Phi(k-\frac{a}{\sigma}) + e^{\frac{ka}{\sigma}} [\Phi(z-k-\frac{a}{\sigma}) - \Phi(-k-\frac{a}{\sigma})]}{e^{\frac{ka}{\sigma}} \Phi(k+\frac{a}{\sigma}) + e^{-\frac{ka}{\sigma}} \Phi(k-\frac{a}{\sigma})} & z > \circ. \end{cases}$$

نکته کلیدی در ساختار توزیع BUN این است که می‌توانیم BUN را به صورت آمیخته‌ای از دو توزیع نرمال بریده‌شده بنویسیم، به طوری که نرمال‌های بریده‌شده بدون هم‌پوشانی هستند. گزاره ۲.۳.۳. تابع چگالی $X \sim \text{BUN}(\mu, \sigma, k, a)$ آمیخته‌ای از دو توزیع نرمال بریده‌شده به صورت زیر است:

$$\alpha \frac{\phi(\frac{x-\mu}{\sigma} + k - \frac{a}{\sigma})}{\sigma \Phi(k - \frac{a}{\sigma})} + (1 - \alpha) \frac{\phi(\frac{x-\mu}{\sigma} - k - \frac{a}{\sigma})}{\sigma \Phi(k + \frac{a}{\sigma})}$$

که در آن

$$\alpha = \frac{e^{-\frac{ka}{\sigma}} \Phi(k - \frac{a}{\sigma})}{e^{\frac{ka}{\sigma}} \Phi(k + \frac{a}{\sigma}) + e^{-\frac{ka}{\sigma}} \Phi(k - \frac{a}{\sigma})}$$

است. به راحتی می‌توان نشان داد که $\frac{\phi(z-k-\frac{a}{\sigma})}{\Phi(k+\frac{a}{\sigma})}$ و $\frac{\phi(z+k-\frac{a}{\sigma})}{\Phi(k-\frac{a}{\sigma})}$ توابع چگالی نرمال بریده‌شده روی بازه‌های به ترتیب $(-\infty, \mu)$ و (μ, ∞) هستند.

تابع مولد گشتاور Z به صورت زیر به دست می‌آید:

$$M_Z(t) = \frac{e^{(k+t)\frac{a}{\sigma} + \frac{(k+t)^{\gamma}}{\gamma}} \Phi(k + \frac{a}{\sigma} + t) + e^{-(k-t)\frac{a}{\sigma} + \frac{(k-t)^{\gamma}}{\gamma}} \Phi(k - \frac{a}{\sigma} - t)}{e^{\frac{ka}{\sigma} + \frac{k^{\gamma}}{\gamma}} \Phi(k + \frac{a}{\sigma}) + e^{-\frac{ka}{\sigma} + \frac{k^{\gamma}}{\gamma}} \Phi(k - \frac{a}{\sigma})}.$$

بنابراین به راحتی می‌توانیم گشتاورهای BUN را محاسبه کنیم. برخی گشتاورهای متغیر تصادفی Z عبارتند از:

$$\begin{aligned} E(Z) &= \frac{p_k p_t - n_k n_t}{\delta} \\ E(Z^2) &= \frac{(1 + p_k^{\gamma}) p_t + (1 + n_k^{\gamma}) n_t + 2 k n_d}{\delta} \\ E(Z^3) &= \frac{(3 p_k + p_k^{\gamma}) p_t - (3 n_k + n_k^{\gamma}) n_t + 6 \frac{ka}{\sigma} n_d}{\delta} \\ E(Z^4) &= \frac{(3 + 6 p_k^{\gamma} + p_k^{\epsilon}) p_t + (3 + 6 n_k^{\gamma} + n_k^{\epsilon}) n_t + [2 k^3 + 10 k + 6 (\frac{a}{\sigma})^2 k] n_d}{\delta} \end{aligned}$$

که در آنها

$$\begin{aligned} p_t &= e^{\frac{ka}{\sigma}} \Phi\left(k + \frac{a}{\sigma}\right) \\ n_t &= e^{\frac{-ka}{\sigma}} \Phi\left(k - \frac{a}{\sigma}\right) \\ \delta &= p_t + n_t \\ p_k &= k + \frac{a}{\sigma} \\ n_k &= k - \frac{a}{\sigma} \\ n_d &= e^{\frac{-ka}{\sigma}} \phi\left(k - \frac{a}{\sigma}\right) \end{aligned}$$

برآوردهای ماکسیمم درست‌نمایی پارامترهای توزیع BUN

فرض کنید $\theta^{BUN} = (\mu, \sigma, k, a)'$ بردار پارامترهای مدل باشد. تابع لگاریتم درست‌نمایی مدل برابر است با

$$\begin{aligned} \ell = \ell(\theta^{BUN}) &= -\frac{nk^2}{2} - n \log(\sigma) - n \log(\delta) - \frac{n}{2} \log(2\pi) \\ &\quad + \frac{k}{\sigma} \sum_{i=1}^n |x_i - \mu| - \frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu - a}{\sigma} \right)^2. \end{aligned}$$

برای ماکسیمم کردن ℓ نسبت به بردار پارامترهای θ^{BUN} ، باید بردار امتیاز

$$S_n^{BUN} = \left(\frac{\partial \ell}{\partial \mu}, \frac{\partial \ell}{\partial \sigma}, \frac{\partial \ell}{\partial k}, \frac{\partial \ell}{\partial a} \right)' = 0$$

را حل کنیم. مولفه‌های بردار امتیاز به صورت زیر محاسبه می‌شوند:

$$\begin{aligned} \frac{\partial \ell}{\partial \mu} &= -\frac{k}{\sigma} \sum_{i=1}^n \text{sign}(x_i - \mu) + \frac{1}{\sigma} \sum_{i=1}^n \left(\frac{x_i - \mu - a}{\sigma} \right) \\ \frac{\partial \ell}{\partial \sigma} &= -\frac{n}{\sigma} + n \frac{ka}{\sigma^2} - \frac{k}{\sigma^2} \sum_{i=1}^n |x_i - \mu| + \frac{1}{\sigma} \sum_{i=1}^n \left(\frac{x_i - \mu - a}{\sigma} \right)^2 \\ \frac{\partial \ell}{\partial k} &= -nk - n \frac{as}{\sigma} - n\rho + \sum_{i=1}^n \left| \frac{x_i - \mu}{\sigma} \right| \\ \frac{\partial \ell}{\partial a} &= -n \frac{ks}{\sigma} + \frac{1}{\sigma} \sum_{i=1}^n \left(\frac{x_i - \mu - a}{\sigma} \right) \end{aligned}$$

که در آنها

$$\begin{aligned} \delta &= e^{\frac{ka}{\sigma}} \Phi\left(k + \frac{a}{\sigma}\right) + e^{-\frac{ka}{\sigma}} \Phi\left(k - \frac{a}{\sigma}\right) \\ s &= \frac{e^{\frac{ka}{\sigma}} \Phi\left(k + \frac{a}{\sigma}\right) - e^{-\frac{ka}{\sigma}} \Phi\left(k - \frac{a}{\sigma}\right)}{\delta} \\ \rho &= \frac{2e^{\frac{ka}{\sigma}} \phi\left(k + \frac{a}{\sigma}\right)}{\delta}. \end{aligned}$$

حل این دستگاه معادلات دارای شکل بسته نیستند. بنابراین باید با روش‌های عددی بهینه‌سازی آن را حل کنیم.

۲.۳.۳ توزیع t استیودنت دومدی-یکمدی

از معروف‌ترین توزیع‌ها برای داده‌های شامل داده پرت، توزیع دم‌سنگین St است. تعمیم‌های متعددی از این توزیع در مقالات و نوشته‌های علمی وجود دارند. برای مثال به جونز و فادی (۲۰۰۳)، آس و هاف^{۱۵} (۲۰۰۶)، ناداراجا و کتز^{۱۶} (۲۰۰۶)، بالاکریشنان و لای^{۱۷} (۲۰۰۹) هانگ^{۱۸} و همکاران (۲۰۱۹) مراجعه کنید. در این بخش، تعمیم جدیدی از توزیع St را با استفاده از قضیه ۳.۲.۳ معرفی و بررسی خواهیم کرد. در قضیه ۳.۲.۳، $g(x)$ را تابع چگالی St با پارامتر مکان k و $h(x) = |x|$ در نظر می‌گیریم. تابع چگالی توزیع جدید برابر

$$f(x) = \frac{d_\nu \left(\left| \frac{x-\mu}{\sigma} \right| - k \right)}{\sqrt{2\sigma} D_\nu(k)} = \frac{\{d_\nu \left(\frac{x-\mu}{\sigma} - k \right)\}^{I(x \geq \mu)} \{d_\nu \left(\frac{x+\mu}{\sigma} + k \right)\}^{I(x < \mu)}}{\sqrt{2\sigma} D_\nu(k)} \quad (11.3)$$

است، که برای $k > 0$ دومدی و برای $k = 0$ همان توزیع St است. این توزیع دارای تابع چگالی متقارن است که برای داده‌های نامتقارن مناسب نیست. می‌توانیم با افزودن یک پارامتر شکل، تابع چگالی (۱۱.۳) را به یک تابع چگالی منعطف چوله تعمیم دهیم. می‌توان نوشت

$$f(x) = \frac{\left\{ s_-^{-\frac{\nu+1}{\sqrt{s_-}}} d_\nu \left(\frac{x-\mu-(a+k)}{\sqrt{s_-}} \right) \right\}^{I(x \geq \mu)} \left\{ s_+^{-\frac{\nu+1}{\sqrt{s_+}}} d_\nu \left(\frac{x-\mu-(a-k)}{\sqrt{s_+}} \right) \right\}^{I(x < \mu)}}{s_-^{-\frac{\nu}{\sqrt{s_-}}} D_\nu \left(\frac{a+k}{\sqrt{s_-}} \right) + s_+^{-\frac{\nu}{\sqrt{s_+}}} D_\nu \left(\frac{k-a}{\sqrt{s_+}} \right)} \quad (12.3)$$

که $d_\nu(\cdot)$ و $D_\nu(\cdot)$ در (۱۱.۳) و (۱۲.۳) به ترتیب تابع چگالی و تابع توزیع St با ν درجه آزادی است و همچنین $s_+ = \frac{\nu\sigma^2 + 2ak}{\nu}$ و $s_- = \frac{\nu\sigma^2 - 2ak}{\nu}$.

اگر متغیر تصادفی X دارای تابع چگالی (۱۲.۳) باشد، می‌گوییم X از توزیع t استیودنت دومدی-یکمدی^{۱۹} (BUST) پیروی می‌کند و با نماد $X \sim \text{BUST}(\mu, \sigma, k, a, \nu)$ نمایش می‌دهیم. اگر $a = 0$ باشد، تابع چگالی (۱۲.۳) معادل (۱۱.۳) است، که در این حالت با نماد $X \sim \text{BUST}(\mu, \sigma, k, \nu)$ نمایش می‌دهیم. ناحیه دومدی بودن تابع چگالی (۱۲.۳) برابر $|a| < k$ است. شکل تابع چگالی (۱۲.۳) برای برخی پارامترها در نمودار ۶.۳ رسم شده است.

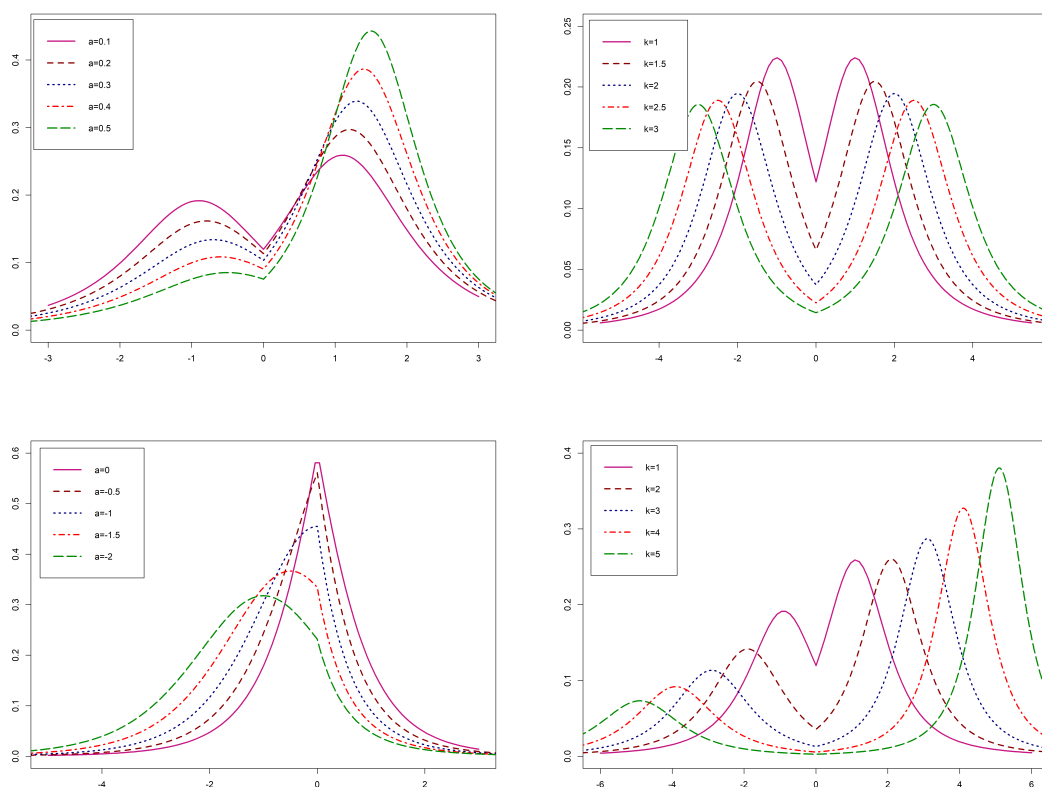
¹⁵Aas and Haff

¹⁶Kotz

¹⁷Balakrishnan and Lai

¹⁸Huang

¹⁹Bimodal-Unimodal Student t



شکل ۶.۳: منحنی‌های (۱۲.۳) برای برخی مقادیر پارامترها. $(\mu = 0, \sigma = 1, k = 0, \nu = 2)$ (بالا سمت راست)، $(\mu = 0, \sigma = 1, k = 1, \nu = 2)$ (بالا سمت چپ)، $(\mu = 0, \sigma = 1, k = 0, \nu = 5)$ (پایین سمت راست) و $(\mu = 0, \sigma = 1, k = -1, \nu = 5)$ (پایین سمت چپ) و $a = 0/1$

گزاره ۳.۳.۳. اگر X از توزیع $\text{BUST}(\mu, \sigma, k, a, \nu)$ پیروی کند، داریم

$$\text{Mode} = \begin{cases} \mu + a \pm k & k > |a| \\ \mu + a - k & a < k < -a \\ \mu + a + k & -a < k < a \\ \mu & k \leq -|a|. \end{cases}$$

گزاره ۴.۳.۳. اگر $X \sim \text{BUST}(\mu, \sigma, k, a, \nu)$ و $Y \sim \text{BUN}(\mu, \sigma, \frac{k}{\sigma}, a)$ آن‌گاه برای $\nu \rightarrow \infty$ داریم

$$X \xrightarrow{d} Y$$

که $\xrightarrow{d}$ به معنی همگرایی در توزیع است.

تابع توزیع تجمعی و گشتاورهای توزیع BUST

همانند BUN، BUST نیز دارای نمایش تصادفی به صورت آمیخته‌ای از دو توزیع St بریده‌شده بدون هم‌پوشانی است.

گزاره ۵.۳.۳. چگالی (۱۲.۳) آمیخته‌ای از دو توزیع St بریده‌شده به صورت زیر است:

$$f(x) = R(s_-) f(x_1) + R(s_+) f(x_2)$$

که در آن $X_2 \sim \text{Tt}_{(-\infty, \mu)}(\mu + (a - k), \sqrt{s_+}; \nu)$ ، $X_1 \sim \text{Tt}_{(\mu, \infty)}(\mu + (a + k), \sqrt{s_-}; \nu)$ و

$$R(s_-) = \frac{s_-^{-\frac{\nu}{2}} D_\nu\left(\frac{a+k}{\sqrt{s_-}}\right)}{s_-^{-\frac{\nu}{2}} D_\nu\left(\frac{a+k}{\sqrt{s_-}}\right) + s_+^{-\frac{\nu}{2}} D_\nu\left(\frac{k-a}{\sqrt{s_+}}\right)}$$

$$R(s_+) = \frac{s_+^{-\frac{\nu}{2}} D_\nu\left(\frac{k-a}{\sqrt{s_+}}\right)}{s_-^{-\frac{\nu}{2}} D_\nu\left(\frac{a+k}{\sqrt{s_-}}\right) + s_+^{-\frac{\nu}{2}} D_\nu\left(\frac{k-a}{\sqrt{s_+}}\right)}.$$

بر اساس کیم^{۲۰} (۲۰۰۸) توابع چگالی متغیرهای تصادفی X_1 و X_2 به صورت زیر هستند:

$$f(x_1) = \frac{s_-^{-\frac{1}{2}} d_\nu\left(\frac{x_1 - \mu - (a+k)}{\sqrt{s_-}}\right)}{D_\nu\left(\frac{a+k}{\sqrt{s_-}}\right)}$$

$$f(x_2) = \frac{s_+^{-\frac{1}{2}} d_\nu\left(\frac{x_2 - \mu - (a-k)}{\sqrt{s_+}}\right)}{D_\nu\left(\frac{k-a}{\sqrt{s_+}}\right)}.$$

از گزاره ۵.۳.۳ برای تولید نمونه از توزیع BUST و محاسبه گشتاورهای این توزیع استفاده می‌کنیم. r -امین گشتاور $X \sim \text{BUST}(\mu, \sigma, k, a, \nu)$ برای $r = 1, 2, 3, 4$ برابر است با

$$E(X^r) = \sum_{i=0}^r \binom{r}{i} R(s_-)^i R(s_+)^{r-i} E(X_1^i) E(X_2^{r-i})$$

که بر اساس کیم (۲۰۰۸) داریم

$$E(X_1^m) = \sum_{j=0}^m \binom{m}{j} (\mu + a + k)^{m-j} (s_-)^{\frac{j}{2}} \eta_m$$

9

$$E(X_2^l) = \sum_{j=0}^l \binom{l}{j} (\mu + a - k)^{l-j} (s_+)^{\frac{j}{2}} \lambda_l$$

به طوری که برای $m, l = 1, 2, 3, 4$

$$\begin{aligned}\eta_1 &= G_\nu(1) \left(\nu + \frac{(a+k)^{\frac{\nu}{\gamma}}}{s_-} \right)^{-\frac{\nu-1}{\gamma}} \quad \nu > 1 \\ \eta_2 &= \frac{\nu}{\nu-\frac{\nu}{\gamma}} - \frac{a+k}{\sqrt{s_-}} G_\nu(1) \left(\nu + \frac{(a+k)^{\frac{\nu}{\gamma}}}{s_-} \right)^{-\frac{\nu-1}{\gamma}} \quad \nu > 2 \\ \eta_3 &= G_\nu(3) \left(\nu + \frac{(a+k)^{\frac{\nu}{\gamma}}}{s_-} \right)^{-\frac{\nu-\frac{\nu}{\gamma}}{\gamma}} + \frac{(a+k)^{\frac{\nu}{\gamma}}}{s_-} G_\nu(1) \left(\nu + \frac{(a+k)^{\frac{\nu}{\gamma}}}{s_-} \right)^{-\frac{\nu-1}{\gamma}} \quad \nu > 3 \\ \eta_4 &= \frac{\nu^{\frac{\nu}{\gamma}}}{(\nu-\frac{\nu}{\gamma})(\nu-\frac{\nu}{\gamma})} - \frac{G_\nu(3)}{\frac{\nu}{\gamma}} \frac{a+k}{\sqrt{s_-}} \left(\nu + \frac{(a+k)^{\frac{\nu}{\gamma}}}{s_-} \right)^{-\frac{\nu-\frac{\nu}{\gamma}}{\gamma}} \\ &\quad - \frac{(a+k)^{\frac{\nu}{\gamma}}}{(s_-)^{\frac{\nu}{\gamma}}} G_\nu(1) \left(\nu + \frac{(a+k)^{\frac{\nu}{\gamma}}}{s_-} \right)^{-\frac{\nu-1}{\gamma}} \quad \nu > 4\end{aligned}$$

9

$$\begin{aligned}\lambda_1 &= -G'_\nu(1) \left(\nu + \frac{(a-k)^{\frac{\nu}{\gamma}}}{s_+} \right)^{-\frac{\nu-1}{\gamma}} \quad \nu > 1 \\ \lambda_2 &= \frac{\nu}{\nu-\frac{\nu}{\gamma}} + \frac{a-k}{\sqrt{s_+}} G'_\nu(1) \left(\nu + \frac{(a-k)^{\frac{\nu}{\gamma}}}{s_+} \right)^{-\frac{\nu-1}{\gamma}} \quad \nu > 2 \\ \lambda_3 &= -G'_\nu(3) \left(\nu + \frac{(a-k)^{\frac{\nu}{\gamma}}}{s_+} \right)^{-\frac{\nu-\frac{\nu}{\gamma}}{\gamma}} - \frac{(a-k)^{\frac{\nu}{\gamma}}}{s_+} G'_\nu(1) \left(\nu + \frac{(a-k)^{\frac{\nu}{\gamma}}}{s_+} \right)^{-\frac{\nu-1}{\gamma}} \quad \nu > 3 \\ \lambda_4 &= \frac{\nu^{\frac{\nu}{\gamma}}}{(\nu-\frac{\nu}{\gamma})(\nu-\frac{\nu}{\gamma})} + \frac{G'_\nu(3)}{\frac{\nu}{\gamma}} \frac{a-k}{\sqrt{s_+}} \left(\nu + \frac{(a-k)^{\frac{\nu}{\gamma}}}{s_+} \right)^{-\frac{\nu-\frac{\nu}{\gamma}}{\gamma}} \\ &\quad + \frac{(a-k)^{\frac{\nu}{\gamma}}}{(s_+)^{\frac{\nu}{\gamma}}} G'_\nu(1) \left(\nu + \frac{(a-k)^{\frac{\nu}{\gamma}}}{s_+} \right)^{-\frac{\nu-1}{\gamma}} \quad \nu > 4\end{aligned}$$

و برای $s = 1, 2$

$$G_\nu(s) = \frac{\Gamma\left(\frac{\nu-s}{\gamma}\right) \nu^{\frac{\nu}{\gamma}}}{\frac{\nu}{\gamma} D_\nu\left(\frac{a+k}{\sqrt{s_-}}\right) \Gamma\left(\frac{\nu}{\gamma}\right) \Gamma\left(\frac{1}{\gamma}\right)}$$

9

$$G'_\nu(s) = \frac{\Gamma\left(\frac{\nu-s}{\gamma}\right) \nu^{\frac{\nu}{\gamma}}}{\frac{\nu}{\gamma} D_\nu\left(\frac{k-a}{\sqrt{s_+}}\right) \Gamma\left(\frac{\nu}{\gamma}\right) \Gamma\left(\frac{1}{\gamma}\right)}.$$

تابع توزیع تجمعی (۱۲.۳) برابر است با

$$F(x) = \begin{cases} \frac{s_+^{-\frac{\nu}{\gamma}} D_\nu \left(\frac{x-\mu-(a-k)}{\sqrt{s_+}} \right)}{s_-^{-\frac{\nu}{\gamma}} D_\nu \left(\frac{a+k}{\sqrt{s_-}} \right) + s_+^{-\frac{\nu}{\gamma}} D_\nu \left(\frac{k-a}{\sqrt{s_+}} \right)} & x \leq \mu \\ \frac{s_-^{-\frac{\nu}{\gamma}} \left(D_\nu \left(\frac{x-\mu-(a+k)}{\sqrt{s_-}} \right) - D_\nu \left(\frac{-k-a}{\sqrt{s_-}} \right) \right) + s_+^{-\frac{\nu}{\gamma}} D_\nu \left(\frac{k-a}{\sqrt{s_+}} \right)}{s_-^{-\frac{\nu}{\gamma}} D_\nu \left(\frac{a+k}{\sqrt{s_-}} \right) + s_+^{-\frac{\nu}{\gamma}} D_\nu \left(\frac{k-a}{\sqrt{s_+}} \right)} & x > \mu \end{cases} \quad (13.3)$$

برآوردهای ماکسیمم درستنمایی پارامترهای توزیع BUSt

فرض کنید $\theta^{BUSt} = (\mu, \sigma, k, a, \nu)^T$ بردار پارامترهای توزیع BUSt باشد. تابع لگ درستنمایی توزیع به صورت زیر به دست می آید:

$$\begin{aligned} \ell = \ell(\theta^{BUSt}) &= -\frac{\nu+1}{\gamma} \log(s_-) \sum_{i=1}^n I(x_i \geq \mu) \\ &- n \log(\delta) - \frac{\nu+1}{\gamma} \log(s_+) \sum_{i=1}^n I(x_i < \mu) \\ &+ n \log \left(\Gamma \left(\frac{\nu+1}{\gamma} \right) \right) - n \log \left(\Gamma \left(\frac{\nu}{\gamma} \right) \right) - n \log(\sqrt{\nu\pi}) \\ &- \frac{\nu+1}{\gamma} \sum_{i=1}^n \left\{ \log \left(1 + \frac{1}{\nu} \left(\frac{u_i^-}{\sqrt{s_-}} \right)^\gamma \right) I(x_i \geq \mu) \right\} \\ &- \frac{\nu+1}{\gamma} \sum_{i=1}^n \left\{ \log \left(1 + \frac{1}{\nu} \left(\frac{u_i^+}{\sqrt{s_+}} \right)^\gamma \right) I(x_i < \mu) \right\}. \end{aligned}$$

مؤلفه های بردار امتیاز $S_n^{BUSt} = \left(\frac{\partial \ell}{\partial \mu}, \frac{\partial \ell}{\partial \sigma}, \frac{\partial \ell}{\partial k}, \frac{\partial \ell}{\partial a}, \frac{\partial \ell}{\partial \nu} \right)^T$ به صورت زیر محاسبه می شوند:

$$\begin{aligned} \frac{\partial \ell}{\partial \mu} &= \sum_{i=1}^n \frac{\frac{\nu+1}{\nu} \frac{u_i^-}{s_-} I(x_i \geq \mu)}{1 + \frac{1}{\nu} \left(\frac{u_i^-}{\sqrt{s_-}} \right)^\gamma} + \sum_{i=1}^n \frac{\frac{\nu+1}{\nu} \frac{u_i^+}{s_+} I(x_i < \mu)}{1 + \frac{1}{\nu} \left(\frac{u_i^+}{\sqrt{s_+}} \right)^\gamma} \\ \frac{\partial \ell}{\partial \sigma} &= n\nu\sigma \frac{s_-^{-\frac{\nu+\gamma}{\gamma}} D_\nu^- + s_+^{-\frac{\nu+\gamma}{\gamma}} D_\nu^+}{\delta} \\ &- n\sigma \frac{(a+k) s_-^{-\frac{\nu+\gamma}{\gamma}} d_\nu^- + (k-a) s_+^{-\frac{\nu+\gamma}{\gamma}} d_\nu^+}{\delta} \\ &+ \frac{\nu+1}{\nu} \sum_{i=1}^n \frac{\frac{\sigma}{s_-^\gamma} \left(\frac{u_i^-}{\sqrt{s_-}} \right)^\gamma I(x_i \geq \mu)}{1 + \frac{1}{\nu} \left(\frac{u_i^-}{\sqrt{s_-}} \right)^\gamma} - (\nu+1) \frac{\sigma}{s_-} \sum_{i=1}^n I(x_i \geq \mu) \end{aligned}$$

$$\begin{aligned}
 & + \frac{\nu+1}{\nu} \sum_{i=1}^n \frac{\frac{\sigma}{s_+} \left(\frac{u_i^+}{\sqrt{s_+}} \right)^{\frac{\nu}{\nu+1}} I(x_i < \mu)}{1 + \frac{1}{\nu} \left(\frac{u_i^+}{\sqrt{s_+}} \right)^{\frac{\nu}{\nu+1}}} - (\nu+1) \frac{\sigma}{s_+} \sum_{i=1}^n I(x_i < \mu) \\
 \frac{\partial \ell}{\partial k} = & -n \frac{as_-^{-\frac{\nu+1}{\nu}} D_\nu^- + s_-^{-\frac{\nu+1}{\nu}} \left((a+k) \frac{as_-^{-1}}{\nu} + 1 \right) d_\nu^-}{\delta} \\
 & - n \frac{as_+^{-\frac{\nu+1}{\nu}} D_\nu^+ + s_+^{-\frac{\nu+1}{\nu}} \left((k-a) \frac{as_+^{-1}}{\nu} - 1 \right) d_\nu^+}{\delta} \\
 & - \frac{\nu+1}{\nu} \sum_{i=1}^n \frac{\left(\frac{u_i^-}{\sqrt{s_-}} \right) \left(\frac{-1 + \frac{as_-^{-1}}{\nu} (u_i^-)}{\sqrt{s_-}} \right) I(x_i \geq \mu)}{1 + \frac{1}{\nu} \left(\frac{u_i^-}{\sqrt{s_-}} \right)^{\frac{\nu}{\nu+1}}} \\
 & - \frac{\nu+1}{\nu} \sum_{i=1}^n \frac{\left(\frac{u_i^+}{\sqrt{s_+}} \right) \left(\frac{1 - \frac{as_+^{-1}}{\nu} (u_i^+)}{\sqrt{s_+}} \right) I(x_i < \mu)}{1 + \frac{1}{\nu} \left(\frac{u_i^+}{\sqrt{s_+}} \right)^{\frac{\nu}{\nu+1}}} \\
 & + a \frac{\nu+1}{\nu s_-} \sum_{i=1}^n I(x_i \geq \mu) - a \frac{\nu+1}{\nu s_+} \sum_{i=1}^n I(x_i < \mu) \\
 \frac{\partial \ell}{\partial a} = & -n \frac{ks_-^{-\frac{\nu+1}{\nu}} D_\nu^- + s_-^{-\frac{\nu+1}{\nu}} \left((a+k) \frac{ks_-^{-1}}{\nu} + 1 \right) d_\nu^-}{\delta} \\
 & - n \frac{ks_+^{-\frac{\nu+1}{\nu}} D_\nu^+ + s_+^{-\frac{\nu+1}{\nu}} \left((k-a) \frac{ks_+^{-1}}{\nu} - 1 \right) d_\nu^+}{\delta} \\
 & - \frac{\nu+1}{\nu} \sum_{i=1}^n \frac{\left(\frac{u_i^-}{\sqrt{s_-}} \right) \left(\frac{-1 + \frac{ks_-^{-1}}{\nu} (u_i^-)}{\sqrt{s_-}} \right) I(x_i \geq \mu)}{1 + \frac{1}{\nu} \left(\frac{u_i^-}{\sqrt{s_-}} \right)^{\frac{\nu}{\nu+1}}} \\
 & + \frac{\nu+1}{\nu} \sum_{i=1}^n \frac{\left(\frac{u_i^+}{\sqrt{s_+}} \right) \left(\frac{1 + \frac{ks_+^{-1}}{\nu} (u_i^+)}{\sqrt{s_+}} \right) I(x_i < \mu)}{1 + \frac{1}{\nu} \left(\frac{u_i^+}{\sqrt{s_+}} \right)^{\frac{\nu}{\nu+1}}} \\
 & + k \frac{\nu+1}{\nu s_-} \sum_{i=1}^n I(x_i \geq \mu) - k \frac{\nu+1}{\nu s_+} \sum_{i=1}^n I(x_i < \mu)
 \end{aligned}$$

$$\begin{aligned}
 \frac{\partial \ell}{\partial \nu} = & -n \frac{\left(-\frac{\log(s_-)}{\nu} - \frac{ak}{\nu s_-} \right) s_-^{-\frac{\nu}{\nu}} D_{\nu}^{-} + s_-^{-\frac{\nu+\Delta}{\nu}} \frac{(a+k)ak}{\nu^{\frac{\nu}{\nu}}} d_{\nu}^{-}}{\delta} \\
 & - n \frac{\left(-\frac{\log(s_+)}{\nu} + \frac{ak}{\nu s_+} \right) s_+^{-\frac{\nu}{\nu}} D_{\nu}^{+} - s_+^{-\frac{\nu+\Delta}{\nu}} \frac{(k-a)ak}{\nu^{\frac{\nu}{\nu}}} d_{\nu}^{+}}{s_+^{-\frac{\nu}{\nu}} D_{\nu} \left(\frac{a+k}{\sqrt{s_-}} \right) + s_+^{-\frac{\nu}{\nu}} D_{\nu} \left(\frac{k-a}{\sqrt{s_+}} \right)} \\
 & - \frac{\log(s_-)}{\nu} \sum_{i=1}^n I(x_i \geq \mu) - \frac{\nu+1}{\nu^{\frac{\nu}{\nu}}} \frac{ak}{s_-} \sum_{i=1}^n I(x_i \geq \mu) \\
 & - \frac{\log(s_+)}{\nu} \sum_{i=1}^n I(x_i < \mu) + \frac{\nu+1}{\nu^{\frac{\nu}{\nu}}} \frac{ak}{s_+} \sum_{i=1}^n I(x_i < \mu) \\
 & + n \left(\psi \left(\frac{\nu+1}{\nu} \right) - \psi \left(\frac{\nu}{\nu} \right) - \frac{1}{\nu} \right) \\
 & - \frac{1}{\nu} \sum_{i=1}^n \left\{ \log \left(1 + \frac{1}{\nu} \left(\frac{u_i^-}{\sqrt{s_-}} \right)^{\frac{\nu}{\nu}} \right) I(x_i \geq \mu) \right\} \\
 & - \frac{\nu+1}{\nu} \sum_{i=1}^n \frac{-\frac{1}{\nu^{\frac{\nu}{\nu}}} \left(\frac{u_i^-}{\sqrt{s_-}} \right)^{\frac{\nu}{\nu}} + \frac{\nu ak}{\nu^{\frac{\nu}{\nu}} s_-} \left(\frac{u_i^-}{\sqrt{s_-}} \right)^{\frac{\nu}{\nu}} I(x_i \geq \mu)}{\left(1 + \frac{1}{\nu} \left(\frac{u_i^-}{\sqrt{s_-}} \right)^{\frac{\nu}{\nu}} \right)} \\
 & - \frac{1}{\nu} \sum_{i=1}^n \left\{ \log \left(1 + \frac{1}{\nu} \left(\frac{u_i^-}{\sqrt{s_+}} \right)^{\frac{\nu}{\nu}} \right) I(x_i < \mu) \right\} \\
 & - \frac{\nu+1}{\nu} \sum_{i=1}^n \frac{-\frac{1}{\nu^{\frac{\nu}{\nu}}} \left(\frac{u_i^-}{\sqrt{s_+}} \right)^{\frac{\nu}{\nu}} - \frac{\nu ak}{\nu^{\frac{\nu}{\nu}} s_+} \left(\frac{u_i^-}{\sqrt{s_+}} \right)^{\frac{\nu}{\nu}} I(x_i < \mu)}{\left(1 + \frac{1}{\nu} \left(\frac{u_i^-}{\sqrt{s_+}} \right)^{\frac{\nu}{\nu}} \right)}
 \end{aligned}$$

که در آنها $u_i^+ = x_i - \mu - a + k$ ، $u_i^- = x_i - \mu - a - k$ ، $\delta = s_-^{-\frac{\nu}{\nu}} D_{\nu} \left(\frac{a+k}{\sqrt{s_-}} \right) + s_+^{-\frac{\nu}{\nu}} D_{\nu} \left(\frac{k-a}{\sqrt{s_+}} \right)$ و $d_{\nu}^+ = d_{\nu} \left(\frac{k-a}{\sqrt{s_+}} \right)$ هستند. همانند توزیع BUN حل معادلات امتیاز توزیع BUST نیز به جواب‌های بسته منتهی نمی‌شوند و باید از روش‌های عددی برای حل آنها استفاده کرد.

۳.۳.۳ توزیع لاپلاس دومدی-یکمدی

یکی از تابع چگالی‌های نیمه‌دم‌سنگین و پرکاربرد در بین توزیع‌های آماری، توزیع لاپلاس است. از آن جا که توزیع لاپلاس متقارن است و برای داده‌های نامتقارن چندان کاربردی ندارد، تعمیم‌های مختلفی روی این توزیع، برای چوله کردن آن، معرفی شده‌اند. به‌عنوان

مثال، کانکر و مکادو^{۲۱} (۱۹۹۹)، آریال^{۲۲} و ناداراجا (۲۰۰۵)، نکوخو و علامت‌ساز (۲۰۱۲)، شمس و علامت‌ساز (۲۰۱۳) ییلماز^{۲۳} (۲۰۱۶) و شاه^{۲۴} و همکاران (۲۰۱۹) را ببینید. در این بخش، تعمیمی مانند تعمیم‌های ایجادشده برای نرمال و St بر روی توزیع لاپلاس انجام می‌دهیم. در قضیه ۳.۲.۳، $g(x)$ را تابع چگالی توزیع لاپلاس با پارامتر مکان صفر و پارامتر مقیاس $\frac{1}{k}$ و $h(x) = 1 + (u - \frac{a}{\sigma})^2$ در نظر می‌گیریم. تابع چگالی هدف به‌صورت زیر است:

$$f_{BUL}(x) = \frac{k}{\sigma c} \left(1 + \left(u - \frac{a}{\sigma} \right)^2 \right) e^{-k|u|}. \quad (14.3)$$

اگر متغیر تصادفی X دارای تابع چگالی (۱۴.۳) باشد، می‌گوییم X دارای توزیع لاپلاس دومی-یک‌مدی^{۲۵} BUL است و با نماد $X \sim \text{BUL}(\mu, \sigma, k, a)$ نمایش می‌دهیم. برای $a = 0$ ، چگالی (۱۴.۳) متقارن است که در این حالت آن را با نماد $X \sim \text{BUL}(\mu, \sigma, k)$ نمایش می‌دهیم. شکل تابع چگالی (۱۴.۳) برای برخی مقادیر پارامترها در نمودار ۷.۳ رسم شده است.

گزاره ۶.۳.۳. دو تا از مدهای BUL برای $k \leq 1$ برابر $\left(\frac{1}{k} \pm \sqrt{\frac{1}{k^2} - 1} \right)$ و $\mu + a \pm \sigma$ هستند.

تابع توزیع و گشتاورهای توزیع BUL

تابع توزیع تجمعی BUL برابر است با

$$F_{BUL}(x) = \begin{cases} \frac{k}{c} \left(\frac{u^2}{k} - \frac{2}{k} \left(\frac{a}{\sigma} + \frac{1}{k} \right) \left(u - \frac{1}{k} \right) + \left(\frac{a^2}{\sigma^2 k} + \frac{1}{k} \right) \right) e^{ku} & x \geq \mu \\ \frac{k}{c} \left(\frac{2}{k^2} \left(\frac{a}{\sigma} + \frac{1}{k} \right) + \left(\frac{a^2}{\sigma^2 k} + \frac{1}{k} \right) \right) + \frac{k}{c} \left(\left(\frac{a^2}{\sigma^2 k} + \frac{1}{k} \right) (1 - e^{-ku}) - \frac{u^2}{k} e^{-ku} \right) & x < \mu \\ -\frac{2k}{c} \left(\frac{a}{\sigma} - \frac{1}{k} \right) \left(\frac{1 - e^{-ku} - k u e^{-ku}}{k^2} \right) & \end{cases} \quad (15.3)$$

که در آن $c = 2 \left(1 + \frac{a^2}{\sigma^2} + \frac{2}{k^2} \right)$ و $u = \frac{x - \mu}{\sigma}$. برای محاسبه گشتاورهای BUL فرض کنید $Z = \frac{X - \mu}{\sigma}$ حالت استاندارد BUL باشد. می‌توان گفت $Z \sim \text{BUL}(0, 1, k, \frac{a}{\sigma})$ که در این حالت r -امین گشتاور Z برابر است با

$$E(Z^r) = \frac{2}{c} \left\{ \left(1 + \frac{a^2}{\sigma^2} \right) E(W_L^r) - 2 \frac{a}{\sigma} E(W_L^{r+1}) + E(W_L^{r+2}) \right\}$$

که در آن W_L متغیر تصادفی لاپلاس با پارامتر مکان صفر و مقیاس $\frac{1}{k}$ است. بدیهی است که s -امین گشتاور W_L برابر $\frac{s!}{k^s} (1 + (-1)^s)$ است.

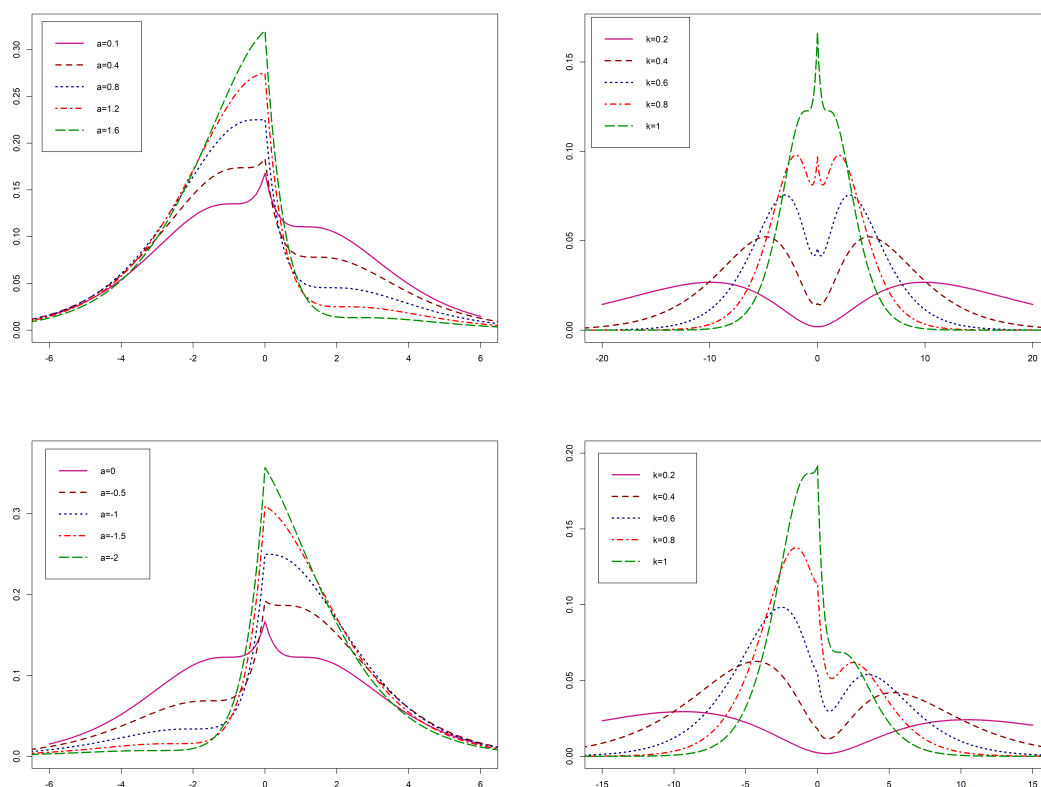
²¹Koenker and Machado

²² Aryal

²³Yilmaz

²⁴Shah

²⁵Bimodal–Unimodal Laplace



شکل ۷.۳: منحنی‌های (۱۴.۳) برای برخی مقادیر پارامترها. ($\mu = 0$ و $\sigma = 1$ و $a = 0$) (بالا سمت راست)، ($\mu = 0$ و $\sigma = 1$ و $k = 0$) (بالا سمت چپ)، ($\mu = 0$ و $\sigma = 1$ و $a = 0.5$) (پایین سمت راست) و ($\mu = 0$ و $\sigma = 1$ و $k = 1$) (پایین سمت چپ).

برآوردگرهای ماکسیمم درست‌نمایی پارامترهای توزیع BUL

بردار پارامترهای $\theta^{BUL} = (\mu, \sigma, k, a)'$ را برای توزیع BUL داریم. لگاریتم تابع درست‌نمایی توزیع BUL به‌صورت زیر است:

$$\begin{aligned} \ell &= \ell(\theta^{BUL}) = n \log \left(\frac{k}{\sqrt{2}\sigma} \right) - n \log \left(1 + \frac{a^2}{\sigma^2} + \frac{2}{k^2} \right) - k \sum_{i=1}^n \left| \frac{x_i - \mu}{\sigma} \right| \\ &\quad + \sum_{i=1}^n \log \left(1 + \left(\frac{x_i - \mu - a}{\sigma} \right)^2 \right). \end{aligned}$$

مؤلفه‌های بردار امتیاز $S_n^{BUL} = \left(\frac{\partial \ell}{\partial \mu}, \frac{\partial \ell}{\partial \sigma}, \frac{\partial \ell}{\partial k}, \frac{\partial \ell}{\partial a} \right)'$ نیز به صورت زیر محاسبه می‌شوند:

$$\begin{aligned}\frac{\partial \ell}{\partial \mu} &= \frac{k}{\sigma} \sum_{i=1}^n \text{sign}(x_i - \mu) - \frac{2}{\sigma} \sum_{i=1}^n \frac{\left(\frac{x_i - \mu - a}{\sigma}\right)}{1 + \left(\frac{x_i - \mu - a}{\sigma}\right)^2} \\ \frac{\partial \ell}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{2na^2}{\sigma^3 \left(1 + \frac{a^2}{\sigma^2} + \frac{2}{k^2}\right)} + \frac{k}{\sigma^2} \sum_{i=1}^n |x_i - \mu| - \frac{2}{\sigma} \sum_{i=1}^n \frac{\left(\frac{x_i - \mu - a}{\sigma}\right)^2}{1 + \left(\frac{x_i - \mu - a}{\sigma}\right)^2} \\ \frac{\partial \ell}{\partial k} &= \frac{n}{k} + \frac{4n}{k^3 \left(1 + \frac{a^2}{\sigma^2} + \frac{2}{k^2}\right)} - \sum_{i=1}^n \left| \frac{x_i - \mu}{\sigma} \right| \\ \frac{\partial \ell}{\partial a} &= -\frac{2na}{\sigma^2 \left(1 + \frac{a^2}{\sigma^2} + \frac{2}{k^2}\right)} - \frac{2}{\sigma} \sum_{i=1}^n \frac{\left(\frac{x_i - \mu - a}{\sigma}\right)}{1 + \left(\frac{x_i - \mu - a}{\sigma}\right)^2}.\end{aligned}$$

مانند مدل‌های قبلی حل $U_n^{BUL} = 0$ دارای جواب‌های بسته نیست و باید از روش‌های عددی برای حل آن استفاده کرد.

۴.۳ مثال واقعی

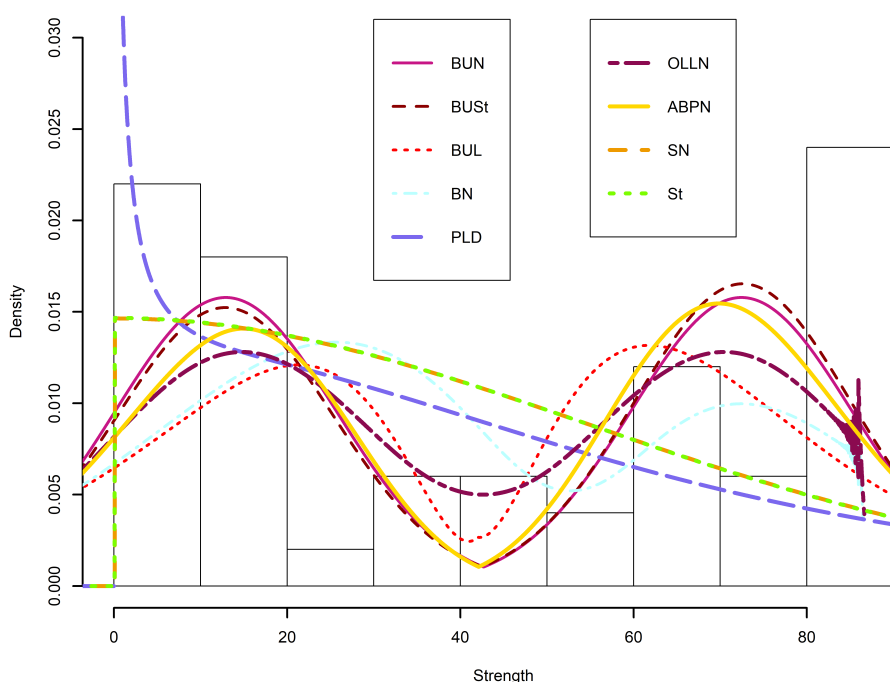
در این بخش عملکرد BUN، BUST و BUL را با سه مثال واقعی ارزیابی می‌کنیم. دو مثال واقعی دومی هستند و یکی از مثال‌ها یک‌مدی چوله است.

مثال ۱.۴.۳. اولین مجموعه داده طول عمر 50° دستگاه است که توسط آرست^{۲۶} (۱۹۸۷) به کار گرفته شده است. اخیراً باکوچ و همکاران (۲۰۱۹) نیز این داده‌ها را مورد بررسی قرار دادند. باکوچ و همکاران (۲۰۱۹) توزیع PLD را بر روی این مجموعه داده برازش دادند. توزیع PLD نیز خاصیت دومی بودن را دارد ولی برآوردهایی که باکوچ و همکاران (۲۰۱۹) از برازش PLD به دست آوردند، در ناحیه دومی PLD قرار ندارند. برای این مجموعه داده توزیع‌های BUN، BUST و BUL را برازش دادیم. توزیع‌های BN، PLD، OLLN، ABPN، SN و SSt نیز به عنوان توزیع‌های رقیب بر روی این داده‌ها برازش شده‌اند. نتایج برآورد ML از قبیل مقادیر برآورد پارامترها، AIC، BIC، مقادیر آماره KS و p -مقدار متناظر آماره‌های KS در جدول ۱.۳ گزارش شده‌اند. با توجه به این که مقادیر کوچک‌تر AIC و BIC برتری مدل را نشان می‌دهند، بر اساس معیارهای AIC و BIC گزارش شده در جدول ۱.۳، توزیع BUN برترین مدل برازش شده برای این مجموعه داده است. طبق ستون مربوط به آماره KS در جدول ۱.۳ تمامی مدل‌ها به درستی برازش شده‌اند. در شکل ۸.۳ بافت نگار داده‌ها و شکل توابع چگالی برازش شده به ازای مقادیر پارامترهای برآورده شده رسم شده‌اند. با توجه به نتایج شکل ۸.۳، توزیع‌های BUN، BUST، BUL، BN، OLLN و ABPN دومی و توزیع‌های PLD، SN و St یک‌مدی برازش شده‌اند.

²⁶ Aarset

جدول ۱.۳: نتایج برآورد ML برای مجموعه داده مثال ۱.۴.۳

مدل	برآوردها	AIC	BIC	KS	-p مقدار
BUN(μ, σ, k)	۲/۳۳۲۵ ۱۲/۷۶۳۲ ۴۲/۶۸۰۰	۴۶۷/۷۸	۴۷۳/۵۲	۰/۰۷۲۸	۰/۵۹
BUST(μ, σ, k, a, ν)	۲۹/۸۷۲۲ ۱۲/۶۳۵۹ ۴۲/۵۰۹۹	۴۷۱/۹۲	۴۸۱/۴۸	۰/۰۷۷۱	۰/۵۵
BUL(μ, σ, k, a)	۰/۳۴۰۴ ۳/۵۵۰۴ ۴۲/۰۱۶۶	۰/۴۴۲۲	۴۹۵/۰۸	۰/۰۱۰۶۷	۰/۳۲
BN(μ, σ, a, b)	۴/۱۰۵۹ ۵۲/۴۷۱۷	۰/۰۳۸۰	۴۸۱/۹۸	۰/۰۱۰۹۷	۰/۳۰
PLD(ν, ρ, ζ)	۱۷۱۰/۲۷۴۳ ۰/۱۵۷۲ ۰/۰۲۴۶	۴۷۴/۰۲	۴۷۹/۷۵	۰/۰۱۵۱۰	۰/۱۰
OLLN(α, μ, σ)	۰/۰۶۶۵ ۵/۳۰۹۷ ۴۲/۶۱۹۳	۴۶۹/۱۹	۴۷۴/۹۳	۰/۰۰۹۰۳	۰/۴۴
ABPN($\alpha, \beta, \mu, \sigma$)	۴۲/۱۱۷۹ ۰/۰۴۷۳ ۶/۸۸۷۹	۲۲/۲۶۵۶	۴۷۴/۶۰	۴۸۲/۲۵	
SN(μ, σ, λ)	۲۳۵۱۱۶ ۵۴/۴۹۶۱ ۰/۰۹۹۱	۴۷۸/۴۵	۴۸۴/۱۹	۰/۱۳۷۵	۰/۱۵
SSt($\mu, \sigma, \lambda, \nu$)	۵۴/۳۹۴۷ ۰/۰۸۶۹	۴۸۰/۴۸	۴۸۸/۱۳	۰/۱۳۷۳	۰/۱۵



شکل ۸.۳: بافت نگار داده‌ها و منحنی توابع چگالی برازش شده مدل‌های رقیب به‌ازای مقادیر پارامترهای برآورد شده برای مجموعه داده مثال ۱.۴.۳.

مثال ۲.۴.۳. مشاهدات متغیر b.weight از مجموعه داده بررسی شده در بولفرونی و همکاران (۲۰۱۸) به‌عنوان دومین مثال واقعی در این قسمت مورد بررسی قرار گرفته است. این مجموعه داده شامل ۵۰۰ مشاهده است و b.weight وزن سونوگرافی (وزن جنین به گرم) است. این مجموعه داده در آدرس اینترنتی

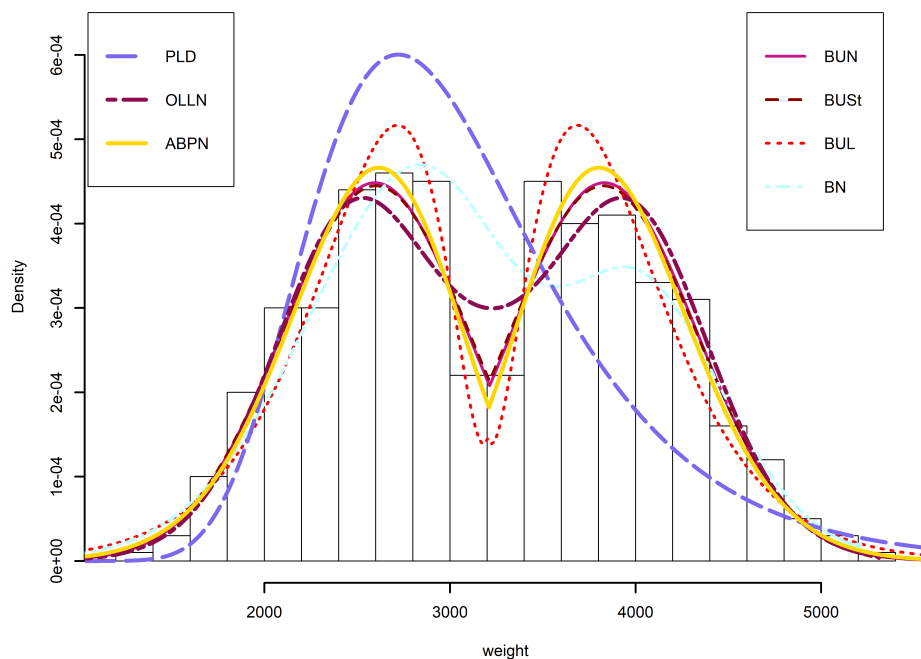
<http://www.mat.uda.cl/hsalinas/data/weight.rar>

برای دانلود موجود هستند. بر اساس بافت نگار مشاهدات، داده‌ها دومی متقارن هستند. مدل‌های برازش شده برای این داده‌ها عبارتند از BUN، BUST، BUL، BN، OLLN، PLD و ABPN. مانند مثال ۱.۴.۳ نتایج برآورد ML شامل مقادیر برآورد پارامترها، AIC، BIC، مقادیر آماره KS و -p مقدار متناظر آماره‌های KS در جدول ۲.۳ گزارش شده‌اند. براساس

معیارهای AIC و BIC در جدول ۲.۳، توزیع BUN عملکرد بهتری نسبت به سایر مدل‌ها دارد. بر اساس آماره‌ی KS جدول ۲.۳ تمامی مدل‌ها به‌نیکویی برازش شده‌اند. بافت نگار داده‌ها و منحنی‌های توابع چگالی برازش‌شده در نمودار ۹.۳ رسم شده‌اند. با توجه به نمودار ۹.۳ همه توزیع‌های برازش‌شده به جز توزیع PLD دومی برازش شده‌اند.

جدول ۲.۳: نتایج برآورد ML برای مجموعه داده مثال ۲.۴.۳

مدل	برآوردها	AIC	BIC	KS	p-مقدار
$BUN(\mu, \sigma, k)$	۱/۲۴۰۲ ۴۹۸/۴۰۴۰ ۳۲۱۳/۳۱۶۶	۸۰۸۷/۴۴	۸۱۰۰/۰۸	۰/۰۲۰۹	۰/۶۴
$BUST(\mu, \sigma, k, \nu)$	۶۱۲/۶۱۶۸ ۵۰۵/۲۵۳۱ ۳۲۱۲/۶۵۲۹	۸۰۸۹/۶۲	۸۱۰۶/۴۸	NAN	NA
$BUL(\mu, \sigma, k)$	۰/۳۹۷۶ ۱۰۰/۵۱۶۵ ۳۲۰۱/۲۸۵۷	۸۰۹۹/۲۱	۸۱۱۱/۸۵	۰/۰۱۴۹	۰/۸
$BN(\mu, \sigma, a, b)$	۰/۰۷۷۳ ۲۰۸/۷۲۷۱ ۵۳۵۰۶/۵۹۹	۰/۱۳۲۹	۸۱۳۴/۴۸	۰/۰۲۷۳	۰/۴۷
$PLD(\nu, \rho, \zeta)$	۱۷۱۰/۳۴۹۶ ۰/۰۰۴۸ ۰/۰۰۱۶	۸۲۲۴/۱۵	۸۲۳۶/۷۹	۰/۰۳۸۵	۰/۲۳
$OLLN(\alpha, \mu, \sigma)$	۰/۱۸۹۴ ۲۵۲/۱۸۸۸ ۳۲۲۶/۲۴۰۴	۸۰۹۶/۷۰	۸۱۰۹/۳۵	۰/۰۳۱۹	۰/۳۶
$ABPN(\alpha, \mu, \sigma)$	۶۶۰/۶۸۶۰ ۳۲۰۹/۴۵۷۰ ۳/۷۴۷۱	۸۰۸۸/۲۱	۸۱۰۰/۸۵		



شکل ۹.۳: بافت نگار داده‌ها و منحنی توابع چگالی برازش‌شده مدل‌های رقیب به‌ازای مقادیر پارامترهای برآوردشده برای مجموعه داده مثال ۲.۴.۳.

مثال ۳.۴.۳. مجموعه داده سوم مربوط به توده بدنی ورزشکاران استرالیا است. این داده‌ها توسط کوک و ویسبرگ^{۲۷} (۲۰۰۹)، اصغرزاده و همکاران (۲۰۱۳)، و ساستری و بهاتی^{۲۸} (۲۰۱۶) بررسی شده‌اند.

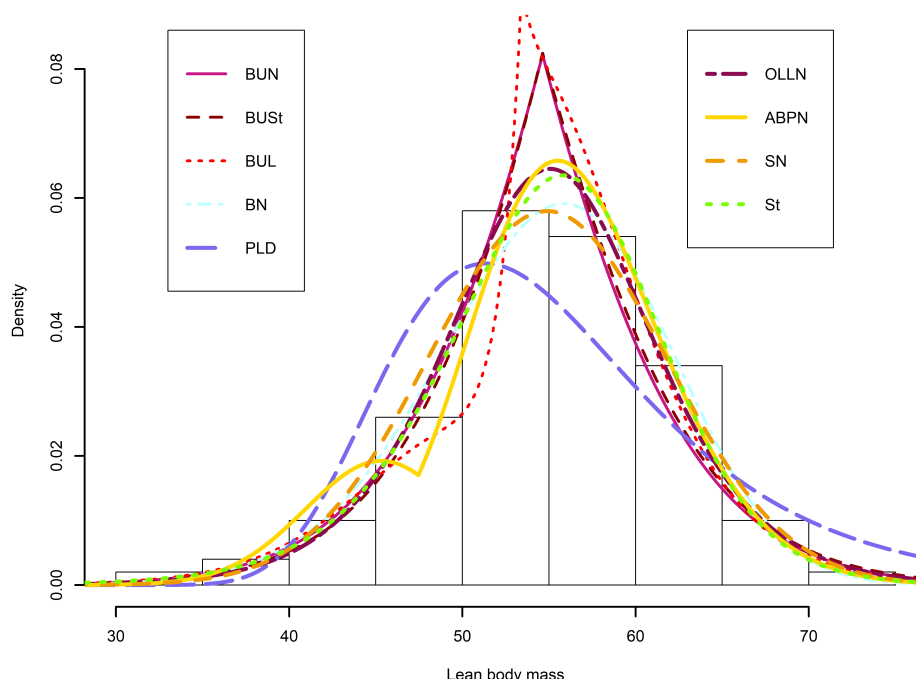
^{۲۷}Cook and Weisberg

^{۲۸} Sastry and Bhati

از مزایای BUN، BUST، و BUL این بود که برای داده‌های یک مدی چوله هم قابل برازش بودند. این مجموعه داده که دارای یک مد است و از طرفی دارای چولگی است، برای نشان دادن این مزیت انتخاب شده است. مدل‌های برازش شده برای این داده‌ها عبارتند از BUN، BUST، BUL، BN، OLLN، SN، SSt، و ABPN. نتایج برآورد ماکسیمم درست‌نمایی در جدول ۳.۳ گزارش شده‌اند. با توجه به جدول ۳.۳ توزیع BUL بهترین عملکرد را نسبت به معیارهای AIC و BIC دارد. مقادیر آماره KS نیکویی برازش تمام مدل‌ها را تأیید می‌کنند. بافت نگار داده‌ها و منحنی توابع چگالی برازش شده در نمودار ۱۰.۳ رسم شده‌اند. با توجه به نمودار ۱۰.۳ همه توزیع‌های برازش شده به جز توزیع ABPN یک‌مدی برازش شده‌اند.

جدول ۳.۳: نتایج برآورد ML برای مجموعه داده مثال ۳.۴.۳

مدل	برآوردها	AIC	BIC	KS	p-مقدار
$BUN(\mu, \sigma, k, a)$	۱۱/۷۳۵۷ ۵۴/۶۲۹۹ -۱/۴۹۰۱ ۰/۷۹۹۶	۶۷۳/۳۹	۶۸۳/۷۱	۰/۰۳۳۴	۰/۸۱
$BUST(\mu, \sigma, k, a, \nu)$	۱۰/۰۰۱۶ ۵۴/۶۲۹۵ -۱۲/۶۹۳۴ ۰/۸۰۳۴	۶۷۵/۵۷	۶۸۸/۶۰	۰/۰۴۱۶	۰/۷۱
$BUL(\mu, \sigma, k, a)$	۳/۲۲۰۳ ۵۳/۴۰۹۷ ۱/۲۲۳۹ -۱/۵۹۲۲	۶۶۶/۶۰	۶۷۷/۰۲	۰/۰۴۶۱	۰/۶۵
$BN(\mu, \sigma, a, b)$	۱۰/۲۶۶۲ ۶۵/۸۹۷۰ ۱/۰۴۹۵ ۴/۵۷۴۹	۶۷۶/۱۸	۶۸۶/۶۰	۰/۰۷۰۹	۰/۳۷
$PLD(\nu, \rho, \zeta)$	۰/۰۴۴۵ ۰/۱۳۴۰ ۱۷۱۰/۳۳۴۹	۷۰۲/۶۶	۷۱۰/۴۷	۰/۱۰۷۶	۰/۱۰
$OLLN(\alpha, \mu, \sigma)$	۱۷/۳۱۳۰ ۵۵/۰۷۰۷ ۲/۷۹۹۹	۶۷۳/۹۲	۶۸۱/۷۳	۰/۰۵۳۴	۰/۵۶
$ABPN(\alpha, \beta, \mu, \sigma)$	۱/۲۶۵۵ ۳/۴۳۲۸ ۴۷/۴۶۴۶	۶۷۴/۳۱	۶۸۴/۷۳	۰/۰۵۸۰	۰/۵۱
$SN(\mu, \sigma, \lambda)$	۶/۸۸۳۲ ۵۴/۷۸۰۵ ۰/۰۲۰۱	۶۷۵/۷۳	۶۸۳/۵۴	۰/۰۵۸۰	۰/۵۱
$SSt(\mu, \sigma, \lambda, \nu)$	۷/۲۳۲۹ ۵۹/۰۸۱۶ -۰/۹۰۲۱	۶۷۵/۴۸	۶۸۵/۹۰	۰/۰۶۲۶	۰/۴۶



شکل ۱۰.۳: بافت نگار داده‌ها و منحنی توابع چگالی برازش شده مدل‌های مختلف به‌ازای مقادیر پارامترهای برآورد شده برای مجموعه داده مثال ۳.۴.۳.

۵.۳ نتیجه‌گیری

در این فصل قضایایی را برای تعمیم توزیع‌ها به نسخه‌های منعطف‌تر و کاراتر بیان کردیم. توزیع‌هایی ساختیم که دارای پارامتر شکل برای کنترل دومدی بودن و فاصله دو مد توزیع هستند. برخی از توزیع‌های ساخته‌شده را با جزئیات بیشتری بررسی کردیم و آنها را برروی مجموعه داده‌های مختلفی که چوله یا دومدی بودند، برازش دادیم و نشان دادیم که نسبت به توزیع‌های رقیب برازش بهتری دارند.

فصل ۴

مدل‌های رگرسیونی مبتنی بر توزیع

BUN

در این فصل توزیع BUN، که در فصل قبل معرفی شده است، برای مدل‌بندی خطاها در مدل رگرسیون خطی به کار گرفته می‌شود. در بخش اول، رگرسیون خطی با پذیره چوله بودن توزیع خطا یا وجود داده‌پرت یک‌طرفه در مانده‌ها را مورد بررسی قرار می‌دهیم. در بخش دوم، با توسعه BUN، به حالت چندمتغیره، کاربرد آن را در رگرسیون فضایی مورد مطالعه قرار می‌دهیم. از این فصل مقالات اونق، باغیشنی و نزاکتی (۱۳۹۶ ب) و اونق، نزاکتی و باغیشنی (۱۳۹۶ الف) استخراج شده‌اند.

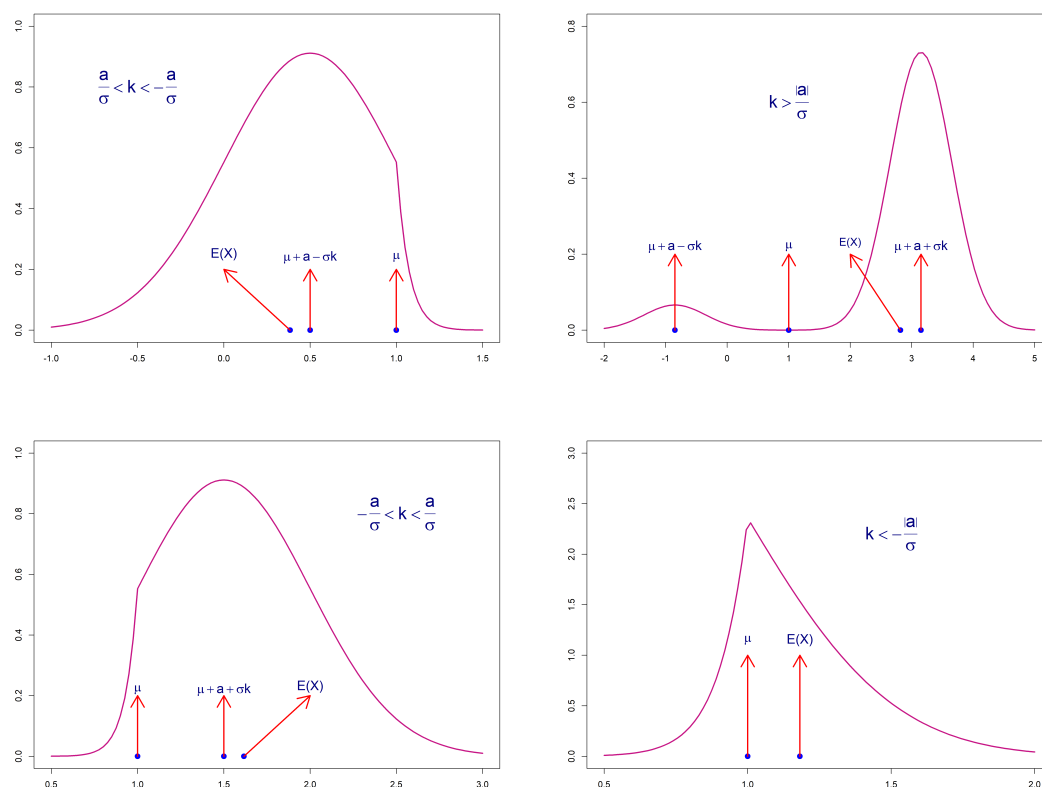
۱.۴ خانواده‌ای از مدل‌های رگرسیون خطی برای مانده‌های شامل داده پرت یک طرفه یا چوله

یکی از ابزارهای پرطرفدار در علوم مختلف مانند شیمی، زمین‌شناسی، اقتصاد، روانشناسی، صوت شناسی و علوم پزشکی، مدل رگرسیون خطی است. پذیره کلاسیک مدل رگرسیون خطی، نرمال در نظر گرفتن توزیع خطاها است که در همه داده‌ها مناسب و کافی نیست. عدم کفایت این پذیره زمانی رخ می‌دهد که خطاها شامل داده پرت یا دارای چولگی باشند.

در مواجهه با داده‌های چوله و داده‌های شامل داده پرت راهکارهای متفاوتی وجود دارند. هنگامی که با داده‌های شامل داده پرت سروکار داریم، می‌توان از توزیع‌های دم‌سنگین مانند St و لاپلاس برای بهبود مدل بهره گرفت. همچنین روش‌های ناپارامتری متعددی مانند رگرسیون چندکی^۱ (کنکر و باست^۲ (۱۹۸۷)) و برآوردگرهای M^۳ (هوبر^۴ (۱۹۶۴)) و (یانگ^۵ و همکاران (۲۰۰۶)) در این زمینه وجود دارند. از طرفی برای داده‌های چوله، سازوکارهای متعددی می‌توان عنوان کرد. به‌عنوان مثال توزیع‌های بیضوی-چوله^۶ (جنتن (۲۰۰۴)) و توزیع چوله-مقارن^۷ (ربیو^۸ و جنتن (۲۰۱۶)) را می‌توان نام برد. در هیچ‌کدام از راهکارهای ارائه‌شده، روش یا مدلی وجود ندارد که هم برای داده‌های شامل داده پرت و هم داده‌های دارای چولگی مناسب باشد. برتری توزیع BUN نسبت به سایر توزیع‌های رقیب این است که BUN به جز پارامترهای مکان و مقیاس دارای دو پارامتر شکل است که چولگی، دومدی بودن و فاصله دو مد را کنترل می‌کنند. این ویژگی‌ها چراغ سبزی است که بتوانیم از BUN برای داده‌های نامقارن چوله یک مدی یا داده‌های شامل داده پرت یک‌طرفه بهره ببریم. این توزیع با انعطاف‌پذیری بالایی که دارد، برای داده‌های چوله یک‌مدی در ناحیه یک‌مدی توزیع برازش می‌شود که می‌توان در این حالت از رگرسیون میانگین بهره برد. این توزیع برای داده‌های شامل داده پرت یک‌طرفه در ناحیه دومدی توزیع برازش می‌شود که بهترین گزینه برای توسعه یک رگرسیون مد^۹ است. برای تشریح بیشتر این مطلب منحنی‌های توزیع BUN را برای برخی از مقادیر پارامترها در شکل ۱.۴ نشان داده‌ایم. از روی این نمودارها به سادگی انعطاف این توزیع در مدل‌بندی داده‌ها با ویژگی‌های اشاره‌شده معلوم می‌شود. همان‌طور که در شکل ۱.۴ مشاهده می‌کنید، شکل چگالی در ناحیه دومدی BUN که برابر $k\sigma > |a|$ است، به‌گونه‌ای است که می‌توان از مد $\mu + a - \sigma k$ به‌عنوان مد داده‌های پرت (یک‌طرفه) صرف‌نظر کرد و مد $\mu + a + \sigma k$ را به‌عنوان برآورد مکان در نظر گرفت. با توجه به شکل ۱.۴ به‌نظر می‌رسد که برای حالت $a < k\sigma < -a$ و $-a < k\sigma < a$ نیز به‌ترتیب مدهای $\mu + a - \sigma k$ و $\mu + a + \sigma k$ مکان توزیع را بهتر بیان می‌کنند. هرچند در این دو حالت میانگین و مد توزیع اختلاف چندانی ندارند. بنابراین هم رگرسیون میانگین و هم رگرسیون مد می‌توانند مورد توجه قرار بگیرند. برای حالت آخر، یعنی $k\sigma < -|a|$ ، میانگین توزیع بهترین گزینه ممکن برای برآورد مکان است.

^۱Quantile Regression^۲Koenker and Bassett^۳M-Estimation^۴Huber^۵Yang^۶Skew Elliptical^۷Skew Symmetric^۸Rubio^۹Modal Regression

خانواده‌ای از مدل‌های رگرسیون خطی برای مانده‌های شامل داده پرت یک طرفه یا چوله ۹۳



شکل ۱.۴: منحنی‌های تابع چگالی BUN برای $\mu = 1$, $\sigma = 0.5$, $k = 4$ و $a = 0.15$ (بالا سمت راست)، $\mu = 1$, $\sigma = 0.5$, $k = -3$ و $a = -2$ (بالا سمت چپ)، $\mu = 1$, $\sigma = 0.5$, $k = -3$ و $a = 2$ (پایین سمت راست) و $\mu = 1$, $\sigma = 0.5$, $k = -3$ و $a = 1$ (پایین سمت چپ)

۱.۱.۴ مدل رگرسیون خطی با جمله خطای BUN

مدل رگرسیون خطی

$$Y_i = \beta' x_i + \epsilon_i \quad i = 1, 2, \dots, n \quad (1.4)$$

را در نظر بگیرید که در آن Y_i متغیر پاسخ، $\beta = (\beta_1, \dots, \beta_p)'$ بردار ضرایب، x_i متغیرهای تبیینی و $\epsilon_i \sim \text{BUN}(0, \sigma, k, a)$ خطاهای تصادفی است که می‌توان نوشت $Y_i \sim \text{BUN}(\beta' x_i, \sigma, k, a)$. فرض کنید $\theta = (\beta', \sigma, k, a)'$ بردار پارامترهای مدل (۱.۴) باشد. تابع لگ درستنمایی مدل (۱.۴) به صورت زیر است:

$$\ell(\theta) = -\frac{nk^2}{2} - n \log(\sqrt{2\pi}\sigma\delta) + k \sum_{i=1}^n \left| \frac{y_i - \beta' x_i}{\sigma} \right| - \frac{1}{2} \sum_{i=1}^n \left(\frac{y_i - \beta' x_i - a}{\sigma} \right)^2.$$

مؤلفه‌های بردار امتیاز $S_n = \left(\left(\frac{\partial \ell}{\partial \beta} \right)', \frac{\partial \ell}{\partial \sigma}, \frac{\partial \ell}{\partial k}, \frac{\partial \ell}{\partial a} \right)'$ برابر

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_j} &= -\frac{k}{\sigma} \sum_{i=1}^n x_{ij} \text{sign}(y_i - \beta' x_i) + \frac{1}{\sigma} \sum_{i=1}^n x_{ij} \left(\frac{y_i - \beta' x_i - a}{\sigma} \right) \\ \frac{\partial \ell}{\partial \sigma} &= -\frac{n}{\sigma} + n \frac{kas}{\sigma^3} - \frac{k}{\sigma^3} \sum_{i=1}^n |y_i - \beta' x_i| + \frac{1}{\sigma} \sum_{i=1}^n \left(\frac{y_i - \beta' x_i - a}{\sigma} \right)^2 \\ \frac{\partial \ell}{\partial k} &= -nk - n \frac{as}{\sigma} - n\rho + \sum_{i=1}^n \left| \frac{y_i - \beta' x_i}{\sigma} \right| \\ \frac{\partial \ell}{\partial a} &= -n \frac{ks}{\sigma} + \frac{1}{\sigma} \sum_{i=1}^n \left(\frac{y_i - \beta' x_i - a}{\sigma} \right) \end{aligned}$$

است، که در آن

$$\begin{aligned} \delta &= e^{\frac{ka}{\sigma}} \Phi\left(k + \frac{a}{\sigma}\right) + e^{-\frac{ka}{\sigma}} \Phi\left(k - \frac{a}{\sigma}\right) \\ s &= \frac{e^{\frac{ka}{\sigma}} \Phi\left(k + \frac{a}{\sigma}\right) - e^{-\frac{ka}{\sigma}} \Phi\left(k - \frac{a}{\sigma}\right)}{\delta} \\ \rho &= \frac{2e^{\frac{ka}{\sigma}} \phi\left(k + \frac{a}{\sigma}\right)}{\delta}. \end{aligned}$$

از حل دستگاه معادلات $S_n = 0$ برآوردهای درست‌نمایی ماکسیمم پارامترهای مدل به دست می‌آیند. اما از آن‌جا که جواب‌های این دستگاه شکل بسته ندارند، برآوردهای ML را باید به کمک روش‌های عددی تکراری به دست آورد. برای بررسی رفتار برآوردهای ML پارامترهای مدل ابتدا باید ماتریس اطلاع فیشر مشاهده‌شده را محاسبه کرد. مؤلفه‌های ماتریس هسی

$$\ell_n^{(\gamma)}(\theta) = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta \partial \beta'} & \frac{\partial^2 \ell}{\partial \beta \partial \sigma} & \frac{\partial^2 \ell}{\partial \beta \partial k} & \frac{\partial^2 \ell}{\partial \beta \partial a} \\ \left(\frac{\partial^2 \ell}{\partial \beta \partial \sigma} \right)' & \frac{\partial^2 \ell}{\partial \sigma^2} & \frac{\partial^2 \ell}{\partial \sigma \partial k} & \frac{\partial^2 \ell}{\partial \sigma \partial a} \\ \left(\frac{\partial^2 \ell}{\partial \beta \partial k} \right)' & \frac{\partial^2 \ell}{\partial \sigma \partial k} & \frac{\partial^2 \ell}{\partial k^2} & \frac{\partial^2 \ell}{\partial k \partial a} \\ \left(\frac{\partial^2 \ell}{\partial \beta \partial a} \right)' & \frac{\partial^2 \ell}{\partial \sigma \partial a} & \frac{\partial^2 \ell}{\partial k \partial a} & \frac{\partial^2 \ell}{\partial a^2} \end{bmatrix}$$

با مشتق‌گیری از بردار امتیاز به صورت زیر نتیجه می‌شوند:

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \beta_j \partial \beta_l} &= -\frac{1}{\sigma^2} \sum_{i=1}^n x_{ij} x_{il} \rightarrow \frac{\partial^2 \ell}{\partial \beta' \partial \beta} = -\frac{1}{\sigma^2} \mathbf{X}_n^T \mathbf{X}_n \\ \frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} &= \frac{k}{\sigma^3} \sum_{i=1}^n x_{ij} \text{sign}(z_i) - \frac{2}{\sigma^3} \sum_{i=1}^n x_{ij} \left(\frac{z_i - a}{\sigma} \right) \\ \frac{\partial^2 \ell}{\partial \beta_j \partial k} &= -\frac{1}{\sigma} \sum_{i=1}^n x_{ij} \text{sign}(z_i) \\ \frac{\partial^2 \ell}{\partial \beta_j \partial a} &= -\frac{1}{\sigma^2} \sum_{i=1}^n x_{ij} \\ \frac{\partial^2 \ell}{\partial \sigma^2} &= \frac{n}{\sigma^3} + \frac{2k}{\sigma^3} \sum_{i=1}^n |z_i| - \frac{2}{\sigma^3} \sum_{i=1}^n \frac{(z_i - a)^2}{\sigma^2} + n \frac{ka}{\sigma^3} (kas^2 - ka - a\rho - 2\sigma s) \end{aligned}$$

خانواده‌ای از مدل‌های رگرسیون خطی برای مانده‌های شامل داده پرت یک طرفه یا چوله ۹۵

$$\begin{aligned}\frac{\partial^{\gamma} \ell}{\partial \sigma \partial k} &= -\frac{1}{\sigma^{\gamma}} \sum_{i=1}^n |z_i| + \frac{na}{\sigma^{\gamma}} \left(s + k \left(\frac{a}{\sigma \delta} - \frac{as^{\gamma}}{\sigma} + \rho s \right) \right) \\ \frac{\partial^{\gamma} \ell}{\partial \sigma \partial a} &= -\frac{\gamma}{\sigma^{\gamma}} \sum_{i=1}^n (z_i - a) + \frac{nk}{\sigma^{\gamma}} \left(s + a \left(\frac{k}{\sigma} + \frac{\rho}{\delta} - \frac{ks^{\gamma}}{\sigma} \right) \right) \\ \frac{\partial^{\gamma} \ell}{\partial k^{\gamma}} &= -n - \frac{na}{\sigma} \left(\frac{a}{\sigma \delta} - \frac{as^{\gamma}}{\sigma} \right) + n\rho(k + \rho) \\ \frac{\partial^{\gamma} \ell}{\partial k \partial a} &= -\frac{n}{\sigma} \left(s + a \left(\frac{k}{\sigma} + \frac{\rho}{\delta} - \frac{ks^{\gamma}}{\sigma} \right) \right) + n\rho \left(\frac{a}{\sigma^{\gamma}} + \frac{ks}{\sigma} \right) \\ \frac{\partial^{\gamma} \ell}{\partial a^{\gamma}} &= -\frac{n}{\sigma^{\gamma}} - \frac{nk}{\sigma} \left(\frac{k}{\sigma} + \frac{\rho}{\delta} - \frac{ks^{\gamma}}{\sigma} \right)\end{aligned}$$

و با گرفتن امید ریاضی از ماتریس هسی، ماتریس اطلاع فیشر به صورت زیر به دست می‌آید:

$$B_n(\theta) = -E \left(\frac{\partial^{\gamma} \ell}{\partial \theta \partial \theta^{\gamma}} \right) = - \begin{bmatrix} E \left(\frac{\partial^{\gamma} \ell}{\partial \beta \partial \beta^{\gamma}} \right) & E \left(\frac{\partial^{\gamma} \ell}{\partial \beta \partial \sigma} \right) & E \left(\frac{\partial^{\gamma} \ell}{\partial \beta \partial k} \right) & E \left(\frac{\partial^{\gamma} \ell}{\partial \beta \partial a} \right) \\ E \left(\frac{\partial^{\gamma} \ell}{\partial \beta \partial \sigma} \right)^{\gamma} & E \left(\frac{\partial^{\gamma} \ell}{\partial \sigma^{\gamma}} \right) & E \left(\frac{\partial^{\gamma} \ell}{\partial \sigma \partial k} \right) & E \left(\frac{\partial^{\gamma} \ell}{\partial \sigma \partial a} \right) \\ E \left(\frac{\partial^{\gamma} \ell}{\partial \beta \partial k} \right)^{\gamma} & E \left(\frac{\partial^{\gamma} \ell}{\partial \sigma \partial k} \right) & E \left(\frac{\partial^{\gamma} \ell}{\partial k^{\gamma}} \right) & E \left(\frac{\partial^{\gamma} \ell}{\partial k \partial a} \right) \\ E \left(\frac{\partial^{\gamma} \ell}{\partial \beta \partial a} \right)^{\gamma} & E \left(\frac{\partial^{\gamma} \ell}{\partial \sigma \partial a} \right) & E \left(\frac{\partial^{\gamma} \ell}{\partial k \partial a} \right) & E \left(\frac{\partial^{\gamma} \ell}{\partial a^{\gamma}} \right) \end{bmatrix}$$

که مؤلفه‌های آن به صورت زیر محاسبه می‌شوند:

$$\begin{aligned}-E \left(\frac{\partial^{\gamma} \ell}{\partial \beta \partial \beta^{\gamma}} \right) &= \frac{1}{\sigma^{\gamma}} \mathbf{X}_n^T \mathbf{X}_n \\ -E \left(\frac{\partial^{\gamma} \ell}{\partial \beta \partial \sigma} \right) &= \frac{ks}{\sigma^{\gamma}} \left(\sum_{i=1}^n \mathbf{x}_{i1}, \sum_{i=1}^n \mathbf{x}_{i\gamma}, \dots, \sum_{i=1}^n \mathbf{x}_{ip} \right)^T \\ -E \left(\frac{\partial^{\gamma} \ell}{\partial \beta \partial k} \right) &= \frac{s}{\sigma} \left(\sum_{i=1}^n \mathbf{x}_{i1}, \sum_{i=1}^n \mathbf{x}_{i\gamma}, \dots, \sum_{i=1}^n \mathbf{x}_{ip} \right)^T \\ -E \left(\frac{\partial^{\gamma} \ell}{\partial \beta \partial a} \right) &= \frac{1}{\sigma^{\gamma}} \left(\sum_{i=1}^n \mathbf{x}_{i1}, \sum_{i=1}^n \mathbf{x}_{i\gamma}, \dots, \sum_{i=1}^n \mathbf{x}_{ip} \right)^T \\ -E \left(\frac{\partial^{\gamma} \ell}{\partial \sigma^{\gamma}} \right) &= n \frac{1}{\sigma^{\gamma}} (\gamma + k^{\gamma} + k\rho) - n \frac{ka^{\gamma}}{\sigma^{\gamma}} (ks^{\gamma} - k - \rho) \\ -E \left(\frac{\partial^{\gamma} \ell}{\partial \sigma \partial k} \right) &= n \frac{1}{\sigma} (k + \rho) - n \frac{ka}{\sigma^{\gamma}} \left(\frac{a}{\sigma \delta} - \frac{as^{\gamma}}{\sigma} + \rho s \right) \\ -E \left(\frac{\partial^{\gamma} \ell}{\partial \sigma \partial a} \right) &= n \frac{ks}{\sigma^{\gamma}} - n \frac{ka}{\sigma^{\gamma}} \left(\frac{k}{\sigma} + \frac{\rho}{\delta} - \frac{ks^{\gamma}}{\sigma} \right)\end{aligned}$$

$$\begin{aligned} -E\left(\frac{\partial^2 \ell}{\partial k^2}\right) &= n + \frac{na}{\sigma} \left(\frac{a}{\sigma\delta} + \frac{as^2}{\sigma} \right) - n\rho(k + \rho) \\ -E\left(\frac{\partial^2 \ell}{\partial k \partial a}\right) &= \frac{n}{\sigma} \left(s + a \left(\frac{k}{\sigma} + \frac{\rho}{\delta} - \frac{ks^2}{\sigma} \right) \right) - n\rho \left(\frac{a}{\sigma^2} + \frac{ks}{\sigma} \right) \\ -E\left(\frac{\partial^2 \ell}{\partial a^2}\right) &= \frac{n}{\sigma^2} + \frac{nk}{\sigma} \left(\frac{k}{\sigma} + \frac{\rho}{\delta} - \frac{ks^2}{\sigma} \right). \end{aligned}$$

اکنون قبل از بیان قضیه اصلی، دو گزاره را مطرح می‌کنیم.

قضیه ۱.۱.۴. اگر $\lim_{n \rightarrow \infty} n^{-1} X_n' X_n = C$ ، که در آن C یک ماتریس متناهی و معین مثبت است، آن‌گاه $\hat{\theta}$ دارای سازگاری ضعیف است و

$$n^{-\frac{1}{2}} (\hat{\theta} - \theta) \xrightarrow{d} N_{p+3} \left(0, I(\theta)^{-1} \right).$$

برهان. برای برهان به پیوست ۳.آ مراجعه کنید. □

منظور از ماتریس متناهی ماتریسی است که همه‌ی مؤلفه‌های آن متناهی هستند.

رگرسیون میانگین

اگر در مدل خطی (۱.۴)، توزیع خطا را BUN فرض کنیم و برآوردهای پارامترهای توزیع در ناحیه $k\sigma < -|a|$ قرار گیرند، به این معنی است که مانده‌ها چوله هستند و گزینه بهتر رگرسیون میانگین است. رگرسیون میانگین را می‌توان به صورت

$$E(Y_i) = \beta' x_i + E(\epsilon_i)$$

تعریف کرد که در آن

$$E(\epsilon_i) = \frac{\left(k + \frac{a}{\sigma}\right) e^{\frac{ka}{\sigma}} \Phi\left(k + \frac{a}{\sigma}\right) - \left(k - \frac{a}{\sigma}\right) e^{-\frac{ka}{\sigma}} \Phi\left(k - \frac{a}{\sigma}\right)}{e^{\frac{ka}{\sigma}} \Phi\left(k + \frac{a}{\sigma}\right) + e^{-\frac{ka}{\sigma}} \Phi\left(k - \frac{a}{\sigma}\right)}.$$

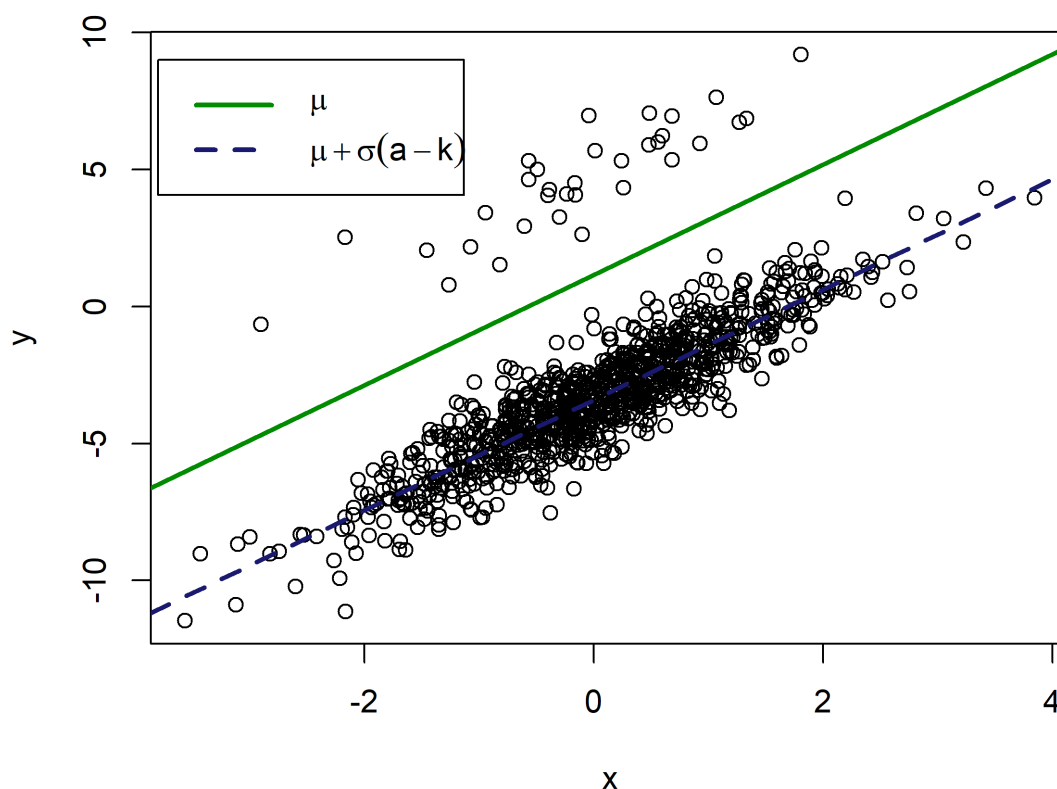
در این حالت عرض از مبدأ برابر $\beta_0 + E(\epsilon_i)$ خواهد بود.

رگرسیون مد

اگر داده‌های مورد مطالعه شامل داده‌های پرت یک‌طرفه باشند، رگرسیون مد کارایی بهتری نسبت به رگرسیون میانگین دارد. دوباره مدل رگرسیون (۱.۴) را در نظر بگیرید. با فرض BUN بودن توزیع خطا، برآوردهای پارامترهای توزیع در ناحیه دومی توزیع قرار گرفته باشند، رگرسیون مد به صورت زیر کارایی بهتری خواهد داشت:

$$\text{Mode}(Y_i) = \beta' x_i + \text{Mode}(\epsilon_i)$$

که در این حالت عرض از مبدأ به صورت $\beta_0 + \text{Mode}(\epsilon_i)$ خواهد بود. مقادیر $\text{Mode}(\epsilon_i)$ در رابطه (۱۰.۳) فصل چهارم داده شده‌اند. به‌عنوان مثالی از حالت رگرسیون مد، ϵ_i را از توزیع $\text{BUN}(0, 1, 4, 0/4)$ و x از توزیع نرمال استاندارد تولید کردیم و پارامترهای مدل $y_i = 1 + 2x_i + \epsilon_i$ را به صورت $\hat{\beta}_0 = 1/0375$ ، $\hat{\beta}_1 = 1/9198$ ، $\hat{\sigma} = 1/0432$ ، $\hat{k} = 3/7654$ و $\hat{a} = 0/4616$ برآورد کردیم. شکل ۲.۴ خط رگرسیون $\hat{\beta}_0 + \hat{\beta}_1 x$ را با خط‌چین نشان می‌دهد. خط رگرسیون مد که برابر $(\hat{\beta}_0 + \hat{a} + \hat{\sigma}\hat{k}) + \hat{\beta}_1 x$ است با خط ممتد رسم شده است که نشان می‌دهد برای داده‌های شامل داده پرت یک طرفه، رگرسیون مد عملکرد خوبی دارد.



شکل ۲.۴: پراکنش و خط‌های رگرسیونی مثالی از رگرسیون مد

برآوردگر M

می‌دانیم که تابع چگالی احتمال BUN برای $a = 0$ متقارن است. همچنین هسته BUN ترکیب خطی از هسته نرمال و لاپلاس به ازای $k < 0$ است. از توزیع نرمال برای داده‌های متقارن و دم‌سبک و توزیع لاپلاس برای داده‌های دم‌سنگین یا داده‌های شامل داده پرت استفاده می‌شود. توزیع BUN که ترکیبی از این دو توزیع است، بدین صورت استفاده می‌شود که در مواجهه با داده‌های دم‌سبک وزن بیشتری به هسته نرمال و در مواقعی که داده پرت داشته

باشیم، وزن بیشتری به هسته لاپلاس می‌دهد. با کمی تغییر در ساختار خطی گفته‌شده، می‌توان تابع زیان جدیدی به صورت زیر ساخت:

$$\rho(\epsilon) = \begin{cases} -k|\epsilon| + \frac{1}{4}\epsilon^2 & |\epsilon| \leq c \\ (c-k)|\epsilon| - \frac{c^2}{4} & |\epsilon| > c \end{cases} \quad (2.4)$$

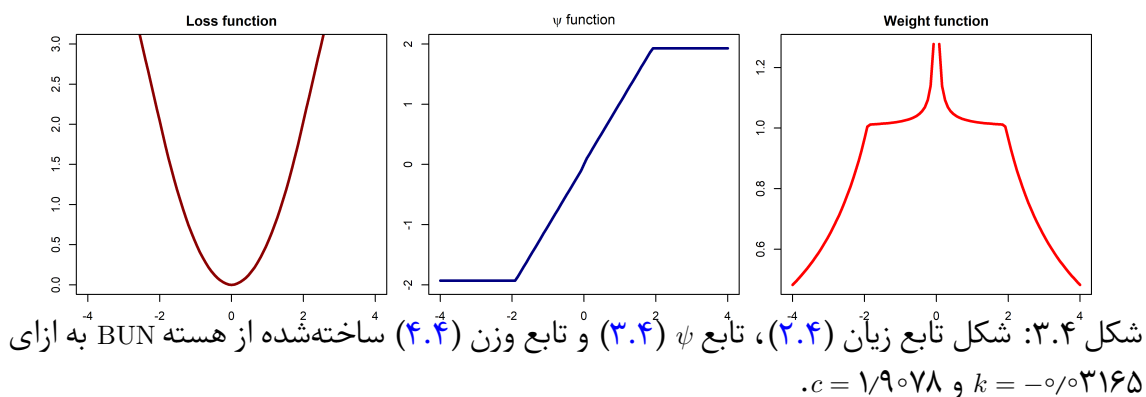
مشتق $\rho(\epsilon)$ نسبت به ϵ برابر با

$$\psi(\epsilon) = \begin{cases} \epsilon - k\text{sign}(\epsilon) & |\epsilon| \leq c \\ (c-k)\text{sign}(\epsilon) & |\epsilon| > c \end{cases} \quad (3.4)$$

و تابع وزن برابر با

$$w(\epsilon) = \begin{cases} 1 - \frac{k}{|\epsilon|} & |\epsilon| \leq c; \\ \frac{c-k}{|\epsilon|} & |\epsilon| > c. \end{cases} \quad (4.4)$$

است. شکل ۳.۴ تابع زیان جدید، ψ و تابع وزن ساخته‌شده از هسته BUN را به ازای $k = -0.3165$ و $c = 1/9078$ نمایش می‌دهد. تابع زیان ساخته‌شده از BUN شرایط ذکرشده



در بخش ۱.۳.۱ را دارد. بنابراین برآوردهای حاصل توسط آن دارای توزیع مجانبی نرمال با میانگین β و واریانس $\frac{\nu}{n}$ هستند به طوری که

$$\nu = \frac{E_F(\psi(\epsilon)^2)}{E_F(\psi'(\epsilon))^2}$$

و

$$\nu = \frac{(\Phi(c) - \Phi(-c))^2}{2(c-k)^2\Phi(-c) + 2(\Phi(c) - c\phi(c)) - 1 + k^2(\Phi(c) - \Phi(-c)) + 4k(\phi(c) - \phi(-c))}.$$

مقادیر k و c طوری انتخاب می‌شوند که کارایی مجانبی ۹۵٪ نرمال داشته باشیم. برای $k = -0.222$ و $c = 1/9078$ کارایی مجانبی تقریباً برابر ۹۵٪ است.

۲.۱.۴ مطالعه شبیه‌سازی

در این مطالعه، چهار نوع سناریو برای تولید داده‌ها به کار بردیم تا عملکرد مدل‌های معرفی شده را ارزیابی کنیم. مدل رگرسیون خطی

$$y = 1 + 2x + \epsilon$$

را در نظر بگیرید که متغیر تبیینی x از توزیع نرمال استاندارد تولید شده است. سناریوهای به کار برده شده عبارتند از

۱. تولید ϵ از نرمال استاندارد و جایگزینی 1° درصد از y_i با $y_i + 1^\circ$ برای مقادیر بالای x .

۲. تولید جمله خطای ϵ از توزیع آمیخته دو نرمال به صورت

$$0.9N(0, 1) + 0.1N(-10, 1).$$

۳. تولید جمله خطای ϵ از توزیع نرمال چوله با پارامترهای مکان صفر و مقیاس ۱ و چولگی ۲.

۴. تولید جمله خطای ϵ از توزیع نرمال چوله با پارامتر مکان صفر، مقیاس ۱ و چولگی ۴.

اندازه نمونه‌های تولید شده برابر 5° ، 10° ، 20° و 40° انتخاب شدند. برای هر سناریو و حجم نمونه 500 تکرار تولید شده است.

ضرایب رگرسیونی را با فرض BUN، نرمال، St و SN بودن توزیع خطاها برآورد کردیم. همچنین برای ارزیابی برآوردگرهای M جدید BUN، ضرایب را با برآوردگرهای M همپل و تابع زیان جدید BUN برآورد کردیم. نتایج در جداول ۱.۴ تا ۳.۴ گزارش شده‌اند. معیارهایی که در جداول گزارش شده‌اند، عبارتند از

$$\begin{aligned} \text{bias}(\hat{\beta}_j) &= \frac{1}{500} \sum_{i=1}^{500} (\hat{\beta}_{ij} - \beta_j), j = 0, 1 \\ \text{MSE}(\hat{\beta}_j) &= \frac{1}{500} \sum_{i=1}^{500} (\hat{\beta}_{ij} - \beta_j)^2, j = 0, 1 \\ \text{MSE} &= \frac{1}{500} \sum_{i=1}^{500} \|\hat{\beta}_i - \beta\|^2 \end{aligned}$$

که در آنها $\beta_0 = 1$ ، $\beta_1 = 2$ ، $\beta = (\beta_0, \beta_1)$ ، $\hat{\beta}_{ij}$ برآورد β_j در تکرار i -ام برای $j = 0, 1$ و $\hat{\beta}_i$ برآورد β در تکرار i -ام است. با توجه به نتایج ارائه شده، می‌توان نتیجه گرفت که مدل‌های پیشنهادی عملکرد خوبی نسبت به مدل‌های رقیب دارند. در سناریوی اول که تولید خطاها شامل داده پرت یکطرفه است مدل رگرسیونی مد و میانگین عملکرد معادلی دارند. توزیع St به علت داشتن دم‌های سنگین، در بین مدل‌های رقیب، برآزش بهتری داشته است اما برتری

رگرسیون مد و میانگین بر پایه توزیع BUN کاملاً مشهود است. در سناریوی دوم نیز خطاها شامل داده پرت یک‌طرفه است. در این سناریو نیز رگرسیون مد عملکرد بهتری نسبت به سایر مدل‌ها دارد. در سناریوی سوم خطاها از توزیع چوله نرمال با پارامتر چولگی کوچک تولید شده است، یعنی خطاها به توزیع نرمال نزدیک‌تر هستند، که مشاهده می‌کنیم همه مدل‌ها عملکرد یکسانی دارند با این حال رگرسیون مد برای حجم نمونه‌های بالا عملکرد خوبی نسبت به سایر مدل‌ها دارد. در سناریوی چهارم نیز توزیع چوله نرمال با پارامتر چولگی بزرگتر برای خطا در نظر گرفته شده است، رگرسیون مد عملکرد بهتری نسبت به سایر مدل‌ها دارد. توجه داشته باشید که در سناریوهای سوم و چهارم توزیع واقعی خطا توزیع SN است اما با این وجود رگرسیون میانگین توزیع BUN عملکردی معادل و رگرسیون مد BUN عملکرد بهتری نسبت به رگرسیون توزیع SN دارند.

جدول ۱۰۴: مقادیر MSE کل بردار ضرایب در مثال شبیه‌سازی شده به ازای روش‌های برآورد مختلف

سناریو	n	BUN _{Modal}	BUN _{Mean}	نرمال	St	SN	M – BUN	M – Hampel
اول	۵۰	۱/۰۴	۱/۰۳	۶/۰۹	۱/۱۱	۲/۶۲	۲/۵۵	۱/۷۳
	۱۰۰	۱/۰۱	۱/۰۱	۵/۹۶	۱/۰۶	۲/۵۱	۲/۵۴	۱/۶۰
	۲۰۰	۱/۰۲	۱/۰۱	۵/۹۱	۱/۰۶	۲/۵۰	۲/۴۴	۱/۲۶
	۴۰۰	۱/۰۰	۱/۰۰	۵/۹۰	۱/۰۶	۲/۵۲	۲/۴۱	۱/۰۶
دوم	۵۰	۱/۰۶	۳/۲۸	۳/۵۳	۱/۱۳	۵/۲۶	۱/۵۴	۱/۱۸
	۱۰۰	۱/۰۱	۳/۱۱	۳/۲۳	۱/۰۵	۴/۹۴	۱/۳۹	۱/۰۴
	۲۰۰	۱/۰۲	۳/۰۶	۳/۱۲	۱/۰۵	۴/۸۸	۱/۳۵	۱/۰۳
	۴۰۰	۱/۰۱	۳/۰۳	۳/۰۶	۱/۰۵	۴/۸۲	۱/۳۴	۱/۰۲
سوم	۵۰	۰/۸۴	۰/۸۲	۰/۸۲	۰/۸۲	۰/۸۲	۰/۸۱	۰/۸۲
	۱۰۰	۰/۸۱	۰/۸۱	۰/۸۱	۰/۸۱	۰/۸۱	۰/۸۱	۰/۸۱
	۲۰۰	۰/۷۹	۰/۸۱	۰/۸۱	۰/۸۰	۰/۸۱	۰/۸۰	۰/۸۰
	۴۰۰	۰/۷۷	۰/۸۰	۰/۸۰	۰/۷۹	۰/۸۰	۰/۷۹	۰/۷۹
چهارم	۵۰	۰/۸۱	۰/۸۵	۰/۸۵	۰/۸۴	۰/۸۵	۰/۸۴	۰/۸۴
	۱۰۰	۰/۷۹	۰/۸۳	۰/۸۳	۰/۸۱	۰/۸۳	۰/۸۲	۰/۸۲
	۲۰۰	۰/۸۰	۰/۸۳	۰/۸۳	۰/۸۲	۰/۸۳	۰/۸۲	۰/۸۲
	۴۰۰	۰/۸۰	۰/۸۲	۰/۸۲	۰/۸۱	۰/۸۲	۰/۸۱	۰/۸۱

خانواده‌ای از مدل‌های رگرسیون خطی برای مانده‌های شامل داده پرت یک طرفه یا چوله ۱۰۱

جدول ۲.۴: مقادیر MSE فردی پارامترها در مثال شبیه‌سازی شده به ازای روش‌های برآورد مختلف

سناریو	پارامتر	n	BUN_{Modal}	BUN_{Mean}	نرمال	St	SN	M – BUN	M – Hampel
اول	$\hat{\beta}_0$	۵۰	-۰/۰۰۶۳	۰/۹۹۸۴	۱/۰۱۵۶	۰/۰۳۹۷	۱/۴۰۳۲	۰/۴۸۵۳	۰/۱۷۱۳
	$\hat{\beta}_1$		-۰/۰۱۴۵	-۰/۰۱۴۵	۱/۷۷۰۲	۰/۰۶۴۹	۰/۵۶۱۷	۰/۸۷۰۲	۰/۳۱۷۸
	$\hat{\beta}_0$	۱۰۰	۱e – ۰۴	۰/۹۹۸۴	۱/۰۰۱۳	۰/۰۵۱	۱/۳۵۵۳	۰/۴۹۳	۰/۱۳۵۸
	$\hat{\beta}_1$		۲e – ۰۴	۲e – ۰۴	۱/۷۶۵	۰/۰۸۸۷	۰/۵۹۸	۰/۸۹۰۴	۰/۲۵۱۹
	$\hat{\beta}_0$	۲۰۰	-۰/۰۰۵۳	۰/۹۹۶۸	۱/۰۰۲۹	۰/۰۴	۱/۳۵۸۴	۰/۴۸۱۵	۰/۰۵۷۷
	$\hat{\beta}_1$		۰/۰۰۱۵	۰/۰۰۱۵	۱/۷۶۲۲	۰/۰۸۱۸	۰/۶۰۶۵	۰/۸۷۱۸	۰/۱۱۸۹
	$\hat{\beta}_0$	۴۰۰	-۰/۰۰۳۷	۰/۹۹۷	۰/۹۹۲۴	۰/۰۴۳۶	۱/۳۴۴۴	۰/۴۷۶۹	۰/۰۱۹۹
دوم	$\hat{\beta}_1$		۰/۰۰۲۸	۰/۰۰۲۸	۱/۷۶۶۲	۰/۰۸۸۱	۰/۶۳۵۷	۰/۸۷۳۳	۰/۰۴۵۲
	$\hat{\beta}_0$	۵۰	۰/۰۰۴۷	-۰/۹۹۳۴	-۰/۹۹۹۸	-۰/۰۲۷۵	-۱/۵۰۱۴	-۰/۲۸۵۷	-۰/۰۴۶۳
	$\hat{\beta}_1$		۰/۰۰۱۷	۰/۰۰۱۷	۰/۰۲۹۴	۴e – ۰۴	-۰/۰۰۶۲	۰/۰۰۷۸	-۰/۰۰۴۱
	$\hat{\beta}_0$	۱۰۰	-۰/۰۰۰۳	-۰/۹۹۰۳	-۰/۹۸۷۸	-۰/۰۲۸۱	-۱/۴۹۳۲	-۰/۲۶۰۷	-۰/۰۱۶۳
	$\hat{\beta}_1$		-۰/۰۰۲۷	-۰/۰۰۲۷	۰/۰۱۲۸	-۰/۰۰۷۳	۰/۰۱	۰/۰۰۳۶	-۰/۰۰۱۶
	$\hat{\beta}_0$	۲۰۰	-۰/۰۰۱۸	-۰/۹۹۴۴	-۰/۹۹۵۳	-۰/۰۳۲۱	-۱/۴۹۶۳	-۰/۲۵۸۶	-۰/۰۱۱۹
	$\hat{\beta}_1$		۱e – ۰۴	۱e – ۰۴	۰/۰۰۷۹	۲e – ۰۴	-۰/۰۰۳۶	۶e – ۰۴	-۰/۰۰۱۵
سوم	$\hat{\beta}_0$	۴۰۰	۰/۰۰۱۸	-۰/۹۹۲۸	-۰/۹۹۳۲	-۰/۰۲۷۸	-۱/۵۰۳۶	-۰/۲۵۱۹	-۰/۰۰۶۱
	$\hat{\beta}_1$		۳e – ۰۴	۳e – ۰۴	-۰/۰۰۸۴	-۰/۰۰۰۳	-۰/۰۰۱۹	-۰/۰۰۳۴	-۹e – ۰۴
	$\hat{\beta}_0$	۵۰	۰/۵۲۰۷	۰/۷۱۵۳	۰/۷۱۸۲	۰/۷۰۴۱	۰/۷۱۴۹	۰/۷۰۴۹	۰/۷۰۷۱
	$\hat{\beta}_1$		۸e – ۰۴	۸e – ۰۴	۴e – ۰۴	۱e – ۰۴	۰/۰۰۱	۸e – ۰۴	۶e – ۰۴
	$\hat{\beta}_0$	۱۰۰	۰/۴۹۴۵	۰/۷۱۰۴	۰/۷۱۱۷	۰/۷	۰/۷۱۰۸	۰/۶۹۸۷	۰/۷۰۱۱
	$\hat{\beta}_1$		۰/۰۰۰۵	۰/۰۰۰۵	۰/۰۰۶۵	۰/۰۰۵۷	۰/۰۰۴۵	۰/۰۰۵۸	۰/۰۰۵۸
	$\hat{\beta}_0$	۲۰۰	۰/۴۴۰۴	۰/۷۱۱۴	۰/۷۱۲۱	۰/۷۰۰۳	۰/۷۱۱۷	۰/۶۹۷۹	۰/۷۰۰۵
چهارم	$\hat{\beta}_1$		-۰/۰۰۱۴	-۰/۰۰۱۴	-۷e – ۰۴	-۸e – ۰۴	-۸e – ۰۴	-۰/۰۰۱	-۰/۰۰۱
	$\hat{\beta}_0$	۴۰۰	۰/۴۱۵	۰/۷۱۴۴	۰/۷۱۴۷	۰/۷۰۳	۰/۷۱۴۶	۰/۷	۰/۷۰۲۷
	$\hat{\beta}_1$		-۰/۰۰۱۷	-۰/۰۰۱۷	-۰/۰۰۱۸	-۰/۰۰۱۸	-۶e – ۰۴	-۰/۰۰۱۷	-۰/۰۰۱۷
	$\hat{\beta}_0$	۵۰	۰/۴۵۴۲	۰/۷۷۵۱	۰/۷۸۱۴	۰/۷۵۱۹	۰/۷۸۰۳	۰/۷۶۲	۰/۷۶۵۷
	$\hat{\beta}_1$		۰/۰۰۲۳	۰/۰۰۲۳	۰/۰۰۱۳	۰/۰۰۳۵	۰/۰۰۱۶	۰/۰۰۲۲	۰/۰۰۲۱
	$\hat{\beta}_0$	۱۰۰	۰/۳۷۳۵	۰/۷۷۳	۰/۷۷۵۵	۰/۷۴۹	۰/۷۷۴۲	۰/۷۵۶۴	۰/۷۶۰۴
	$\hat{\beta}_1$		۰/۰۰۱۶	۰/۰۰۱۶	-۹e – ۰۴	-۰/۰۰۱۲	۹e – ۰۴	-۰/۰۰۱۲	-۰/۰۰۱۳
	$\hat{\beta}_0$	۲۰۰	۰/۳۲۹	۰/۷۷۴۳	۰/۷۷۵۵	۰/۷۴۷۹	۰/۷۷۴۶	۰/۷۵۵۸	۰/۷۵۹۸
	$\hat{\beta}_1$		-۰/۰۰۰۲	-۰/۰۰۰۲	۱e – ۰۴	-۲e – ۰۴	-۰/۰۰۱۶	۱e – ۰۴	۱e – ۰۴
	$\hat{\beta}_0$	۴۰۰	۰/۲۹۸۷	۰/۷۷۳۴	۰/۷۷۳۹	۰/۷۴۵۳	۰/۷۷۳۹	۰/۷۵۳۷	۰/۷۵۷۷
	$\hat{\beta}_1$		-۰/۰۰۲۲	-۰/۰۰۲۲	-۰/۰۰۰۲	-۰/۰۰۱۶	-۰/۰۰۱۲	-۰/۰۰۱۷	-۰/۰۰۱۷

۳.۱.۴ مثال‌های واقعی

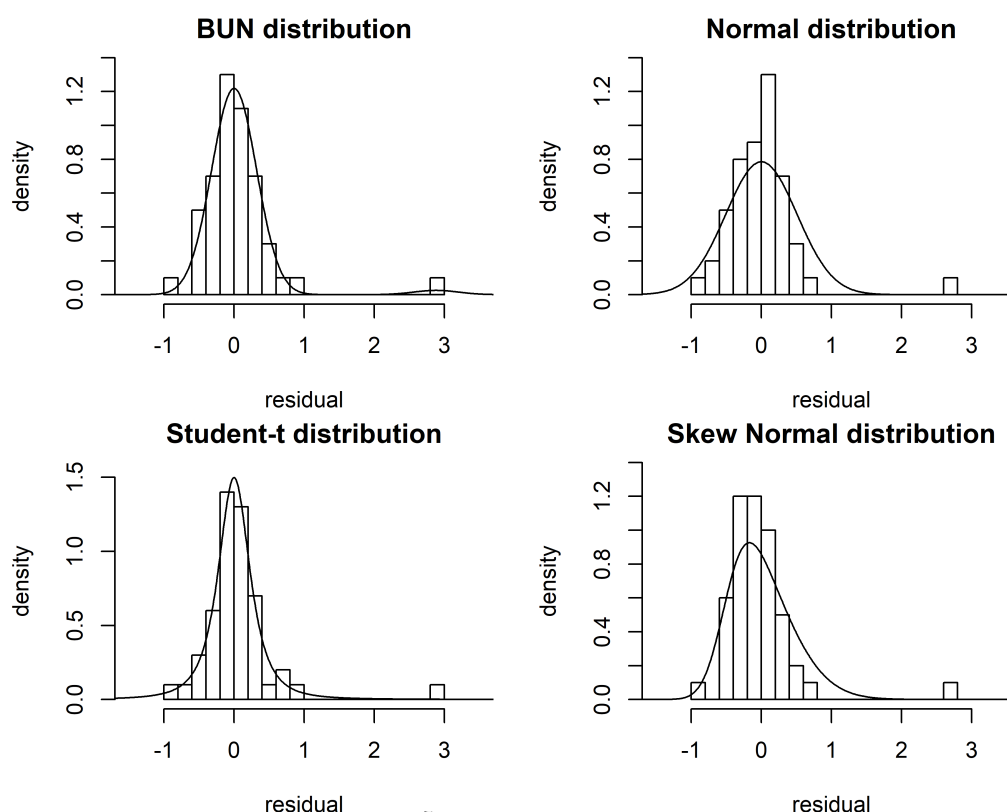
برای ارزیابی عملکرد مدل‌های رگرسیون پیشنهادی بر پایه توزیع BUN در برازش داده‌های واقعی، دو مجموعه داده واقعی را تحلیل کردیم. از سه رقیب پارامتری نرمال، SN و St و یک رقیب ناپارامتری برآوردگر M همپل برای مقایسه با مدل‌های پیشنهادی استفاده کردیم.

جدول ۳.۴: مقادیر اریبی فردی پارامترها در مثال شبیه‌سازی شده به ازای روش‌های برآورد مختلف

سناریو	پارامتر	n	BUN_{Modal}	BUN_{Mean}	نرمال	St	SN	M – BUN	M – Hampel
اول	$\hat{\beta}_0$	۵۰	۰/۰۲۲۴	۰/۰۱۵۹	۱/۱۱۹۶	۰/۰۳۵۸	۲/۰۱۸۸	۰/۲۹۵	۰/۱۵۵۸
	$\hat{\beta}_1$		۰/۰۳۳۶	۰/۰۳۳۶	۳/۲۱۳۶	۰/۰۵۷	۰/۴۵۲۴	۰/۸۸۹۸	۰/۴۹۰۵
	$\hat{\beta}_0$	۱۰۰	۰/۰۱۲۷	۰/۰۰۷۷	۱/۰۴۱۲	۰/۰۱۹۳	۱/۸۶	۰/۲۷۳۵	۰/۱۰۳۴
	$\hat{\beta}_1$		۰/۰۱۶	۰/۰۱۶	۳/۱۵۵۵	۰/۰۳۲۹	۰/۴۱۶۶	۰/۸۶۲۹	۰/۳۳۰۱
	$\hat{\beta}_0$	۲۰۰	۰/۰۰۶۵	۰/۰۹۹۹۱	۱/۰۲۵۶	۰/۰۱۰۵	۱/۸۵۵۸	۰/۲۴۶۴	۰/۰۴
	$\hat{\beta}_1$		۰/۰۰۷۶	۰/۰۰۷۶	۳/۱۲۲۸	۰/۰۱۷۴	۰/۳۹۲۹	۰/۷۹۲۲	۰/۱۳۱
	$\hat{\beta}_0$	۴۰۰	۰/۰۰۳۱	۰/۰۹۹۶۵	۰/۹۹۵۴	۰/۰۰۵۸	۱/۸۱۲۵	۰/۲۳۴۵	۰/۰۱۱۲
	$\hat{\beta}_1$		۰/۰۰۳۵	۰/۰۰۳۵	۳/۱۲۸۸	۰/۰۱۲۷	۰/۴۱۵۴	۰/۷۷۷۳	۰/۰۲۹۸
دوم	$\hat{\beta}_0$	۵۰	۰/۰۲۵۴	۱/۱۸۰۷	۱/۱۹۸۶	۰/۰۳۲	۲/۵۷۱۵	۰/۱۵۷۲	۰/۰۷۹۴
	$\hat{\beta}_1$		۰/۰۱۹۷	۰/۰۱۹۷	۰/۲۲۳۶	۰/۰۳۱۲	۰/۰۹۰۲	۰/۰۵۰۲	۰/۰۳۲۳
	$\hat{\beta}_0$	۱۰۰	۰/۰۱۱۷	۱/۰۷۱۴	۱/۰۶۶۷	۰/۰۱۷۹	۲/۳۶۹۶	۰/۰۹۱۳	۰/۰۱۳۳
	$\hat{\beta}_1$		۰/۰۱۱۱	۰/۰۱۱۱	۰/۱۰۳۵	۰/۰۱۶۵	۰/۰۳۶۸	۰/۰۲۰۵	۰/۰۱۲۵
	$\hat{\beta}_0$	۲۰۰	۰/۰۰۵۶	۱/۰۴۴۱	۱/۰۴۶۳	۰/۰۰۸۸	۲/۳۱۸۹	۰/۰۷۸۷	۰/۰۰۶۲
	$\hat{\beta}_1$		۰/۰۰۴۷	۰/۰۰۴۷	۰/۰۴۸۳	۰/۰۰۷۴	۰/۰۱۴۴	۰/۰۰۸۹	۰/۰۰۵۳
	$\hat{\beta}_0$	۴۰۰	۰/۰۰۲۶	۱/۰۰۷	۱/۰۰۷۹	۰/۰۰۴۵	۲/۲۹۱۷	۰/۰۶۸۳	۰/۰۰۲۸
	$\hat{\beta}_1$		۰/۰۰۲۴	۰/۰۰۲۴	۰/۰۲۵۴	۰/۰۰۳۷	۰/۰۰۷۲	۰/۰۰۴۳	۰/۰۰۲۷
سوم	$\hat{\beta}_0$	۵۰	۰/۳۲۸۸	۰/۵۲۲۴	۰/۵۲۶۴	۰/۵۰۷۴	۰/۵۲۱۵	۰/۵۰۷۸	۰/۵۱۰۹
	$\hat{\beta}_1$		۰/۰۱۲	۰/۰۱۲	۰/۰۱۰۷	۰/۰۱۰۸	۰/۰۱۰۵	۰/۰۱۰۶	۰/۰۱۰۶
	$\hat{\beta}_0$	۱۰۰	۰/۲۸۷۱	۰/۵۱	۰/۵۱۱۷	۰/۴۹۵۶	۰/۵۱۰۵	۰/۴۹۳۵	۰/۴۹۶۸
	$\hat{\beta}_1$		۰/۰۰۵۸	۰/۰۰۵۸	۰/۰۰۵۵	۰/۰۰۵۵	۰/۰۰۵	۰/۰۰۵۴	۰/۰۰۵۴
	$\hat{\beta}_0$	۲۰۰	۰/۲۱۸۵	۰/۵۰۸۵	۰/۵۰۹۶	۰/۴۹۳۱	۰/۵۰۸۹	۰/۴۸۹۵	۰/۴۹۳۲
	$\hat{\beta}_1$		۰/۰۰۲۸	۰/۰۰۲۸	۰/۰۰۲۶	۰/۰۰۲۶	۰/۰۰۲۴	۰/۰۰۲۶	۰/۰۰۲۶
	$\hat{\beta}_0$	۴۰۰	۰/۱۸۹۷	۰/۵۱۱۶	۰/۵۱۲	۰/۴۹۵۵	۰/۵۱۱۹	۰/۴۹۱۳	۰/۴۹۵
	$\hat{\beta}_1$		۰/۰۰۱۲	۰/۰۰۱۲	۰/۰۰۱۲	۰/۰۰۱۲	۰/۰۰۱	۰/۰۰۱۲	۰/۰۰۱۲
چهارم	$\hat{\beta}_0$	۵۰	۰/۲۵۳۲	۰/۶۱۰۷	۰/۶۱۸۳	۰/۵۷۵۳	۰/۶۱۶۸	۰/۵۸۸۸	۰/۵۹۴۵
	$\hat{\beta}_1$		۰/۰۰۷۶	۰/۰۰۷۶	۰/۰۰۷۸	۰/۰۰۷۹	۰/۰۰۷۲	۰/۰۰۷۶	۰/۰۰۷۶
	$\hat{\beta}_0$	۱۰۰	۰/۱۶۹۵	۰/۶۰۱۶	۰/۶۰۵۵	۰/۵۶۶۴	۰/۶۰۳۵	۰/۵۷۶۵	۰/۵۸۲۵
	$\hat{\beta}_1$		۰/۰۰۳۶	۰/۰۰۳۶	۰/۰۰۴۱	۰/۰۰۳۹	۰/۰۰۳	۰/۰۰۳۹	۰/۰۰۳۹
	$\hat{\beta}_0$	۲۰۰	۰/۱۲۹۴	۰/۶۰۱۴	۰/۶۰۳۲	۰/۵۶۲	۰/۶۰۱۸	۰/۵۷۳۱	۰/۵۷۹۳
	$\hat{\beta}_1$		۰/۰۰۱۸	۰/۰۰۱۸	۰/۰۰۲۱	۰/۰۰۲	۰/۰۰۱۴	۰/۰۰۲	۰/۰۰۲
	$\hat{\beta}_0$	۴۰۰	۰/۱۰۵۵	۰/۵۹۹۱	۰/۶	۰/۵۵۶۹	۰/۵۹۹۹	۰/۵۶۹۱	۰/۵۷۵۳
	$\hat{\beta}_1$		$8e-04$	$8e-04$	$9e-04$	$9e-04$	$7e-04$	$9e-04$	$9e-04$

معیارهای مقایسه AIC، RSE، RMSPE و ADE را نیز برای مقایسه مدل‌ها به کار بردیم.

داده‌های صوت‌شناسی را در بخش ۱.۲.۲ بررسی کردیم. جزئیات داده‌ها و متغیر پاسخ و متغیرهای تبیینی که انتخاب شده‌اند، در بخش ۱.۲.۲ شرح داده شده است. نتایج برازش مدل‌ها از جمله مقادیر پارامترهای رگرسیونی و توزیعی برآوردشده به همراه se و معیارهای مقایسه AIC ، RSE ، $RMSPE$ و ADE در جدول ۴.۴ گزارش شده‌اند. با توجه به جدول ۴.۴ از نظر معیار انتخاب مدل AIC توزیع BUN بهترین برازش را دارد. کوچک بودن مقادیر معیارهای RSE ، $RMSPE$ و ADE نشان‌دهنده برتری مدل است. بنابراین از نظر این سه معیار نیز رگرسیون مد BUN به‌عنوان بهترین مدل انتخاب می‌شود. به دلیل این که برآوردهای توزیعی BUN در ناحیه دومی قرار گرفته‌اند، ما در جدول ۴.۴ رگرسیون مد را گزارش کردیم. با استفاده از بافت نگار مانده‌ها و منحنی‌های چگالی‌های برآوردشده (شکل ۴.۴) وجود داده پرت یک‌طرفه کاملاً مشهود است. همچنین با توجه به شکل ۴.۴ مشاهده می‌شود که تنها توزیع BUN همه مانده‌ها را پوشش داده است.



شکل ۴.۴: بافت نگار مانده‌ها و منحنی‌های تابع چگالی برآوردشده برای داده‌های صوت‌شناسی در مدل‌های رگرسیونی مختلف پیشنهادی

جدول ۴.۴: نتایج برآورد ML مدل‌های رگرسیون با جمله خطای BUN برای داده‌های صوت شناسی

M – Hampel	M – BUN	SN	St	نرمال	BUN _{Modal}	
پارامترهای رگرسیونی						
-۰/۶۹۵۲	-۰/۵۵۵۸	-۱/۳۰۹۱	-۰/۶۸۹۷	-۰/۰۰۶۹	۰/۸۷۷۷	$\hat{\beta}_0$
۰/۴۵۸۷	۰/۴۸۱۷	۰/۶۱۷۴	۰/۴۲۸۳	۰/۶۹۰۱	۰/۴۴۳۷	se
-۰/۶۹۵۲	-۰/۵۵۵۸	-۰/۷۹۷۵	-۰/۶۸۹۷	-۰/۰۰۶۹	-۰/۷۰۱۲	ضریب ثابت
۰/۴۵۸۷	۰/۴۸۱۷		۰/۴۲۸۳	۰/۶۹۰۱		se
۰/۰۴۸۷	۰/۰۴۸۵	۰/۰۴۷۲	۰/۰۵۵۰	۰/۰۴۷۴	۰/۰۴۸۷	loudness
۰/۰۰۵۶	۰/۰۰۵۸	۰/۰۰۶۵	۰/۰۰۶۲	۰/۰۰۸۴	۰/۰۰۵۳	se
۰/۶۱۱۵	۰/۵۹۹۹	۰/۷۰۷۰	۰/۴۳۱۰	۰/۵۵۸۵	۰/۶۱۳۴	sharpness
۰/۲۰۱۲	۰/۲۱۱۲	۰/۲۴۴۱	۰/۱۹۹۷	۰/۳۰۲۶	۰/۱۹۱۰	se
۱/۷۲۷۳	۱/۶۲۷۷	۱/۷۰۴۰	۱/۹۳۱۲	۱/۲۱۹۰	۱/۷۲۸۹	impulsiveness
۰/۳۱۱۲	۰/۳۲۶۸	۰/۴۰۰۵	۰/۲۹۵۷	۰/۴۶۸۲	۰/۳۰۰۶	se
پارامترهای توزیعی						
		۰/۶۸۶۹	۰/۲۴۰۳	۰/۵۰۸۲	۰/۳۲۰۶	مقیاس
		۰/۰۸۳۱	۰/۰۴۸۱	۰/۰۵۰۸	۰/۰۳۱۸	se
		۲/۶۰۲۹			۴/۴۹۳۶	شکل ۱
		۰/۸۸۳۲			۰/۶۷۷۸	
					-۰/۱۳۸۳	شکل ۲
					۰/۰۴۷۶	se
			۲/۳۶۶۰			درجه آزادی
			۰/۹۹۲۷			se
معیارهای مقایسه						
		۷۲/۶۹۵۲	۵۶/۷۰۳۹	۸۴/۱۹۹	۵۲/۴۳۹۵	AIC
۰/۳۰۲۲	۰/۳۰۰۲	۰/۳۲۱۷	۰/۲۹۳۱	۰/۳۴۳۳	۰/۲۹۲۰	RSE
۰/۵۳۶۸	۰/۵۳۹۰	۰/۵۵۲۲	۰/۵۷۱۴	۰/۵۴۶۵	۰/۵۳۴۰	RMSPE
۰/۳۳۰۱	۰/۳۳۳۵	۰/۳۴۱۷	۰/۳۷۸۹	۰/۳۴۳۲	۰/۳۲۶۹	ADE

داده‌های فشار خون سیستولیک

داده‌های فشار خون سیستولیک^{۱۰} (SBP) در یک مطالعه موردی در نشست سالانه انجمن آماری کانادا در سال ۲۰۰۵ بررسی شده است. داده‌ها شامل ۵۰۰ مشاهده از ۵۰۱ متغیر است. متغیر پاسخ SBP و ۵۰۰ متغیر تبیینی دیگر شامل ۱۷ متغیر بالینی و ۴۸۳ مشخصه ژنتیکی هستند. هدف این مطالعه موردی مشخص کردن متغیرهای رگرسیونی از بین ۵۰۰ متغیر تبیینی است. داده‌ها در آدرس <https://ssc.ca/en/case-study/blood-pressure> قابل دسترسی هستند. ما در این بخش، قسمت کوچکی از این داده‌ها را که شامل متغیر پاسخ SBP، وضعیت سیگاری بودن (غیرسیگاری = N ، سیگاری = Y) و وضعیت اضافه وزن (چاق

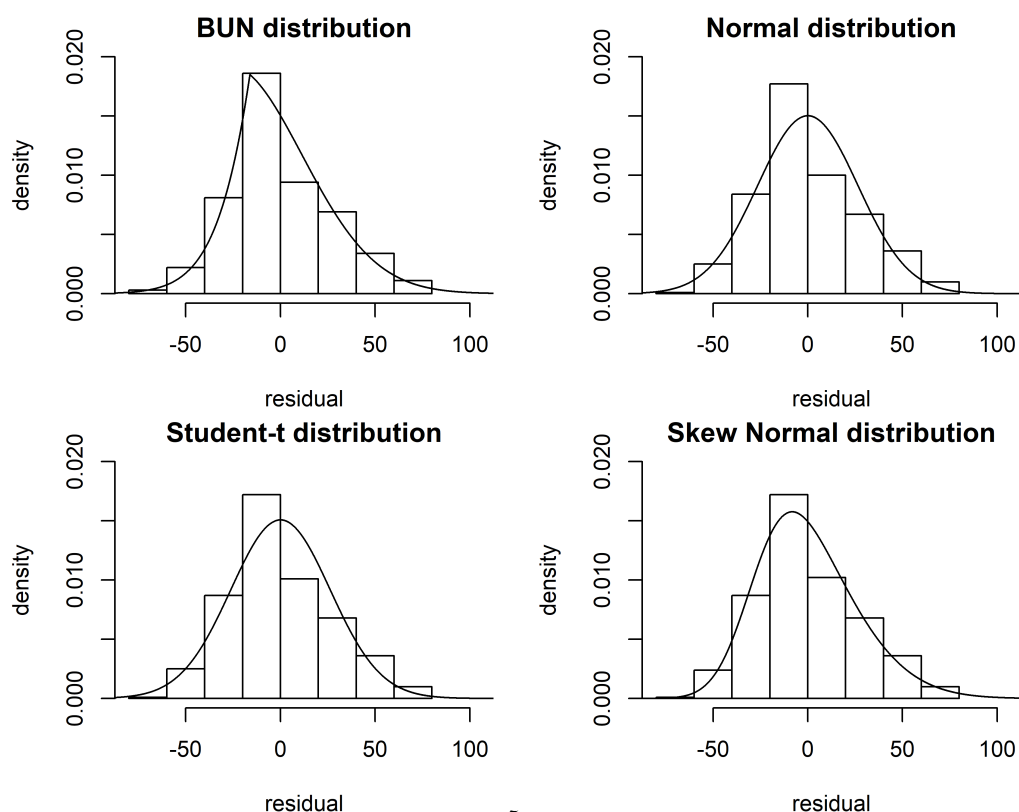
¹⁰Systolic Blood Pressure

خانواده‌ای از مدل‌های رگرسیون خطی برای مانده‌های شامل داده پرت یک طرفه یا چوله ۱۰۵

= ۳، اضافه وزن = ۲ و نرمال = ۱) هستند را مورد بررسی قرار دادیم.

این قسمت از داده‌ها در پکیج *Stat2Data* در نرم‌افزار R (کانون^{۱۱} و همکاران (۲۰۱۹)) قابل دسترس هستند. به نظر می‌رسد که در این داده‌ها خطای مدل چولگی دارد. بنابراین از رگرسیون میانگین BUN بهره بردیم. متغیرهای دودویی (وضعیت سیگاری) و طبقه‌ای (وضعیت اضافه وزن) را با تعریف متغیرهای ساختگی به مدل اضافه کردیم. غیرسیگاری و وزن نرمال را به عنوان سطوح پایه برای متغیر وضعیت سیگاری و اضافه وزن انتخاب کردیم. برآوردهای ضرایب و پارامترهای توزیع خطا به همراه se آنها در جدول ۵.۴ برای مدل‌های مختلف گزارش شده‌اند.

معیارهای مقایسه AIC، RMAPE، RSE و ADE نیز در جدول ۵.۴ گزارش شده‌اند. طبق نتایج جدول ۵.۴، بر اساس معیارهای AIC و RSE بهترین مدل رگرسیون میانگین BUN است. معیارهای ADE و RMSPE در این جدول برای مدل‌های مختلف نزدیک هم هستند. بافت نگار مانده‌ها و منحنی‌های چگالی‌های برآوردشده در شکل ۵.۴ رسم شده‌اند. با توجه به شکل ۵.۴ مشاهده می‌شود که مانده‌ها دارای چولگی هستند که به همین علت رگرسیون میانگین BUN گزارش شده است.



شکل ۵.۴: بافت نگار و منحنی‌های تابع چگالی برآورد ده بری داده‌های فشار خون سیستمیک بر حسب مدل‌های پیشنهادی مختلف

¹¹Cannon

جدول ۵.۴: نتایج برآورد ML مدل‌های رگرسیون با جمله خطای BUN برای داده‌های فشار خون سیستمیک

M – Hampel	M – BUN	SN	St	نرمال	BUN _{Mean}	
پارامترهای رگرسیونی						
۱۳۱/۸۶۰۹	۱۳۱/۹۰۸۷	۱۰۳/۰۸۰۱	۱۳۱/۷۴۶۴	۱۳۱/۸۹۷۷	۱۱۷/۷۲۶۰	$\hat{\beta}_0$
۲/۳۱۲۰	۲/۳۶۸۳	۳/۰۵۵۳	۲/۲۷۰۱	۲/۲۶۱۸	۲/۶۶۲۴	se
۱۳۱/۸۶۰۹	۱۳۱/۹۰۸۷	۱۳۱/۴۸۱۲	۱۳۱/۷۴۶۴	۱۳۱/۸۹۷۷	۱۳۳/۸۳۰۷	ضریب ثابت
۲/۳۱۲۰	۲/۳۶۸۳		۲/۲۷۰۱	۲/۲۶۱۸		se
۹/۱۴۳۴	۹/۰۴۷۹	۹/۰۴۲۳	۹/۵۹۴۰	۹/۱۷۹۸	۶/۲۷۴۰	smoker
۲/۴۵۹۵	۲/۵۱۹۴	۲/۲۵۴۸	۲/۴۱۲۳	۲/۴۰۶۱	۲/۶۶۲۸	se
۷/۹۹۲۸	۷/۷۱۵۰	۱۰/۰۳۵۸	۸/۸۹۳۱	۸/۲۵۹۲	۹/۰۳۱۸	overweight
۳/۲۷۳۲	۳/۳۵۲۹	۳/۰۴۴۹	۳/۲۱۲۳	۳/۲۰۲۱	۵/۷۸۴۳	se
۱۵/۲۵۰۵	۱۵/۰۴۸۹	۱۵/۷۶۸۸	۱۵/۱۵۵۷	۱۵/۶۱۰۲	۱۴/۲۷۴۰	obese
۲/۷۶۹۹	۲/۸۳۷۳	۲/۵۶۵۱	۲/۷۲۱۸	۲/۷۰۹۷	۲/۶۶۲۷	se
پارامترهای توزیعی						
		۳۸/۹۲۷۱	۲۶/۴۱۴۵	۲۶/۵۶۷۹	۳۹/۰۵۳۳	مقیاس
		۲/۰۲۷۸	۰/۸۹۲۷	۰/۸۴۰۰	۴/۸۸۰۷	se
		۲/۲۵۹۰			-۱/۲۹۰۲	شکل ۱
		۰/۳۶۲۸			۰/۴۰۲۳	se
					۳۸/۷۷۶۱	شکل ۲
					۹/۷۹۰۹	se
						درجه آزادی
			۱۱۷/۹۲۶۰			
			۱۶۶/۳۱۹۵			se
معیارهای مقایسه						
		۴۶۹۰/۵۸۷	۴۷۱۱/۳۱۷	۴۷۰۸/۷۴۱	۴۶۸۸/۰۸۱	AIC
۲۶/۶۹	۲۶/۶۸	۲۴/۸۷	۲۵/۲۸	۲۵/۵۳	۲۳/۷۴	RSE
۲۶/۷۸۸۰	۲۶/۷۸۵۳	۲۶/۷۶۰۷	۲۶/۸۰۵۵	۲۶/۷۷۹۱	۲۶/۷۷۶۶	RMSPE
۲۱/۲۹۳۷	۲۱/۲۵۷۵	۲۱/۳۶۳۴	۲۱/۳۲۵۶	۲۱/۳۳۴۹	۲۱/۲۷۴۳	ADE

۲.۴ رگرسیون فضایی برای داده‌های دومدی و چوله

تحلیل داده‌های فضایی دومدی که در موقعیت‌های کاربردی مختلف مشاهده می‌شوند، چندان مورد بررسی پژوهشگران قرار نگرفته است. در این بخش با معرفی یک توزیع چندمتغیره منعطف، مدل رگرسیونی فضایی برای تحلیل این نوع داده‌ها را ارائه می‌کنیم. برای برازش مدل، رهیافت درست‌نمایی را به کار خواهیم برد. با مثال‌های شبیه‌سازی و واقعی عملکرد مدل پیشنهادی را نسبت به رگرسیون فضایی معمول ارزیابی می‌کنیم.

۱.۲.۴ مقدمه

داده‌های فضایی به داده‌هایی گفته می‌شود که بر حسب موقعیت قرار گرفتن آنها، وابسته هستند. بنابراین در تحلیل رگرسیونی این نوع داده‌ها پذیره ناهمبسته بودن خطاها برقرار نیست و بین مشاهدات مجاور همبستگی قوی‌تری وجود دارد. در حالت کلی، برحسب آن که جمله خطا، متغیرهای تبیینی یا ضرایب رگرسیونی تابعی از موقعیت‌های فضایی باشند یا نه، انواع مختلفی از مدل‌های رگرسیون خطی برای داده‌های فضایی قابل تعریف هستند (محمدزاده ۱۳۹۴). در این بخش تمرکز ما بر روی مدلی است که در آن ساختار وابستگی جمله خطا تابعی از موقعیت‌های فضایی است. معمولاً در تحلیل رگرسیونی این نوع داده‌ها فرض بر این است که توزیع خطا نرمال چندمتغیره است. اما در موقعیت‌های مختلف با مواردی برخورد می‌کنیم که خطاها از توزیع نرمال تبعیت نمی‌کنند. به عنوان مثال در یک معدن طلا ممکن است دو ناحیه مجزا دارای رگه‌هایی با عیار طلای قابل ملاحظه و متفاوت باشند؛ در این حالت توزیع داده‌ها دومی خواهد بود. به عنوان مثالی دیگر، در داده‌های خاک‌شناسی، توانایی تبادل کاتیوم به عنوان تابعی از موقعیت مکانی ناحیه تحت مطالعه، با فراوانی بیشتری نزدیک به دو مقدار عددی مشاهده می‌شود.

برای تحلیل پاسخ‌های فضایی چوله، کارهای متعددی انجام شده‌اند. به عنوان نمونه می‌توان به کارهای ساهو^{۱۲} و همکاران (۲۰۰۳)، آرنولد وال^{۱۳} و همکاران (۲۰۰۵)، بولفرونی و لوکاس^{۱۴} (۲۰۰۷)، کریمی و همکاران (۲۰۱۰)، حسینی و همکاران (۲۰۱۱) و حسینی و محمدزاده (۲۰۱۲) که بر پایه رده توزیع‌های پارامتری مانند چوله نرمال شکل گرفته‌اند، اشاره کرد. اما برای پاسخ‌های دومی، تا حدی که نویسندگان مطلع هستند، تنها مدل‌هایی معرفی شده‌اند که بر اساس فرآیندهای آمیخته^{۱۵} پارامتری و ناپارامتری ساخته شده‌اند. به عنوان نمونه می‌توان فلاح و سدودی (۲۰۱۵) برای مدل آمیخته متناهی نرمال (پارامتری) و گلفاند^{۱۶} و همکاران (۲۰۰۵)، و گریفین^{۱۷} و استیل (۲۰۰۶) برای مدل فرآیند آمیخته دیریکله (ناپارامتری) را نام برد.

ما در این بخش، یک توزیع چندمتغیره منعطف معرفی می‌کنیم که ویژگی‌های مطلوبی دارد و از آن برای تحلیل پاسخ‌های فضایی غیرنرمال پیوسته شامل پاسخ‌های چوله و دومی استفاده می‌کنیم. این توزیع که آن را نرمال دومی-یک‌مدی چندمتغیره^{۱۸} MBUN می‌نامیم، می‌تواند ویژگی‌های متنوع موجود در داده‌ها شامل یک‌مدی یا دومی بودن، تقارن، چولگی یا ترکیب آنها را توسط پارامترهای خود کنترل و مدل‌بندی کند. همچنین تحت مقادیر مشخصی

¹²Sahu

¹³Arellano-Valle

¹⁴Lachos

¹⁵Mixture

¹⁶Gelfand

¹⁷Griffin

¹⁸Multivariate Bimodal-Unimodal Normal

برای پارامترهای توزیع MBUN، توزیع نرمال چندمتغیره نتیجه خواهد شد. بنابراین رده توزیع‌های نرمال چندمتغیره زیررده‌ای از رده توزیع‌های MBUN است.

۲.۲.۴ توزیع نرمال دومی-یک‌مدی چندمتغیره

توزیع نرمال چندمتغیره از پرکاربردترین توزیع‌های چندمتغیره است که در زمینه‌های مختلف از آن استفاده می‌شود. البته این توزیع برای داده‌هایی که چوله، دم‌سنگین یا چندمدی هستند، مناسب نیست. تلاش‌های مختلفی برای معرفی توزیع‌های چندمتغیره مناسب این نوع داده‌ها صورت گرفته‌اند. نمونه‌هایی از این توسعه‌های چندمتغیره در بخش ۴.۲ مطرح شدند. توزیع‌های معرفی شده در مدل‌بندی چولگی داده‌های چندمتغیره ممکن است موفق باشند، اما برای داده‌های دومی انتخاب‌های مناسبی نیستند. توزیع MBUN پیشنهادی ما می‌تواند ویژگی‌های متنوع موجود در داده‌ها شامل یک‌مدی یا دومی بودن، تقارن، و چولگی را مدل‌بندی کند. توزیع نرمال چندمتغیره نیز یکی از اعضای این خانواده محسوب می‌شود. توزیع MBUN برای بردار تصادفی p متغیره X دارای تابع چگالی به شکل

$$f_X(x) = \frac{e^{-\frac{k'x}{\Sigma}} e^{\left\{k' \left| \Sigma^{-\frac{1}{2}}(x-\mu) \right| \right\}} \phi_p(x, \mu + \Sigma^{\frac{1}{2}}a, \Sigma)}{\prod_{i=1}^p (e^{k_i a_i} \Phi(k_i + a_i) + e^{-k_i a_i} \Phi(k_i - a_i))} \quad (۵.۴)$$

است که در آن $\phi_p(\cdot; \mu, \Sigma)$ تابع چگالی نرمال چندمتغیره با بردار میانگین μ و ماتریس کوواریانس Σ و $\Phi(\cdot)$ تابع توزیع نرمال یک‌متغیره استاندارد است. پارامترهای این توزیع سه بردار p بعدی $\mu = (\mu_1, \dots, \mu_p)'$ و $k = (k_1, \dots, k_p)'$ و $a = (a_1, \dots, a_p)'$ با مولفه‌های حقیقی مقدار و ماتریس $p \times p$ معین مثبت Σ است. اگر بردار تصادفی X از این توزیع پیروی کند، برای نمایش آن از نماد $X \sim \text{MBUN}(\mu, \Sigma, k, a)$ استفاده می‌کنیم. همان‌طور که از شکل تابع چگالی (۵.۴) قابل استنباط است، پارامترهای μ_i ، k_i و a_i مختص متغیر i ام، $i = 1, \dots, p$ ، هستند. مقدار پارامتر k_i دومی یا یک‌مدی بودن و در صورت دومی بودن فاصله دو مد را مشخص می‌کند. پارامترهای μ_i و a_i نیز به ترتیب پارامترهای مکان و شکل متغیر متناظر هستند. برای تعیین وابستگی بین متغیرها نیز از ماتریس Σ استفاده می‌شود.

ویژگی اصلی داده‌های فضایی وجود ساختار وابستگی مکانی داده‌ها بر اساس موقعیت قرار گرفتن آنها است. برای پرهیز از پیچیدگی غیرقابل کنترل، در این بخش فرض می‌کنیم تعداد مدها، چولگی و شکل داده‌ها به موقعیت مکانی داده‌ها وابسته نیستند و ساختار وابستگی فضایی را از طریق ماتریس Σ وارد مدل می‌کنیم. البته این پذیره ساده‌سازی بیش از اندازه نیست، زیرا در عمل نیز ساختار وابستگی را معمولاً از طریق پارامترهای مقیاس و مکان وارد مدل می‌کنند. بنابراین به ازای $i = 1, \dots, p$ می‌پذیریم $k_i = k$ و $a_i = a$. در این صورت تابع چگالی (۵.۴) به شکل زیر تبدیل می‌شود:

$$f_X(x) = \frac{e^{-\frac{pk^2}{\Sigma}} e^{\left\{k' \left| \Sigma^{-\frac{1}{2}}(x-\mu) \right| \right\}} \phi_p(x, \mu + \Sigma^{\frac{1}{2}}a, \Sigma)}{(e^{ka} \Phi(k+a) + e^{-ka} \Phi(k-a))^p}$$

که در آن ۱ برداری از ۱ها با بعد p است.

برای سهولت در به دست آوردن برخی ویژگی‌های توزیع از نسخه استاندارد استفاده می‌کنیم که در آن $Z = \Sigma^{-\frac{1}{2}}(X - \mu)$. گشتاور اول و دوم متغیر Z به ترتیب

$$E[Z_j] = \frac{(a_j + k_j) e^{k_j a_j} \Phi(k_j + a_j) + (a_j - k_j) e^{-k_j a_j} \Phi(k_j - a_j)}{e^{k_j a_j} \Phi(k_j + a_j) + e^{-k_j a_j} \Phi(k_j - a_j)}$$

و

$$E[Z_j^2] = \frac{\left(1 + (a_j + k_j)^2\right) e^{k_j a_j} \Phi(k_j + a_j) + \left(1 + (a_j - k_j)^2\right) e^{-k_j a_j} \Phi(k_j - a_j)}{e^{k_j a_j} \Phi(k_j + a_j) + e^{-k_j a_j} \Phi(k_j - a_j)} + \frac{\frac{k_j}{\sqrt{\pi}} e^{-\frac{1}{2}(a_j^2 + k_j^2)}}{e^{k_j a_j} \Phi(k_j + a_j) + e^{-k_j a_j} \Phi(k_j - a_j)}$$

هستند. بردار میانگین و ماتریس کوواریانس MBUN نیز با توجه به $X = \Sigma^{\frac{1}{2}}Z + \mu$ عبارتند از

$$\begin{aligned} E[X] &= \Sigma^{\frac{1}{2}} E[Z] + \mu \\ Var[X] &= \Sigma^{\frac{1}{2}} Var[Z] \Sigma^{\frac{1}{2}}. \end{aligned}$$

با فرض $a_i = a$ و $k_i = k$ ، می‌توان نوشت

$$\begin{aligned} E[X] &= h_1(k, a) \Sigma^{\frac{1}{2}} \mathbf{1} + \mu \\ Var[X] &= h_2(k, a) \Sigma \end{aligned}$$

که در آن

$$h_1(k, a) = \frac{(a + k) e^{ka} \Phi(k + a) + (a - k) e^{-ka} \Phi(k - a)}{e^{ka} \Phi(k + a) + e^{-ka} \Phi(k - a)}$$

و

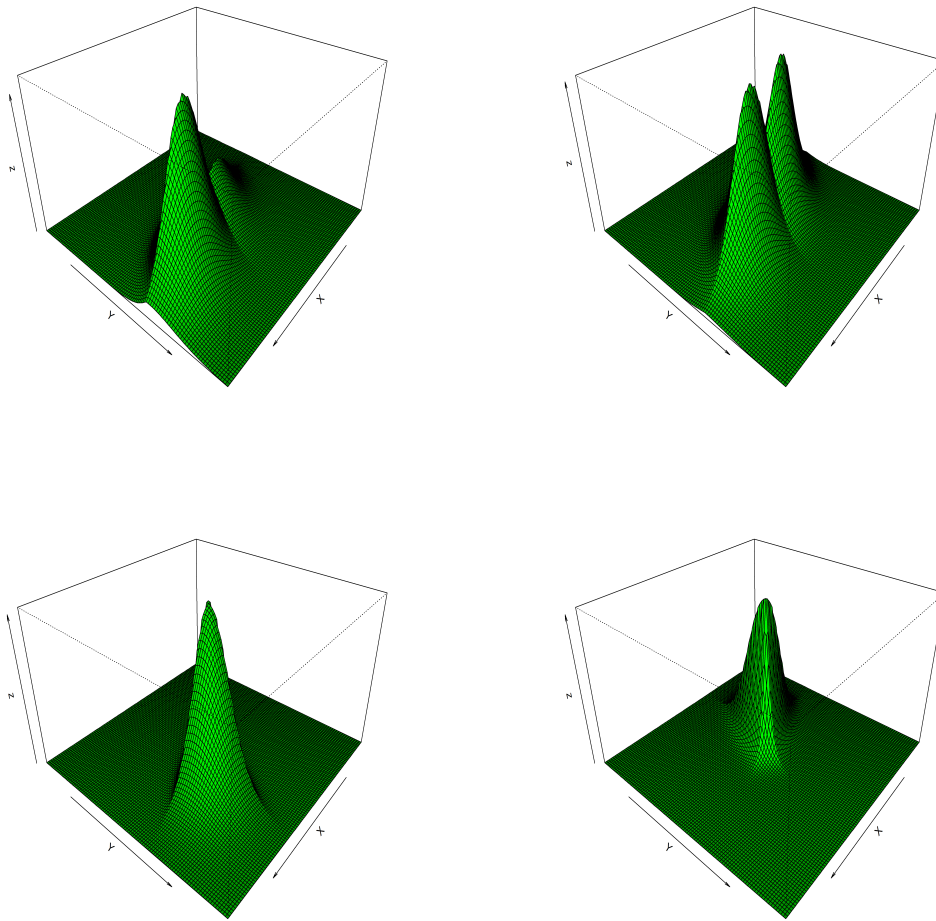
$$\begin{aligned} h_2(k, a) &= \frac{\left(1 + (a + k)^2\right) e^{ka} \Phi(k + a) + \left(1 + (a - k)^2\right) e^{-ka} \Phi(k - a)}{e^{ka} \Phi(k + a) + e^{-ka} \Phi(k - a)} \\ &\quad + \frac{\frac{k}{\sqrt{\pi}} e^{-\frac{1}{2}(a^2 + k^2)}}{e^{ka} \Phi(k + a) + e^{-ka} \Phi(k - a)} - h_1(k, a)^2. \end{aligned}$$

منحنی تابع چگالی (۵.۴) با دو متغیر را در شکل ۶.۴ برای مقادیر مختلف پارامترها می‌توانید مشاهده کنید.

۳.۲.۴ رگرسیون فضایی پاسخ‌های دومی

در مدل (۶.۱) فرض می‌کنیم

$$\delta \sim \text{MBUN}(\circ, \Sigma, k \setminus_n, a \setminus_n)$$



شکل ۶.۴: منحنی تابع چگالی توزیع MBUN برای $p = 2$ به‌ازای مقادیر مختلف پارامترهای توزیع. $a = (0, 0)'$ و $k = (1, -1)'$ (بالا سمت راست)، $a = (0, 0)'$ و $k = (1, -1)'$ (بالا سمت چپ)، $a = (-2, -2)'$ و $k = (-2, -2)'$ (پایین سمت راست)، $a = (2, 2)'$ و $k = (-2, -2)'$ (پایین سمت چپ) و در همه نمودارها پارامترهای $\mu = (0, 0)'$ و $\text{vech}(\Sigma) = (1, 0, 0, 1)'$ هستند.

که در آن $\delta = (\delta(s_1), \dots, \delta(s_n))'$ و اندیس n در 1_n برای تاکید بر تعداد موقعیت‌های فضایی مشاهده‌شده است.

ساختار وابستگی فضایی را از طریق ماتریس مقیاس معین مثبت Σ وارد مدل می‌کنیم. همچنین فرض می‌کنیم ساختار وابستگی فضایی، مانا^{۱۹} باشد. برای این کار، مشابه مدل رگرسیون با فرآیند خطای فضایی نرمال، از توابع پارامتری هم‌تغییرنگار

$$\sigma(h; \theta) = \text{Cov}_{\theta}[\delta(s), \delta(s+h)]$$

استفاده می‌کنیم. به‌عنوان مثالی از این نوع توابع، می‌توان دو مدل نمایی و گاوسی را نام برد

¹⁹Stationary

که به ترتیب به صورت زیر تعریف می‌شوند (کرسی^{۲۰}) (۱۹۹۳):

$$\sigma(h; \theta_1, \theta_2) = \theta_1 e^{-\|h\|/\theta_2}$$

و

$$\sigma(h; \theta_1, \theta_2) = \theta_1 e^{-\|h\|^2/\theta_2^2}$$

که در آنها پارامترهای وابستگی θ_1 و θ_2 به ترتیب ازاره^{۲۱} و دامنه هستند.

برازش مدل: رهیافت ماکسیمم درستنمایی

مدل خطی (۶.۱) را در نظر بگیرید. تابع لگاریتم درستنمایی مدل با فرآیند خطای فضایی $\delta \sim \text{MBUN}(\circ, \Sigma_h^\theta, k \mathbf{1}_n, a \mathbf{1}_n)$ به صورت

$$\begin{aligned} \ell(\zeta) = & -\frac{nk^2}{2} - n \log \left(e^{ka} \Phi(k+a) + e^{-ka} \Phi(k-a) \right) + k \mathbf{1}_n' \left| \Sigma_h^{\theta^{-1}} (Z - X\beta) \right| \\ & - \frac{1}{2} \log \left(\det \left(\Sigma_h^\theta \right) \right) - \frac{1}{2} \left(Z - X\beta - a \Sigma_h^{\theta^{-1}} \mathbf{1}_n \right)' \Sigma_h^{\theta^{-1}} \left(Z - X\beta - a \Sigma_h^{\theta^{-1}} \mathbf{1}_n \right) \\ & - \frac{n}{2} \log(\pi) \end{aligned}$$

است، که در آن $\zeta = (\beta', \theta', k, a)'$ بردار پارامترها و Σ_h^θ ماتریس مقیاس معتبر با بردار پارامتر θ است. بردار امتیاز این مدل

$$U_n = \left(\left(\frac{\partial \ell(\zeta)}{\partial \beta} \right)', \left(\frac{\partial \ell(\zeta)}{\partial \theta} \right)', \frac{\partial \ell(\zeta)}{\partial k}, \frac{\partial \ell(\zeta)}{\partial a} \right)'$$

است، به طوری که

$$\begin{aligned} \frac{\partial \ell(\zeta)}{\partial \beta} &= -k X' \Sigma_h^{\theta^{-1}} \text{sign} \left(\Sigma_h^{\theta^{-1}} (Z - X\beta) \right) + X' \Sigma_h^{\theta^{-1}} \left(Z - X\beta - a \Sigma_h^{\theta^{-1}} \mathbf{1}_n \right) \\ \frac{\partial \ell(\zeta)}{\partial \theta_j} &= - \left(k \mathbf{1}_n' \text{diag} \left(\text{sign} \left(\Sigma_h^{\theta^{-1}} (Z - X\beta) \right) \right) \Sigma_h^{\theta^{-1}} \frac{\partial \Sigma_h^{\theta^{-1}}}{\partial \theta_j} \Sigma_h^{\theta^{-1}} (Z - X\beta) \right. \\ &\quad \left. - \frac{1}{2} (Z - X\beta)' \Sigma_h^{\theta^{-1}} \frac{\partial \Sigma_h^{\theta^{-1}}}{\partial \theta_j} \Sigma_h^{\theta^{-1}} (Z - X\beta) + a (Z - X\beta)' \Sigma_h^{\theta^{-1}} \frac{\partial \Sigma_h^{\theta^{-1}}}{\partial \theta_j} \Sigma_h^{\theta^{-1}} \mathbf{1}_n \right. \\ &\quad \left. + \frac{1}{2} \text{tr} \left(\Sigma_h^{\theta^{-1}} \frac{\partial \Sigma_h^{\theta^{-1}}}{\partial \theta_j} \right) \right) \\ \frac{\partial \ell(\zeta)}{\partial k} &= -nk - na \frac{ae^{ka} \Phi(k+a) - ae^{-ka} \Phi(k-a) + 2ke^{ka} \phi(k+a)}{e^{ka} \Phi(k+a) + e^{-ka} \Phi(k-a)} \\ &\quad + \mathbf{1}_n' \left| \Sigma_h^{\theta^{-1}} (Z - X\beta) \right| \\ \frac{\partial \ell(\zeta)}{\partial a} &= -na - nk \frac{e^{ka} \Phi(k+a) - e^{-ka} \Phi(k-a)}{e^{ka} \Phi(k+a) + e^{-ka} \Phi(k-a)} + (Z - X\beta)' \Sigma_h^{\theta^{-1}} \mathbf{1}_n \end{aligned}$$

²⁰Cressie

²¹Sill

که در آن X ماتریس طرح حاصل از متغیرهای تبیینی با بعد $n \times q$ است و $\text{sign}(\cdot)$ تابع علامت است. بدیهی است که برآوردهای ماکسیمم درستنمایی پارامترهای مدل صورت بسته وجود ندارند و با استفاده از روش‌های عددی حل دستگاه معادلات $U_n = 0$ برآوردها را محاسبه می‌کنیم.

۴.۲.۴ ارزیابی کارایی مدل پیشنهادی

در این بخش با استفاده از مثال‌های شبیه‌سازی و واقعی، عملکرد مدل پیشنهادی را برای داده‌های فضایی دومی نسبت به مدل جانشین نرمال مقایسه می‌کنیم.

مطالعه شبیه‌سازی

در این مثال داده‌ها از مدل

$$z(s_i) = x_1(s_i) + 2x_2(s_i) + \delta(s_i), \quad i = 1, \dots, n$$

تولید شدند، که در آن $\delta = (\delta(s_1), \dots, \delta(s_n))'$ توسط یک فرآیند آمیخته از دو توزیع نرمال n متغیره به صورت

$$\frac{1}{2}N_n(\mu_1, \Sigma) + \frac{1}{2}N_n(\mu_2, \Sigma)$$

شبیه‌سازی شد، که در آن μ_1 و μ_2 بردارهای n بعدی ثابت با مولفه‌های به ترتیب ۲ و ۲- هستند و ماتریس کوواریانس Σ توسط یک مدل نمایی با پارامترهای $\theta_1 = 1/5$ و $\theta_2 = 0/5$ روی یک شبکه منظم ساخته شده است. این نوع شبیه‌سازی فرآیند $\delta(\cdot)$ منجر به تولید داده‌های دومی می‌شود. متغیرهای تبیینی x_1 و x_2 از توزیع نرمال استاندارد شبیه‌سازی شدند و سه حجم نمونه $n = 50, 100, 200$ نیز در نظر گرفته شدند.

برای برازش مدل، باید پارامتر مکانی مدل MBUN را برابر $z - \beta_1 x_1 - \beta_2 x_2$ قرار دهیم. برای ارزیابی برآوردهای مدل MBUN و مقایسه با برآوردهای حاصل از مدل نرمال (MN)، شبیه‌سازی را ۱۰۰ بار تکرار کردیم. مقادیر اریبی و MSE برآوردهای ماکسیمم درستنمایی ضرایب رگرسیونی برای هر دو مدل در جدول ۶.۴ گزارش شده‌اند. با توجه به مقادیر جدول، واضح است که مدل MBUN، بر حسب هر دو معیار اریبی و MSE، برآوردهای کاراتری نسبت به مدل MN ارائه کرده است. با مقایسه مقادیر MSE، مدل MBUN حداقل ۶۶ درصد کارایی بیشتری دارد. این وضعیت با افزایش حجم نمونه نیز بارزتر می‌شود.

دو مدل را از منظر دو معیار انتخاب مدل AIC و BIC نیز مورد مقایسه قرار دادیم. شکل ۷.۴ نمودار جعبه‌ای مقادیر این دو معیار را برای ۱۰۰ تکرار شبیه‌سازی برای دو مدل MBUN و MN نمایش می‌دهد. پراکندگی مقادیر هر دو معیار گویای توصیف بهتر داده‌ها توسط مدل MBUN است. بنابراین عدم توانایی مدل نرمال و در مقابل کارایی مدل MBUN در برازش پاسخ‌های فضایی دومی قابل مشاهده و نتیجه‌گیری است.

جدول ۶.۴: مقادیر اریبی و MSE برآوردهای پارامترها در مثال شبیه‌سازی

$\hat{a}$	$\hat{k}$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	n
۰/۰۰۰۸۸	۰/۲۲۳۲	۰/۰۰۰۸۳	۰/۰۰۶۵۷	۰/۰۰۱۸۱	۰/۰۰۰۲۱	۰/۰۰۰۸۲	۵۰
۰/۰۰۰۰۵	۰/۱۷۱۸	۰/۰۰۲۵۶	۰/۰۰۱۶۶	۰/۰۰۱۲۰	۰/۰۰۱۱۲	۰/۰۰۲۷۳	MSE
۰/۰۰۰۴۳	۰/۰۰۶۱۳	۰/۰۰۰۳۳	۰/۰۰۱۶۶	۰/۰۰۰۱۴	۰/۰۰۰۵۳	۰/۰۰۱۴۴	۱۰۰
۰/۰۰۰۰۲	۰/۰۰۵۸۷	۰/۰۰۰۸۲	۰/۰۰۰۷۹	۰/۰۰۰۴۰	۰/۰۰۰۴۹	۰/۰۰۱۹۰	MSE
۰/۰۰۰۲۸	۰/۰۰۵۵۷	۰/۰۰۰۱۲	۰/۰۰۱۸۹	۰/۰۰۰۲۰	۰/۰۰۰۳۷	۰/۰۰۰۶۸	۲۰۰
۰/۰۰۰۰۱	۰/۰۰۳۱۱	۰/۰۰۰۳۹	۰/۰۰۰۴۳	۰/۰۰۰۱۶	۰/۰۰۰۲۲	۰/۰۰۱۴۶	MSE
۰/۰۰۰۰۲	۰/۰۰۲۰۹	۰/۰۰۰۸۱	۰/۰۰۱۲۷	۰/۰۰۰۰۹	۰/۰۰۰۱۷	۰/۰۰۱۸۵	۵۰۰
۰/۰۰۰۰۰	۰/۰۰۰۹۲	۰/۰۰۰۱۷	۰/۰۰۰۲۰	۰/۰۰۰۰۴	۰/۰۰۰۰۴	۰/۰۰۱۵۳	MSE

جدول ۷.۴: مقادیر اریبی و MSE برآوردهای ضرایب رگرسیونی در مثال شبیه‌سازی

n		۵۰		۱۰۰		۲۰۰		مدل
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	
اریبی		۰/۰۰۰۸۴	۰/۰۰۴۳۷	۰/۰۰۱۹۲	۰/۰۰۲۱۶	۰/۰۰۰۱۲	۰/۰۰۱۴۴	MBUN
MSE		۰/۰۰۵۰۱	۰/۰۰۵۷۸	۰/۰۰۲۲۶	۰/۰۰۱۸۱	۰/۰۰۰۹۳	۰/۰۰۱۳۰	
اریبی		۰/۰۰۶۰۸	۰/۰۰۶۰۱	۰/۰۰۵۴۴	۰/۰۰۲۳۷	۰/۰۰۱۲۴	۰/۰۰۰۱۴	MN
MSE		۰/۰۰۱۳۰۸	۰/۰۰۱۰۱۱	۰/۰۰۹۸۹	۰/۰۰۴۲۸	۰/۰۰۲۵۸	۰/۰۰۲۲۲	

مثال‌های واقعی

در این بخش عملکرد و کاربرد مدل MBUN را در برازش دو مثال واقعی مورد بررسی قرار می‌دهیم و از دو معیار AIC و BIC برای مقایسه با مدل نرمال استفاده می‌کنیم. اولین مثال مجموعه داده غلظت فلزات سنگین است که در نزدیکی رودخانه میوز^{۲۲} نمونه‌برداری شده‌اند و در بسته georob^{۲۳} در نرم‌افزار R با نام Meuse در دسترس است. این مجموعه داده شامل ۱۴ متغیر و ۱۶۴ مشاهده است. مدل خطی که در نظر گرفتیم، به صورت

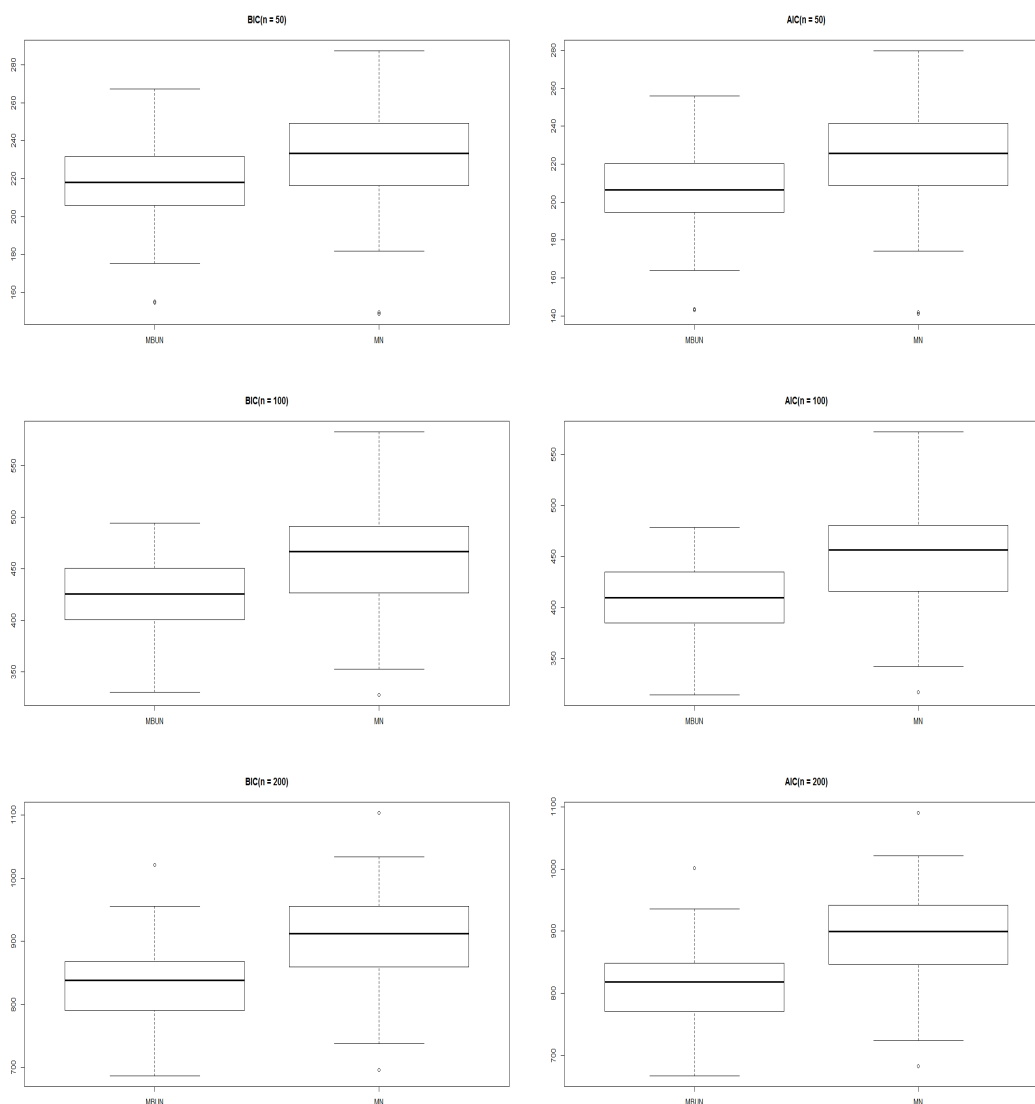
$$\log(\text{zinc}) = \beta_1 + \beta_2 \sqrt{\text{dist}} + \beta_3 \text{cadmium} + \delta$$

است که در آن zinc غلظت عنصر روی، dist فاصله تا رودخانه و cadmium غلظت کادمیم سطح خاک است.

مجموعه داده دوم مربوط به خواص شیمیایی اندازه‌گیری شده خاک با ۲۵۰ مشاهده و ۲۲

^{۲۲}Meuse

^{۲۳}<https://cran.r-project.org/package=georob>



شکل ۷.۴: مقادیر AIC (ستون سمت راست) و BIC (ستون سمت چپ) برای دو مدل MBUN و MN به ازای حجم‌های نمونه مختلف

متغیر است که در بسته ^{۲۴}geoR در نرم‌افزار R با نام soil250 موجود است. مدل مورد نظر برای این داده‌ها به صورت

$$CTC = \beta_1 + \beta_2 K + \beta_3 Ca + \delta$$

است که در آن CTC توانایی تبادل کاتیوم، K مقدار پتاسیم و Ca مقدار کلسیم موجود در خاک است. برای هر دو مثال از مدل هم‌تغییرنگار نمایی استفاده کردیم. نتایج برآورد پارامترهای رگرسیونی هر دو مدل به همراه معیارهای AIC و BIC در جدول ۸.۴ گزارش شده‌اند. با توجه به نتایج جدول، برای هر دو مثال، مدل MBUN نسبت به مدل MN برتر است. با استفاده از بافت نگار مانده‌های استاندارد شده و منحنی‌های دو تابع چگالی برآورد شده (شکل ۸.۴)،

²⁴<https://cran.r-project.org/package=geoR>

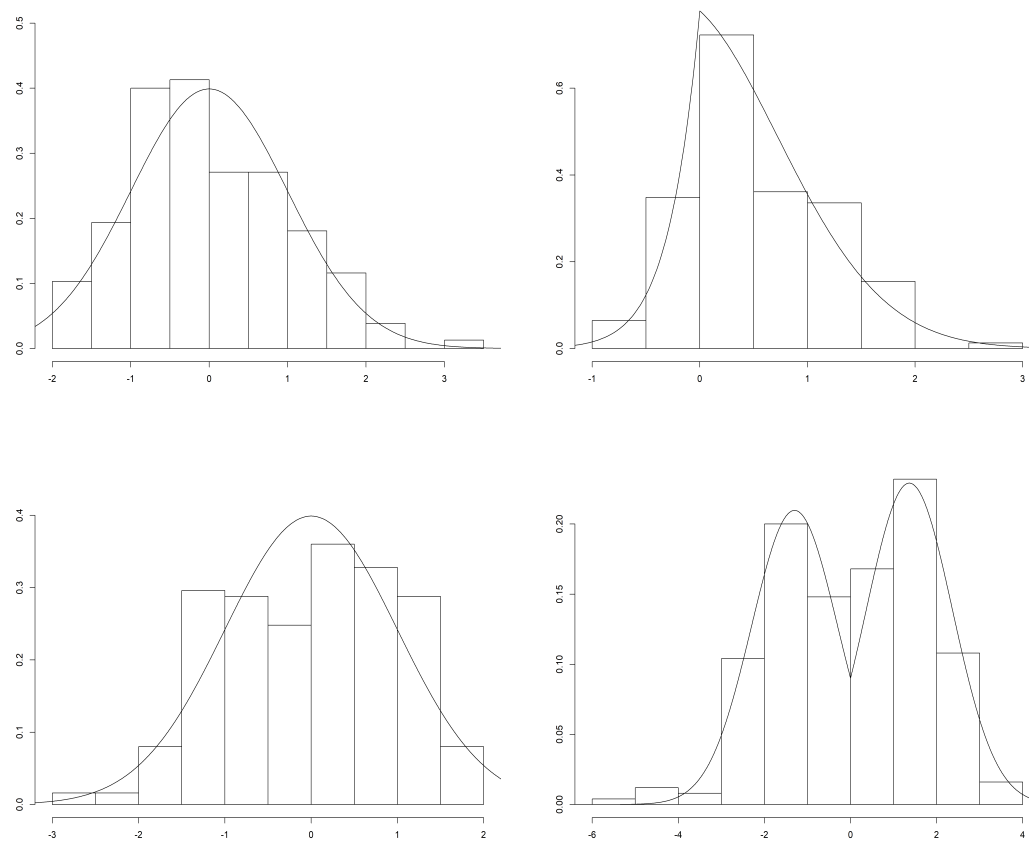
جدول ۸.۴: برآورد پارامترهای رگرسیونی و معیارهای انتخاب مدل دو مجموعه داده واقعی

BIC	AIC	$\hat{\beta}_3$	$\hat{\beta}_2$	$\hat{\beta}_1$	مدل	مجموعه داده
۱۲۶/۴۲	۱۰۵/۱۲	۰/۱۱	-۱/۱۴	۵/۷۵	MBUN	Mesue
۱۲۷/۳۰	۱۱۲/۰۸	۰/۱۲	-۱/۱۹	۶/۰۲	MN	
۳۴۲/۰۴	۳۱۷/۳۹	۱/۲۷	۱/۳۵	۲/۷۰	MBUN	soil250
۴۷۳/۹۳	۴۵۶/۳۳	۱/۰۲	۰/۶۱	۳/۸۵	MN	

نیکویی برازش مدل MBUN کاملاً مشهود است. افزون بر این، عدم نیکویی برازش مدل نرمال در هر دو مثال نیز واضح است؛ در مجموعه داده Meuse مانده‌های مدل نرمال چوله هستند و در مجموعه داده soil250 مانده‌های مدل نرمال دومی هستند. بنابراین واضح است که مدل MBUN انعطاف لازم برای مدل‌بندی هر دو ویژگی چولگی و دومی بودن داده‌ها را داراست.

۵.۲.۴ نتیجه‌گیری

در این فصل، در مورد تحلیل رگرسیون خطی داده‌های چوله و داده‌های شامل مشاهده پرت یک‌طرفه بحث کردیم و کاربردهای توزیع BUN را در این نوع داده‌ها تشریح کردیم. رفتار مجانبی برآوردگرهای ML را بررسی کرده و نشان دادیم سازگار ضعیف و دارای توزیع مجانبی نرمال هستند. با مثال‌های شبیه‌سازی و واقعی نشان دادیم که توزیع BUN کارایی خوبی دارد. همچنین BUN را به نسخه چندمتغیره گسترش دادیم و در مورد داده‌های فضایی چوله و دومی بحث کردیم و با مثال‌های شبیه‌سازی و واقعی عملکرد کارا و بهتر مدل پیشنهادی را نسبت به مدل نرمال زمانی که داده‌ها چوله یا دومی هستند، نشان دادیم.



شکل ۸.۴: برازش توابع چگالی برآوردشده MBUN و MN بر مانده‌های استانداردشده دو مدل در دو مثال واقعی. مجموعه داده اول (بالا) و مجموعه داده دوم (پایین)

فصل ۵

پیشنهادهای برای پژوهش‌های آینده

- می‌توان توزیع‌های معرفی‌شده در این رساله را به حالت چندمتغیره و حتی ماتریسی توسعه داده و در تحلیل داده‌های چندمتغیره مانند داده‌های فضایی و طولی به کار برد.
- می‌توان توزیع‌های معرفی‌شده در این رساله را به نسخه‌های لگاریتمی توسعه داد و در داده‌های بقا و قابلیت اعتماد به کار برد.
- می‌توان مدل‌های جمعی تعمیم‌یافته برای مکان، مقیاس و شکل بر پایه توزیع‌های معرفی‌شده را در این رساله بررسی کرد و توسعه داد.
- یکی از مباحثی که در تحلیل داده‌های زمین‌آماري مطرح و مهم است، پیشگویی فضایی در مکان‌هایی است که فاقد مشاهده هستند. این مساله برای داده‌هایی که دومدی یا چوله هستند، می‌تواند انگیزه‌ای برای پژوهش بر روی آن توسط مدل پیشنهادی باشد.

پیوست آ

اثبات قضایا

در این فصل به اختصار به اثبات قضایایی که در فصل‌های ۳، ۴ و ۵ ارائه شده‌اند، می‌پردازیم.

۱. آ اثبات قضایای فصل سوم

اثبات گزاره ۱.۱.۲:

$$F(z) = \int_{-\infty}^z c_{\nu} \exp(t) \left(1 + \frac{1}{\nu} \exp(2t)\right)^{-\frac{\nu+1}{2}} dt.$$

با تغییر متغیر $u = \frac{\exp(\Upsilon t)}{\nu + \exp(\Upsilon t)}$ داریم

$$\begin{aligned}
 F(z) &= \frac{\exp(\Upsilon z)}{\nu + \exp(\Upsilon z)} \int_0^1 c_\nu \frac{\nu^{\frac{1}{\Upsilon}} u^{\frac{1}{\Upsilon}}}{(1-u)^{\frac{1}{\Upsilon}}} \nu^{\frac{\nu+1}{\Upsilon}} \left(\frac{\nu}{1-u} \right)^{-\frac{\nu+1}{\Upsilon}} \frac{1}{\Upsilon u(1-u)} du \\
 &= \frac{\exp(\Upsilon z)}{\nu + \exp(\Upsilon z)} \int_0^1 \frac{c_\nu}{\Upsilon} \nu^{\frac{1}{\Upsilon}} u^{\frac{1}{\Upsilon}-1} (1-u)^{\frac{\nu}{\Upsilon}-1} du \\
 &= \frac{\exp(\Upsilon z)}{\nu + \exp(\Upsilon z)} \int_0^1 \frac{1}{\Upsilon} \nu^{\frac{1}{\Upsilon}} \frac{\Upsilon \Gamma\left(\frac{\nu+1}{\Upsilon}\right)}{\sqrt{\nu\pi} \Gamma\left(\frac{\nu}{\Upsilon}\right)} u^{\frac{1}{\Upsilon}-1} (1-u)^{\frac{\nu}{\Upsilon}-1} du \\
 &= \frac{\exp(\Upsilon z)}{\nu + \exp(\Upsilon z)} \int_0^1 \frac{1}{B\left(\frac{1}{\Upsilon}, \frac{\nu}{\Upsilon}\right)} u^{\frac{1}{\Upsilon}-1} (1-u)^{\frac{\nu}{\Upsilon}-1} du \\
 &= I_{\frac{\exp(\Upsilon z)}{\nu + \exp(\Upsilon z)}} \left(\frac{1}{\Upsilon}, \frac{\nu}{\Upsilon} \right).
 \end{aligned}$$

اثبات گزاره ۲.۱.۲:

$$M(t) = \int_{-\infty}^{\infty} c_\nu \exp(tz) \exp(z) \left(1 + \frac{1}{\nu} \exp(\Upsilon z) \right)^{-\frac{\nu+1}{\Upsilon}} dz.$$

با تغییر متغیر $u = \frac{\exp(\Upsilon z)}{\nu + \exp(\Upsilon z)}$ داریم

$$\begin{aligned}
 M(t) &= \int_0^1 c_\nu \frac{\nu^{\frac{t+1}{\Upsilon}} u^{\frac{t+1}{\Upsilon}}}{(1-u)^{\frac{t+1}{\Upsilon}}} \nu^{\frac{\nu+1}{\Upsilon}} \left(\frac{\nu}{1-u} \right)^{-\frac{\nu+1}{\Upsilon}} \frac{1}{\Upsilon u(1-u)} du \\
 &= \int_0^1 \frac{c_\nu}{\Upsilon} \nu^{\frac{t+1}{\Upsilon}} u^{\frac{t+1}{\Upsilon}-1} (1-u)^{\frac{\nu-t}{\Upsilon}-1} du \\
 &= \frac{c_\nu}{\Upsilon} \nu^{\frac{t+1}{\Upsilon}} B\left(\frac{t+1}{\Upsilon}, \frac{\nu-t}{\Upsilon}\right) \\
 &= \frac{\nu^{\frac{t}{\Upsilon}} \Gamma\left(\frac{1+t}{\Upsilon}\right) \Gamma\left(\frac{\nu-t}{\Upsilon}\right)}{\sqrt{\pi} \Gamma\left(\frac{\nu}{\Upsilon}\right)}.
 \end{aligned}$$

چون اثبات تابع مشخصه شبیه این است از اثبات آن صرف نظر می کنیم.

اثبات گزاره ۳.۱.۲: فرض کنید X دارای تابع چگالی

$$f(x) = \frac{\cos\left(\frac{\pi}{\Upsilon} \alpha\right)}{\pi \sigma} \frac{e^{\alpha\left(\frac{x-\mu}{\sigma}\right)}}{\cosh\left(\frac{x-\mu}{\sigma}\right)}$$

باشد، آن گاه تابع چگالی $Y = -X$ به صورت زیر داده می شود:

$$\begin{aligned} f(y) &= \frac{\cos\left(\frac{\pi}{\gamma}\alpha\right)}{\pi\sigma} \frac{e^{\alpha\left(\frac{-y-\mu}{\sigma}\right)}}{\cosh\left(\frac{-y-\mu}{\sigma}\right)} \\ &= \frac{\cos\left(\frac{\pi}{\gamma}\alpha\right)}{\pi\sigma} \frac{e^{\alpha\left(\frac{-(y+\mu)}{\sigma}\right)}}{\cosh\left(\frac{-(y+\mu)}{\sigma}\right)}. \end{aligned}$$

با استفاده از زوج بودن توابع $\cos(\cdot)$ و $\cosh(\cdot)$ داریم

$$f(y) = \frac{\cos\left(-\frac{\pi}{\gamma}\alpha\right)}{\pi\sigma} \frac{e^{-\alpha\left(\frac{(y-(-\mu))}{\sigma}\right)}}{\cosh\left(\frac{(y-(-\mu))}{\sigma}\right)},$$

که این تابع چگالی $\text{SHS}_{\gamma}(-\mu, \sigma, -\alpha)$ است.

اثبات گزاره ۴.۱.۲:

$$F(z) = \int_{-\infty}^t \frac{\cos\left(\frac{\pi}{\gamma}\alpha\right)}{\pi} \frac{e^{\alpha t}}{\cosh(t)} dt$$

با تغییر متغیر $u = \frac{\exp(\gamma t)}{1 + \exp(\gamma t)}$ داریم

$$F(z) = \int_{\circ} \frac{\exp(\gamma z)}{1 + \exp(\gamma z)} \frac{\cos\left(\frac{\pi}{\gamma}\alpha\right)}{\pi} u^{\frac{1+\alpha}{\gamma}-1} (1-u)^{\frac{1-\alpha}{\gamma}-1} du.$$

می دانیم

$$\cos\left(\frac{\pi}{\gamma}\alpha\right) = \sin\left(\frac{\pi}{\gamma} - \frac{\pi}{\gamma}\alpha\right) = \sin\left(\pi\left(\frac{1-\alpha}{\gamma}\right)\right)$$

و

$$\frac{\sin(\pi\gamma)}{\pi} = \frac{1}{B(\gamma, 1-\gamma)}$$

و در نتیجه

$$\frac{\cos\left(\frac{\pi}{\gamma}\alpha\right)}{\pi} = \frac{1}{B\left(\frac{1-\alpha}{\gamma}, \frac{1+\alpha}{\gamma}\right)}.$$

بنابراین

$$F(z) = \int_{\circ} \frac{\exp(\gamma z)}{1 + \exp(\gamma z)} \frac{1}{B\left(\frac{1-\alpha}{\gamma}, \frac{1+\alpha}{\gamma}\right)} u^{\frac{1+\alpha}{\gamma}-1} (1-u)^{\frac{1-\alpha}{\gamma}-1} du = I_{\frac{\exp(\gamma z)}{1 + \exp(\gamma z)}} \left(\frac{1+\alpha}{\gamma}, \frac{1-\alpha}{\gamma}\right)$$

اثبات گزاره ۵.۱.۲:

$$\begin{aligned}
C(t) &= \int_{-\infty}^{\infty} e^{itz} \frac{\cos\left(\frac{\pi}{\gamma}\alpha\right)}{\pi\sigma} \frac{e^{\alpha z}}{\cosh(z)} dz \\
&= \frac{\cos\left(\frac{\pi}{\gamma}\alpha\right)}{\cos\left(\frac{\pi}{\gamma}(it+\alpha)\right)} \int_{-\infty}^{\infty} \frac{\cos\left(\frac{\pi}{\gamma}(it+\alpha)\right)}{\pi\sigma} \frac{e^{(it+\alpha)z}}{\cosh(z)} dz \\
&= \frac{\cos\left(\frac{\pi}{\gamma}\alpha\right)}{\cos\left(\frac{\pi}{\gamma}(it+\alpha)\right)}.
\end{aligned}$$

چون اثبات تابع مولد گشتاور شبیه این است از اثبات آن صرف نظر می‌کنیم. **اثبات قضیه ۱.۱.۲:** فرض کنید $Z \sim \text{SHS}_1(\circ, 1, \nu)$ باشد. تابع چگالی Z را می‌توان به صورت زیر نوشت:

$$f(z) \propto \exp(-z)h(z)$$

که در آن

$$h(z) = \frac{\left(\frac{1}{\nu} + \exp(-\gamma z)\right)^{-1}}{\left(1 + \frac{1}{\nu} \exp(\gamma z)\right)^{\frac{\nu-1}{\gamma}}}.$$

از این رو

$$\lim_{z \rightarrow \infty} \frac{h(\gamma z)}{h(z)} = 1.$$

بنابراین SHS_1 تابع چگالی نیمه دم‌سنگین با $\eta = 1$ و $\gamma = \circ$ است. برای $\text{SHS}_2(\circ, 1, \alpha)$ با $\eta = 1$ و $\gamma = \circ$

$$h(z) = \frac{\exp(\alpha z)}{1 + \exp(-\gamma z)}$$

نتیجه مشابهی به دست می‌آید.

اثبات قضیه ۲.۱.۲: برای اثبات ۲.۱.۲، به لم زیر نیاز داریم.

لم ۱.۱.۱. تحت شرایط ۳، ۴ و ۵ داریم،

$$-\frac{1}{n}\ell(\gamma) \xrightarrow{P} B = \begin{bmatrix} B(\beta) & B(\beta, \sigma) & B(\beta, \nu) \\ B(\beta, \sigma) & B(\sigma) & B(\sigma, \nu) \\ B(\beta, \nu) & B(\sigma, \nu) & B(\nu) \end{bmatrix}$$

که در آن

$$\begin{aligned}
\lim_{n \rightarrow \infty} n^{-1} B_n(\beta) &= B(\beta) = \frac{c'_\nu}{\sigma^\nu} \text{Beta} \left(\frac{\nu}{\sigma}, \frac{\nu+1}{\sigma} \right) C \\
\lim_{n \rightarrow \infty} n^{-1} B_n(\beta_j, \sigma) &= B(\beta_j, \sigma) = \frac{c'_\nu}{\sigma^\nu} \left(\text{Beta} \left(\frac{\nu}{\sigma}, \frac{\nu+1}{\sigma} \right) (\log(\nu) + \psi_\nu) + \right. \\
&\quad \left. \frac{1}{\sigma} \text{Beta} \left(\frac{\nu}{\sigma}, \frac{\nu+1}{\sigma} \right) \right) c_j - \frac{1}{\sigma^\nu} c_j \\
\lim_{n \rightarrow \infty} n^{-1} B_n(\beta_j, \nu) &= B(\beta_j, \nu) = \frac{c'_\nu}{\sigma^\nu} \text{Beta} \left(\frac{\nu}{\sigma}, \frac{\nu+1}{\sigma} \right) - \frac{c'_\nu}{\sigma(\nu+1)} \text{Beta} \left(\frac{\nu}{\sigma}, \frac{\nu+1}{\sigma} \right) c_j \\
\lim_{n \rightarrow \infty} n^{-1} B_n(\sigma) &= B(\sigma) = n^{-1} I_n(\sigma) \\
\lim_{n \rightarrow \infty} n^{-1} B_n(\sigma, \nu) &= B(\sigma, \nu) = n^{-1} I_n(\sigma, \nu) \\
\lim_{n \rightarrow \infty} n^{-1} B_n(\nu) &= B(\nu) = n^{-1} I_n(\nu).
\end{aligned}$$

اثبات لم ۱.۱.۴: می توان نوشت

$$-n^{-1} \frac{\partial^2 \ell(\theta)}{\partial \beta_j \partial \beta_k} = -n^{-1} \frac{\partial^2 \ell(\theta)}{\partial \beta_j \partial \beta_k} - n^{-1} B_n(\beta)_{(jk)} + n^{-1} B_n(\beta)_{(jk)}.$$

بنابراین

$$\begin{aligned}
&\lim_{n \rightarrow \infty} n^{-\nu} \mathbb{E} \left(\frac{\partial^2 \ell(\theta)}{\partial \beta_j \partial \beta_k} + B_n(\beta)_{(jk)} \right)^\nu \\
&= \lim_{n \rightarrow \infty} n^{-\nu} \text{Var} \left(\frac{\nu+1}{\sigma^\nu} \sum_{i=1}^n x_{ij} x_{ik} \left(\frac{\frac{1}{\nu} \exp \left(\nu \frac{y_i - \beta' x_i}{\sigma} \right)}{\left(1 + \frac{1}{\nu} \exp \left(\nu \frac{y_i - \beta' x_i}{\sigma} \right) \right)^\nu} \right) \right).
\end{aligned}$$

با استفاده از این که $\{\epsilon, \epsilon_i = y_i - \beta' x_i\}$ برای $i = 1, 2, \dots, n$ مستقل و هم توزیع هستند، داریم

$$\text{cov} \left(\frac{\frac{1}{\nu} \exp \left(\nu \frac{\epsilon_i}{\sigma} \right)}{\left(1 + \frac{1}{\nu} \exp \left(\nu \frac{\epsilon_i}{\sigma} \right) \right)^\nu}, \frac{\frac{1}{\nu} \exp \left(\nu \frac{\epsilon_j}{\sigma} \right)}{\left(1 + \frac{1}{\nu} \exp \left(\nu \frac{\epsilon_j}{\sigma} \right) \right)^\nu} \right) = 0, \quad i \neq j$$

و در نتیجه

$$\begin{aligned}
&\lim_{n \rightarrow \infty} n^{-\nu} \mathbb{E} \left(\frac{\partial^2 \ell(\theta)}{\partial \beta_j \partial \beta_k} + B_n(\beta)_{(jk)} \right)^\nu \\
&= \lim_{n \rightarrow \infty} n^{-\nu} \frac{(\nu+1)^\nu}{\sigma^\nu} \text{Var} \left(\frac{\frac{1}{\nu} \exp \left(\nu \frac{\epsilon}{\sigma} \right)}{\left(1 + \frac{1}{\nu} \exp \left(\nu \frac{\epsilon}{\sigma} \right) \right)^\nu} \right) \sum_{i=1}^n x_{ij} x_{ik} = 0.
\end{aligned}$$

بنابراین

$$-n^{-1} \frac{\partial^2 \ell(\theta)}{\partial \beta_j \partial \beta_k} = o_p(1) + n^{-1} B_n(\beta)_{(j,k)}.$$

یعنی

$$-n^{-1} \frac{\partial^2 \ell(\theta)}{\partial \beta_j \partial \beta_k} \xrightarrow{P} B(\beta)_{(j,k)}.$$

به طور مشابه می توان نشان داد که

$$\begin{aligned} -n^{-1} \frac{\partial^2 \ell(\theta)}{\partial \beta_j \partial \sigma} &\xrightarrow{P} B(\beta, \sigma)_{(j)} \\ -n^{-1} \frac{\partial^2 \ell(\theta)}{\partial \beta_j \partial \nu} &\xrightarrow{P} B(\beta, \nu)_{(j)} \\ -n^{-1} \frac{\partial^2 \ell(\theta)}{\partial \sigma^2} &\xrightarrow{P} B(\sigma) \\ -n^{-1} \frac{\partial^2 \ell(\theta)}{\partial \sigma \partial \nu} &\xrightarrow{P} B(\sigma, \nu) \\ -n^{-1} \frac{\partial^2 \ell(\theta)}{\partial \nu^2} &\xrightarrow{P} B(\nu). \end{aligned}$$

که این اثبات را کامل می کند. اکنون با استفاده از لم ۱.۱.۴ می توانیم اثبات قضیه ۲.۱.۲ را شروع کنیم. بنابر سویتینگ^۱ (۱۹۸۰) و ماردیا و مارشال^۲ (۱۹۸۴) نتیجه عمومی برای سازگاری و توزیع مجانبی نرمال برآوردهای درستنمایی ماکسیمم مستلزم برقراری سه شرط نظم (است: ۱) پیوستگی، (۲) رشد و (۳) همگرایی ماتریس اطلاع فشر. شرط پیوستگی با دو بار مشتق پذیری لگاریتم تابع درستنمایی در فضای پارامتر و مشتق دوم پیوسته آن در مدل پیشنهادی برقرار است. شرط رشد ماتریس اطلاع فشر نیز برقرار است، اگر برای $n \rightarrow \infty$ داشته باشیم $\lim_{n \rightarrow \infty} B_n(\theta) = \infty$ که این شرط برقرار است، چون

$$\begin{aligned} \lim_{n \rightarrow \infty} n^{-1} B_n(\beta) &= B(\beta) & \rightarrow \lim_{n \rightarrow \infty} B_n(\beta) &= \infty \\ \lim_{n \rightarrow \infty} n^{-1} B_n(\beta_j, \sigma) &= B(\beta_j, \sigma) & \rightarrow \lim_{n \rightarrow \infty} B_n(\beta_j, \sigma) &= \infty \\ \lim_{n \rightarrow \infty} n^{-1} B_n(\beta_j, \nu) &= B(\beta_j, \nu) & \rightarrow \lim_{n \rightarrow \infty} B_n(\beta_j, \nu) &= \infty \\ \lim_{n \rightarrow \infty} n^{-1} B_n(\sigma) &= B(\sigma) & \rightarrow \lim_{n \rightarrow \infty} B_n(\sigma) &= \infty \\ \lim_{n \rightarrow \infty} n^{-1} B_n(\sigma, \nu) &= B(\sigma, \nu) & \rightarrow \lim_{n \rightarrow \infty} B_n(\sigma, \nu) &= \infty \\ \lim_{n \rightarrow \infty} n^{-1} B_n(\nu) &= B(\nu) & \rightarrow \lim_{n \rightarrow \infty} B_n(\nu) &= \infty. \end{aligned}$$

شرط سوم یعنی همگرایی ماتریس اطلاع فشر مشاهده شده یعنی $\ell_n^{(2)}$ برقرار خواهد شد هرگاه $B_n(\theta)^{-\frac{1}{2}} \ell_n^{(2)} B_n(\theta)^{-\frac{1}{2}}$ در احتمال به ماتریس همبستگی میل کند. با استفاده از لم ۱.۱.۴ داریم $B \rightarrow -n^{-1} \ell^{(2)}$ و همچنین $n^{-1} B_n \rightarrow B$ بنابراین

$$\lim_{n \rightarrow \infty} -B_n(\theta)^{-\frac{1}{2}} \ell_n^{(2)} B_n(\theta)^{-\frac{1}{2}} = \lim_{n \rightarrow \infty} n^{\frac{1}{2}} B_n(\theta)^{-\frac{1}{2}} \left(-n^{-1} \ell_n^{(2)} \right) n^{\frac{1}{2}} B_n(\theta)^{-\frac{1}{2}} = I.$$

که این اثبات را کامل می کند.

اثبات قضیه ۳.۱.۲: اثبات این قضیه مشابه اثبات قضیه ۲.۱.۲ با جایگزاری ℓ زیر به جای ℓ ۱.۱.۴ است.

¹Sweeting

²Mardia and Kent

لم ۲.۱.آ. تحت شرایط ۳، ۴ و ۵ داریم،

$$-\frac{1}{n}\tilde{\ell}^{(2)} \xrightarrow{P} \tilde{B} = \begin{bmatrix} \tilde{B}(\beta) & \tilde{B}(\beta, \sigma) & \tilde{B}(\beta, \alpha) \\ \tilde{B}(\beta, \sigma) & \tilde{B}(\sigma) & \tilde{B}(\sigma, \alpha) \\ \tilde{B}(\beta, \alpha) & \tilde{B}(\sigma, \alpha) & \tilde{B}(\alpha) \end{bmatrix},$$

که در آن

$$\begin{aligned} \lim_{n \rightarrow \infty} n^{-1} \tilde{B}_n(\beta) &= \tilde{B}(\beta) = \frac{\cos\left(\frac{\pi}{4}\alpha\right)}{\pi\sigma^2} \text{Beta}\left(\frac{\alpha+3}{2}, \frac{3-\alpha}{2}\right) C', \\ \lim_{n \rightarrow \infty} n^{-1} \tilde{B}_n(\beta_j, \sigma) &= \tilde{B}(\beta_j, \sigma) = \alpha \frac{c_j}{\sigma^2} + \frac{\cos\left(\frac{\pi}{4}\alpha\right)}{\pi} \left(2 \text{Beta}\left(\frac{\alpha+3}{2}, \frac{1-\alpha}{2}\right) \right. \\ &\quad \left. - \text{Beta}\left(\frac{\alpha+1}{2}, \frac{1-\alpha}{2}\right) \right) \frac{c_j}{\sigma^2} \\ &\quad + 2 \frac{\cos\left(\frac{\pi}{4}\alpha\right)}{\pi} \text{Beta}\left(\frac{\alpha+3}{2}, \frac{3-\alpha}{2}\right) \left(\psi\left(\frac{\alpha+3}{2}\right) \right. \\ &\quad \left. - \psi\left(\frac{3-\alpha}{2}\right) \right) \frac{c_j}{\sigma^2}, \\ \lim_{n \rightarrow \infty} n^{-1} \tilde{B}_n(\beta_j, \alpha) &= \tilde{B}(\beta_j, \alpha) = -\frac{1}{\sigma^2} c_j, \\ \lim_{n \rightarrow \infty} n^{-1} \tilde{B}_n(\sigma) &= \tilde{B}(\sigma) = \frac{\cos\left(\frac{\pi}{4}\alpha\right)}{\pi} \left(\text{Beta}\left(\frac{\alpha+3}{2}, \frac{1-\alpha}{2}\right) \left(\psi\left(\frac{\alpha+3}{2}\right) - \psi\left(\frac{1-\alpha}{2}\right) \right) \right. \\ &\quad \left. - \text{Beta}\left(\frac{\alpha+1}{2}, \frac{1-\alpha}{2}\right) \left(\psi\left(\frac{\alpha+1}{2}\right) - \psi\left(\frac{1-\alpha}{2}\right) \right) \right) + \frac{\alpha\pi}{\sigma^2} \tan\left(\frac{\pi}{4}\alpha\right) \\ &\quad + 2 \frac{\cos\left(\frac{\pi}{4}\alpha\right)}{\pi} \text{Beta}\left(\frac{\alpha+3}{2}, \frac{3-\alpha}{2}\right) \left(\psi\left(\frac{\alpha+3}{2}\right) - \psi\left(\frac{3-\alpha}{2}\right) \right) + \frac{1}{\sigma^2}, \\ \lim_{n \rightarrow \infty} n^{-1} \tilde{B}_n(\sigma, \alpha) &= \tilde{B}(\sigma, \alpha) = \frac{\pi}{4\sigma} \tan\left(\frac{\pi}{4}\alpha\right), \\ \lim_{n \rightarrow \infty} n^{-1} \tilde{B}_n(\alpha) &= \tilde{B}(\alpha) = \frac{\pi^2}{4} \left(1 + \tan^2\left(\frac{\pi}{4}\alpha\right) \right). \end{aligned}$$

اثبات لم ۲.۱.آ: اثبات این لم مشابه اثبات لم ۱.۱.آ است.

۲.آ اثبات قضایای فصل چهارم

اثبات قضیه ۱.۲.۳:

(الف) از متقارن بودن $g(\cdot)$ و $w(\cdot)$ نتیجه می شود.

(ب) چون $f_X(x)$ تابع چگالی با مقادیر $\mathbb{R}$ است، حداقل دارای یک مد است، فرض کنید m_1 مد $f_X(x)$ باشد. با توجه به محدب بودن $w(\cdot)$ می توان نشان داد $m_1 \neq c$. بدون کاستن از

کلیت مسئله می‌توان فرض کرد $m_1 > c$ است. با توجه به اینکه $f_X(x)$ حول c متقارن است $m_2 = 2c - m_1$ نیز مد دوم $f_X(x)$ است.

(پ) از قسمت ب داریم $m_1 + m_2 = m_1 + 2c - m_1 = 2c$

اثبات قضیه ۲.۲.۳: در این قسمت ثابت می‌کنیم که (۳.۳) و (۴.۳) توابع چگالی هستند. برای اثبات تابع چگالی بودن (۳.۳) داریم

$$\int_{-\infty}^{\infty} e^{kG(|x|)} g(x) dx = \int_{-\infty}^0 e^{kG(-x)} g(x) dx + \int_0^{\infty} e^{kG(|x|)} g(x) dx.$$

از تساوی $G(-x) = 1 - G(x)$ داریم

$$\int_{-\infty}^0 e^{k(1-G(x))} g(x) dx + \int_0^{\infty} e^{kG(x)} g(x) dx$$

و با تغییر متغیر $u = G(x)$ داریم

$$\int_0^1 e^{k(1-u)} du + \int_0^1 e^{ku} du = \frac{2 \left(e^k - e^{\frac{k}{2}} \right)}{k}.$$

برای اثبات تابع چگالی بودن (۴.۳) داریم

$$\begin{aligned} \int_{-\infty}^{\infty} G(|x|)^k g(x) dx &= \int_{-\infty}^0 G(-x)^k g(x) dx + \int_0^{\infty} G(x)^k g(x) dx \\ &= \int_{-\infty}^0 (1 - G(x))^k g(x) dx + \int_0^{\infty} G(x)^k g(x) dx \\ &= \int_{-\infty}^0 (1 - u)^k du + \int_0^{\infty} u^k du = \frac{2 \left(1 - \frac{1}{2^{k+1}} \right)}{k+1}. \end{aligned}$$

دومدی بودن (۳.۳) و (۴.۳) از قضیه ۱.۲.۳ به دلیل این که $w(x) = e^{kG(|x|)}$ و $w(x) = G(|x|)^k$ اکیدا مثبت و محدب هستند، نتیجه می‌شود.

اثبات قضیه ۳.۲.۳:

الف) چون $f_X(x) \propto g(h(x))$ است، مد $f_X(x)$ را با مشتق گرفتن از $g(h(x))$ و برابر صفر قرار دادن آن به دست می‌آوریم. داریم

$$h'(x)g'(h(x)) = 0$$

که نتیجه می‌دهد $h'(x) = 0$ یا $g'(h(x)) = 0$. چون $h(x)$ محدب و حول 0 متقارن است، جواب معادله $h'(x) = 0$ برابر 0 است. یعنی 0 مد (۵.۳) است. برای معادله $g'(h(x)) = 0$ چون مد $g(x)$ برابر صفر است، نتیجه می‌شود $h(x) = 0$ در نتیجه $x = 0$ یعنی تنها مد (۵.۳) است.

(ب) اگر $k \neq 0$ باشد، معادله $h'(x) = 0$ و $g'(h(x)) = 0$ قسمت الف برای $h'(x) = 0$ جواب $x = 0$ و جواب $g'(h(x)) = 0$ حل معادله $h(x) = k$ است که چون $h(x)$ محدب است دو جواب دارد. تا اینجا نشان دادیم که

$$h'(x)g'(h(x)) = 0$$

سه جواب دارد. با آزمون مشتق دوم می‌توان تایید کرد که دو جواب حاصل از $h(x) = k$ ماکسیمم نسبی تابع هستند. مشتق دوم (۵.۳) برابر است با

$$h''(x)g'(h(x)) + h'(x)g''(h(x)).$$

برای $h(x) = k$ داریم

$$h''(x)g'(k) + h'(x)g''(k).$$

چون k مد تابع چگالی $g(\cdot)$ است، $g'(k)$ برابر صفر و $g''(k)$ مثبت است. بنابراین

$$h'(x)g''(k) > 0.$$

(پ) از قسمت ب نتیجه می‌شود.

اثبات قضیه ۴.۲.۳: با قرار دادن $h(x) = |x|$ در (۵.۳) داریم

$$f_X(x) = \frac{g(|x|)}{\int_{-\infty}^{\infty} g(|w|) dw}$$

که به صورت زیر می‌توان نوشت

$$f_X(x) = \begin{cases} \frac{g(x)}{\int_0^{\infty} g(w) dw} & \text{if } x \geq 0; \\ \frac{g(-x)}{\int_{-\infty}^0 g(-w) dw} & \text{if } x < 0. \end{cases}$$

یا

$$f_X(x) = \begin{cases} \frac{g(x)}{1 - G_X(0)} & \text{if } x \geq 0; \\ \frac{g(-x)}{G_{-X}(0)} & \text{if } x < 0. \end{cases}$$

که در آن G_X و G_{-X} به ترتیب توابع توزیع X و $-X$ هستند. با توجه به اینکه $G_{-X}(0) = 1 - G_X(0)$ است، داریم

$$f_X(x) = \begin{cases} \frac{g(x)}{1 - G_X(0)} & \text{if } x \geq 0; \\ \frac{g(-x)}{1 - G_X(0)} & \text{if } x < 0 \end{cases}$$

و چون

$$1 - G_X(\circ) = P(X > \circ) = P(X - k > -k) = P(X - k < k) = G_{(X-k)}(k)$$

داریم

$$f_X(x) = \frac{1}{2} \frac{g(|x|)}{G_{(X-k)}(k)}.$$

اثبات گزاره ۱.۳.۳: برای اثبات گزاره ۱.۳.۳ باید نقاط بحرانی (۱.۳) را به دست بیاوریم. یکی از نقاط بحرانی برابر $x = \mu$ است. چون در این نقطه مشتق وجود ندارد. $x = \mu$ زمانی مد (۱.۳) است که علامت مشتق (۱.۳) برای $x < \mu$ مثبت و برای $x > \mu$ منفی باشد. مشتق تابع چگالی (۱.۳) متناسب عبارت زیر است:

$$k\sigma \text{sign}(x - \mu) - x + \mu + a$$

که برای $x < \mu$ مثبت است اگر

$$k\sigma \text{sign}(x - \mu) - x + \mu + a > \circ.$$

بنابراین داریم

$$-k\sigma + \mu + a > x$$

که در صورتی برقرار است که داشته باشیم

$$-k\sigma + a \geq \circ \longrightarrow k\sigma \leq a. \quad (4-آ)$$

با استدلال مشابه می توان نشان داد که مشتق (۱.۳) برای $x > \mu$ منفی است، اگر داشته باشیم

$$k\sigma + a \leq \circ \longrightarrow k\sigma \leq -a.$$

بنابراین در صورتی که شرایط (۱.۳) و (۱.۳) همزمان برقرار باشند، یعنی داشته باشیم

$$k\sigma \leq -|a| \quad (4-آ)$$

مد (۱.۳) برابر $x = \mu$ است. اکنون حالت های دیگر یعنی

$$k\sigma > -|a| \quad (3-آ)$$

را بررسی می کنیم. ابتدا نقاط بحرانی دیگر (۱.۳) را به دست می آوریم. برای این کار باید معادله زیر را حل کنیم:

$$k\sigma \text{sign}(x - \mu) - x + \mu + a = \circ.$$

داریم

$$\begin{cases} \frac{k}{\sigma} - \frac{x-\mu-a}{\sigma^2} & x > \mu \\ \frac{-k}{\sigma} - \frac{x-\mu-a}{\sigma^2} & x < \mu \end{cases}$$

و

$$\begin{cases} x = \mu + a + k\sigma & x > \mu \\ x = \mu + a - k\sigma & x < \mu \end{cases}$$

که شرط وجود مد $x = \mu + a + k\sigma$ این است که داشته باشیم $x > \mu$ که در صورتی شرط برقرار است که

$$k\sigma > -a. \quad (5-آ)$$

از اشتراک این شرط با (3-آ) داریم

$$-a < k\sigma < a.$$

همچنین شرط وجود مد $x = \mu + a - k\sigma$ برابر $x < \mu$ است که در صورتی برقرار است که داشته باشیم

$$k\sigma > a. \quad (5-آ)$$

که با اشتراک این شرط با (3-آ) داریم

$$a < k\sigma < -a.$$

بنابراین اگر دو شرط (5-آ) و (5-آ) همزمان برقرار باشند، یعنی $k\sigma > |a|$ ، دو مد $x = \mu + a \pm k\sigma$ را خواهیم داشت. این اثبات را کامل می‌کند.

اثبات گزاره ۲.۳.۳: تابع چگالی متغیر تصادفی X به صورت زیر است:

$$f(x) = c_{\sigma,k,a} \exp\left(k \left| \frac{x-\mu}{\sigma} \right| \right) \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma} - \frac{a}{\sigma}\right) \quad x \in \mathbb{R}$$

این تابع چگالی را می‌توان به صورت دوضابطه‌ای زیر نوشت.

$$f(x) = \begin{cases} c_{\sigma,k,a} \exp\left(-k \frac{x-\mu}{\sigma}\right) \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma} - \frac{a}{\sigma}\right) & x < \mu \\ c_{\sigma,k,a} \exp\left(k \frac{x-\mu}{\sigma}\right) \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma} - \frac{a}{\sigma}\right) & x > \mu \end{cases}$$

با محاسبات ساده به رابطه زیر می‌رسیم

$$f(z) = \begin{cases} \frac{\exp\left(-\frac{ka}{\sigma}\right) \Phi\left(k - \frac{a}{\sigma}\right)}{\exp\left(\frac{ka}{\sigma}\right) \Phi\left(k + \frac{a}{\sigma}\right) + \exp\left(-\frac{ka}{\sigma}\right) \Phi\left(k - \frac{a}{\sigma}\right)} \frac{\phi\left(\frac{x-\mu}{\sigma} + k - \frac{a}{\sigma}\right)}{\sigma \Phi\left(k - \frac{a}{\sigma}\right)} & x < \mu \\ \frac{\exp\left(\frac{ka}{\sigma}\right) \Phi\left(k + \frac{a}{\sigma}\right)}{\exp\left(\frac{ka}{\sigma}\right) \Phi\left(k + \frac{a}{\sigma}\right) + \exp\left(-\frac{ka}{\sigma}\right) \Phi\left(k - \frac{a}{\sigma}\right)} \frac{\phi\left(\frac{x-\mu}{\sigma} - k - \frac{a}{\sigma}\right)}{\sigma \Phi\left(k + \frac{a}{\sigma}\right)} & x > \mu \end{cases}$$

و این برابر است با

$$p \frac{\phi\left(\frac{x-\mu}{\sigma} + k - \frac{a}{\sigma}\right)}{\sigma \Phi\left(k - \frac{a}{\sigma}\right)} + (1-p) \frac{\phi\left(\frac{x-\mu}{\sigma} - k - \frac{a}{\sigma}\right)}{\sigma \Phi\left(k + \frac{a}{\sigma}\right)}$$

و اثبات کامل می‌شود.

اثبات گزاره ۳.۳.۳: اثبات گزاره ۳.۳.۳ مشابه اثبات گزاره ۴.۳.۳ است لذا از اثبات آن صرف‌نظر می‌کنیم.

اثبات گزاره ۴.۳.۳: داریم

$$\begin{aligned} \lim_{\nu \rightarrow \infty} s_- &= \lim_{\nu \rightarrow \infty} s_+ = \sigma^2 \\ \lim_{\nu \rightarrow \infty} s_-^{-\frac{\nu}{\sigma^2}} &= \exp\left(\frac{ak}{\sigma^2}\right) \\ \lim_{\nu \rightarrow \infty} s_+^{-\frac{\nu}{\sigma^2}} &= \exp\left(-\frac{ak}{\sigma^2}\right) \\ \lim_{\nu \rightarrow \infty} s_-^{-\frac{1}{\sigma^2}} d_\nu\left(\frac{x-\mu-(a+k)}{\sqrt{s_-}}\right) &= \frac{1}{\sigma} \phi\left(\frac{x-\mu-(a+k)}{\sigma}\right) \\ \lim_{\nu \rightarrow \infty} s_+^{-\frac{1}{\sigma^2}} d_\nu\left(\frac{x-\mu-(a-k)}{\sqrt{s_+}}\right) &= \frac{1}{\sigma} \phi\left(\frac{x-\mu-(a-k)}{\sigma}\right) \\ \lim_{\nu \rightarrow \infty} D_\nu\left(\frac{a+k}{\sqrt{s_-}}\right) &= \Phi\left(\frac{a+k}{\sigma}\right) \\ \lim_{\nu \rightarrow \infty} D_\nu\left(\frac{k-a}{\sqrt{s_+}}\right) &= \Phi\left(\frac{k-a}{\sigma}\right). \end{aligned}$$

بنابراین

$$\begin{aligned} & \lim_{\nu \rightarrow \infty} \frac{\left\{ s_+^{-\frac{\nu+1}{\sigma^2}} d_\nu\left(\frac{x-\mu-(a-k)}{\sqrt{s_+}}\right) \right\}^{I(x<\mu)} \left\{ s_-^{-\frac{\nu+1}{\sigma^2}} d_\nu\left(\frac{x-\mu-(a+k)}{\sqrt{s_-}}\right) \right\}^{I(x \geq \mu)}}{s_-^{-\frac{\nu}{\sigma^2}} D_\nu\left(\frac{a+k}{\sqrt{s_-}}\right) + s_+^{-\frac{\nu}{\sigma^2}} D_\nu\left(\frac{k-a}{\sqrt{s_+}}\right)} \\ &= \frac{\left\{ \exp\left(-\frac{ak}{\sigma^2}\right) \frac{1}{\sigma} \phi\left(\frac{x-\mu-(a-k)}{\sigma}\right) \right\}^{I(x<\mu)} \left\{ \exp\left(\frac{ak}{\sigma^2}\right) \frac{1}{\sigma} \phi\left(\frac{x-\mu-(a+k)}{\sigma}\right) \right\}^{I(x \geq \mu)}}{\exp\left(\frac{ak}{\sigma^2}\right) \Phi\left(\frac{a+k}{\sigma}\right) + \exp\left(-\frac{ak}{\sigma^2}\right) \Phi\left(\frac{k-a}{\sigma}\right)} \\ &= \begin{cases} \frac{\exp\left(-\frac{ka}{\sigma^2}\right) \Phi\left(\frac{k-a}{\sigma}\right)}{\exp\left(\frac{ka}{\sigma^2}\right) \Phi\left(\frac{k+a}{\sigma}\right) + \exp\left(-\frac{ka}{\sigma^2}\right) \Phi\left(\frac{k-a}{\sigma}\right)} \frac{\phi\left(\frac{x-\mu-(a-k)}{\sigma}\right)}{\sigma \Phi\left(\frac{k-a}{\sigma}\right)} & x < \mu \\ \frac{\exp\left(\frac{ka}{\sigma^2}\right) \Phi\left(\frac{k+a}{\sigma}\right)}{\exp\left(\frac{ka}{\sigma^2}\right) \Phi\left(\frac{k+a}{\sigma}\right) + \exp\left(-\frac{ka}{\sigma^2}\right) \Phi\left(\frac{k-a}{\sigma}\right)} \frac{\phi\left(\frac{x-\mu-(a+k)}{\sigma}\right)}{\sigma \Phi\left(\frac{k+a}{\sigma}\right)} & x \geq \mu \end{cases} \end{aligned}$$

که تساوی آخر طبق گزاره ۲.۳.۳ تابع چگالی $\text{BUN}(\mu, \sigma, \frac{k}{\sigma}, a)$ است.

اثبات گزاره ۵.۳.۳: اثبات گزاره ۵.۳.۳ مشابه اثبات گزاره ۲.۳.۳ است، لذا از اثبات آن صرف‌نظر می‌کنیم.

اثبات گزاره ۶.۳.۳: مشتق تابع چگالی (۱۴.۳) متناسب است با

$$(19-.\bar{A}) \quad 2t - k\sigma \left(1 + \frac{t^2}{\sigma^2}\right) \text{sign}(t + a)$$

که در آن $t = x - \mu - a$. در ادامه (۱۹-.\bar{A}) را مساوی صفر قرار داده و حل می‌کنیم. داریم

$$\begin{cases} t^2 - 2\frac{\sigma}{k}t + \sigma^2 = 0 & x > \mu \\ t^2 + 2\frac{\sigma}{k}t + \sigma^2 = 0 & x < \mu \end{cases}$$

دلتای هر دو قسمت معادله اخیر برابر است با $\Delta = \frac{4\sigma^2}{k^2} - 4\sigma^2$ که برای $k \leq 1$ شرط $\Delta \geq 0$ برقرار است. در ادامه داریم

$$(19-.\bar{A}) \quad \begin{cases} x = \mu + a + \sigma \left(\frac{1}{k} + \sqrt{\frac{1}{k^2} - 1}\right) & x > \mu \\ x = \mu + a + \sigma \left(\frac{1}{k} - \sqrt{\frac{1}{k^2} - 1}\right) & x > \mu \\ x = \mu + a - \sigma \left(\frac{1}{k} - \sqrt{\frac{1}{k^2} - 1}\right) & x < \mu \\ x = \mu + a - \sigma \left(\frac{1}{k} + \sqrt{\frac{1}{k^2} - 1}\right) & x < \mu \end{cases}$$

با استفاده از آزمون مشتق دوم مشخص می‌کنیم که کدام یک از جواب‌های فوق ماکسیمم نسبی و کدام یک مینیمم نسبی هستند. مشتق دوم تابع چگالی (۱۴.۳) متناسب است با

$$\begin{cases} 2 - 2\frac{k}{\sigma}t & x > \mu \\ 2 + 2\frac{k}{\sigma}t & x < \mu \end{cases}$$

که با جای‌گذاری جواب‌های موجود در (۱۹-.\bar{A}) با علامت منفی (ماکسیمم نسبی) مشتق دوم به‌عنوان مد عبارتند از

$$\begin{cases} x = \mu + a + \sigma \left(\frac{1}{k} + \sqrt{\frac{1}{k^2} - 1}\right) & x > \mu \\ x = \mu + a - \sigma \left(\frac{1}{k} - \sqrt{\frac{1}{k^2} - 1}\right) & x < \mu \end{cases}$$

یعنی برای $k \leq 1$ مدهای توزیع برابر $\mu + a \pm \sigma \left(\frac{1}{k} - \sqrt{\frac{1}{k^2} - 1}\right)$ است.

۳.آ اثبات قضایای فصل پنجم

اثبات قضیه ۱.۱.۴: برای اثبات قضیه ۱.۱.۴ نیاز به بیان و اثبات دو لم زیر داریم.

لم ۱.۳.آ. اگر $\lim_{n \rightarrow \infty} n^{-1} X_n^T X_n = C$ ، ماتریس معین مثبت با مولفه‌های متناهی باشد، داریم

$$\lim_{n \rightarrow \infty} n^{-1} I_n(\theta) = I(\theta)$$

اثبات لم ۱.۳.آ: داریم

$$I(\theta) = -\lim_{n \rightarrow \infty} n^{-1} E \left(\frac{\partial^2 \ell}{\partial \theta \partial \theta^T} \right) = \begin{bmatrix} I(\beta) & I(\beta, \sigma) & I(\beta, k) & I(\beta, a) \\ I(\beta, \sigma)^T & I(\sigma) & I(\sigma, k) & I(\sigma, a) \\ I(\beta, k)^T & I(\sigma, k) & I(k) & I(k, a) \\ I(\beta, a)^T & I(\sigma, a) & I(k, a) & I(a) \end{bmatrix}$$

که در آن

$$\begin{aligned} I(\beta) &= \frac{1}{\sigma^2} C \\ I(\beta, \sigma) &= \frac{ks}{\sigma^2} (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)^T \\ I(\beta, k) &= \frac{s}{\sigma} (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)^T \\ I(\beta, a) &= \frac{1}{\sigma^2} (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)^T \\ I(\sigma) &= \frac{1}{\sigma^2} (\gamma + k\gamma + k\rho) - \frac{ka\gamma}{\sigma^2} (ks\gamma - k - \rho) \\ I(\sigma, k) &= \frac{1}{\sigma} (k + \rho) - \frac{ka}{\sigma^2} \left(\frac{a}{\sigma\delta} - \frac{as\gamma}{\sigma} + \rho s \right) \\ I(\sigma, a) &= \frac{ks}{\sigma^2} - \frac{ka}{\sigma^2} \left(\frac{k}{\sigma} + \frac{\rho}{\delta} - \frac{ks\gamma}{\sigma} \right) \\ I(k) &= 1 + \frac{a}{\sigma} \left(\frac{a}{\sigma\delta} + \frac{as\gamma}{\sigma} \right) - \rho(k + \rho) \\ I(k, a) &= \frac{1}{\sigma} \left(s + a \left(\frac{k}{\sigma} + \frac{\rho}{\delta} - \frac{ks\gamma}{\sigma} \right) \right) - \rho \left(\frac{a}{\sigma^2} + \frac{ks}{\sigma} \right) \\ I(a) &= \frac{1}{\sigma^2} + \frac{k}{\sigma} \left(\frac{k}{\sigma} + \frac{\rho}{\delta} - \frac{ks\gamma}{\sigma} \right). \end{aligned}$$

لم ۲.۳.آ. اگر $\lim_{n \rightarrow \infty} n^{-1} X_n^T X_n = C$ ، ماتریس معین مثبت با مولفه‌های متناهی باشد داریم

$$-n^{-1} \frac{\partial^2 \ell}{\partial \theta \partial \theta^T} \xrightarrow{p} I(\theta)$$

اثبات لم ۲.۳.آ: داریم

$$-\lim_{n \rightarrow \infty} n^{-1} \frac{\partial^2 \ell}{\partial \theta \partial \theta^T} = -\lim_{n \rightarrow \infty} n^{-1} \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta \partial \beta^T} & \left(\frac{\partial^2 \ell}{\partial \beta \partial \sigma} \right) & \left(\frac{\partial^2 \ell}{\partial \beta \partial k} \right) & \left(\frac{\partial^2 \ell}{\partial \beta \partial a} \right) \\ \left(\frac{\partial^2 \ell}{\partial \beta \partial \sigma} \right)^T & \frac{\partial^2 \ell}{\partial \sigma^2} & \frac{\partial^2 \ell}{\partial \sigma \partial k} & \frac{\partial^2 \ell}{\partial \sigma \partial a} \\ \left(\frac{\partial^2 \ell}{\partial \beta \partial k} \right)^T & \frac{\partial^2 \ell}{\partial \sigma \partial k} & \frac{\partial^2 \ell}{\partial k^2} & \frac{\partial^2 \ell}{\partial k \partial a} \\ \left(\frac{\partial^2 \ell}{\partial \beta \partial a} \right)^T & \frac{\partial^2 \ell}{\partial \sigma \partial a} & \frac{\partial^2 \ell}{\partial k \partial a} & \frac{\partial^2 \ell}{\partial a^2} \end{bmatrix}$$

$$-\lim_{n \rightarrow \infty} n^{-1} \frac{\partial^2 \ell}{\partial \beta^T \beta} = \frac{1}{\sigma^2} \lim_{n \rightarrow \infty} n^{-1} X_n^T X_n = \frac{1}{\sigma^2} C = I(\beta).$$

بنابراین

$$n^{-1} \frac{\partial^2 \ell}{\partial \beta^T \beta} \xrightarrow{p} I(\beta)$$

با تکنیک مشابه می‌توان نشان داد که

$$-n^{-1} \frac{\partial^2 \ell}{\partial \beta \partial a} \xrightarrow{p} I(\beta, a)$$

$$-n^{-1} \frac{\partial^2 \ell}{\partial k^2} \xrightarrow{p} I(k)$$

$$-n^{-1} \frac{\partial^2 \ell}{\partial k \partial a} \xrightarrow{p} I(k, a)$$

$$-n^{-1} \frac{\partial^2 \ell}{\partial a^2} \xrightarrow{p} I(a)$$

و

$$\begin{aligned} -\lim_{n \rightarrow \infty} n^{-1} \frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} &= -\lim_{n \rightarrow \infty} n^{-1} \left(\frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} - E \left(\frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} \right) + E \left(\frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} \right) \right) \\ &= -\lim_{n \rightarrow \infty} n^{-1} \left(\frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} - E \left(\frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} \right) \right) - \lim_{n \rightarrow \infty} n^{-1} E \left(\frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} \right) \end{aligned}$$

و داریم

$$\begin{aligned} &\lim_{n \rightarrow \infty} E \left[\left\{ n^{-1} \left(\frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} - E \left(\frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} \right) \right) \right\}^2 \right] \\ &= \lim_{n \rightarrow \infty} n^{-2} \text{var} \left[\frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} \right] \\ &= \lim_{n \rightarrow \infty} n^{-2} \text{var} \left[\frac{k}{\sigma^2} \sum_{i=1}^n x_{ij} \text{sign}(Z_i) - \frac{2}{\sigma^2} \sum_{i=1}^n x_{ij} \left(\frac{Z_i - a}{\sigma} \right) \right]. \end{aligned}$$

متغیرهای تصادفی $Z_i = Y_i - \beta^T x_i$ برای $i = 1, 2, \dots, n$ مستقل و هم‌توزیع هستند. بنابراین

$$\begin{aligned} &\lim_{n \rightarrow \infty} n^{-2} \text{var} \left[\frac{k}{\sigma^2} \sum_{i=1}^n x_{ij} \text{sign}(Z_i) - \frac{2}{\sigma^2} \sum_{i=1}^n x_{ij} \left(\frac{Z_i - a}{\sigma} \right) \right] \\ &= \lim_{n \rightarrow \infty} n^{-2} \text{var} \left[\frac{k}{\sigma^2} \sum_{i=1}^n x_{ij} \text{sign}(Z) - \frac{2}{\sigma^2} \sum_{i=1}^n x_{ij} \left(\frac{Z - a}{\sigma} \right) \right] \\ &= \lim_{n \rightarrow \infty} n^{-2} \text{var} \left[\frac{k}{\sigma^2} \sum_{i=1}^n x_{ij} \text{sign}(Z) \right] - \lim_{n \rightarrow \infty} n^{-2} \text{var} \left[\frac{2}{\sigma^2} \sum_{i=1}^n x_{ij} \left(\frac{Z - a}{\sigma} \right) \right] \\ &= \frac{k^2}{\sigma^4} \text{var} [\text{sign}(Z)] \lim_{n \rightarrow \infty} n^{-2} \sum_{i=1}^n x_{ij}^2 - \frac{4}{\sigma^4} \text{var} \left[\left(\frac{Z - a}{\sigma} \right) \right] \lim_{n \rightarrow \infty} n^{-2} \sum_{i=1}^n x_{ij}^2 = 0. \end{aligned}$$

داریم

$$-\lim_{n \rightarrow \infty} n^{-1} \frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} = o_n(1) - \lim_{n \rightarrow \infty} n^{-1} E \left(\frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} \right).$$

یعنی

$$-n^{-1} \frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} \xrightarrow{p} -\lim_{n \rightarrow \infty} n^{-1} E \left(\frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} \right) = I(\beta_j, \sigma)$$

که در آن $j-I(\beta_j, \sigma)$ امین مولفه $I(\beta, \sigma)$ است. براساس این

$$-n^{-1} \frac{\partial^2 \ell}{\partial \beta \partial \sigma} \xrightarrow{p} I(\beta, \sigma).$$

با تکنیک مشابه می توان نشان داد

$$-n^{-1} \frac{\partial^2 \ell}{\partial \beta \partial k} \xrightarrow{p} I(\beta, k)$$

$$-n^{-1} \frac{\partial^2 \ell}{\partial \sigma^2} \xrightarrow{p} I(\sigma)$$

$$-n^{-1} \frac{\partial^2 \ell}{\partial \sigma \partial k} \xrightarrow{p} I(\sigma, k)$$

$$-n^{-1} \frac{\partial^2 \ell}{\partial \sigma \partial a} \xrightarrow{p} I(\sigma, a).$$

بنابراین

$$-n^{-1} L^{(2)} \xrightarrow{p} \lim_{n \rightarrow \infty} n^{-1} I_n(\theta) = I(\theta).$$

اکنون با استفاده از لم ۱.۳.آ و ۲.۳.آ، اثبات قضیه ۱.۱.۴ را شروع می کنیم. همان طور که در اثبات قضیه ۲.۱.۲ بر اساس سوییپینگ (۱۹۸۰) و ماردیا و مارشال (۱۹۸۴) گفته شد، نتیجه عمومی برای سازگاری و توزیع مجانبی نرمال برآوردگرهای درستنمایی ماکسیمم مستلزم برقراری سه شرط نظم است: (۱) پیوستگی، (۲) رشد و (۳) همگرایی ماتریس اطلاع فیشر. شرط پیوستگی با دو بار مشتق پذیری لگاریتم تابع درستنمایی در فضای پارامتر و مشتق دوم پیوسته آن در مدل پیشنهادی برقرار است. شرط رشد ماتریس اطلاع فیشر نیز برقرار است. اگر برای $n \rightarrow \infty$ داشته باشیم $\lim_{n \rightarrow \infty} I_n(\theta)^{-1} = 0$ که این شرط بنابر لم ۱.۳.آ برقرار است. شرط سوم یعنی همگرایی ماتریس اطلاع فیشر مشاهده شده یعنی $L_n^{(2)}$ برقرار خواهد شد، هرگاه $-I_n(\theta)^{-\frac{1}{2}} L_n^{(2)} I_n(\theta)^{-\frac{1}{2}}$ در احتمال به ماتریس همانی میل کند. با استفاده از لم ۲.۳.آ داریم $-n^{-1} L^{(2)} \rightarrow I(\theta)$ و همچنین $-n^{-1} I_n(\theta) \rightarrow I(\theta)$. بنابراین

$$\begin{aligned} \lim_{n \rightarrow \infty} -I_n(\theta)^{-\frac{1}{2}} L_n^{(2)} I_n(\theta)^{-\frac{1}{2}} &= \lim_{n \rightarrow \infty} n^{\frac{1}{2}} I_n(\theta)^{-\frac{1}{2}} \left(-n^{-1} L_n^{(2)} \right) n^{\frac{1}{2}} I_n(\theta)^{-\frac{1}{2}} \\ &= \lim_{n \rightarrow \infty} I(\theta)^{-\frac{1}{2}} I(\theta) I(\theta)^{-\frac{1}{2}} = I \end{aligned}$$

که این اثبات را کامل می کند.

مراجع

- [1] Aarset, M. V. (1987). How to identify a bathtub hazard rate. *IEEE Transactions on Reliability*, 36(1), 106-108.
- [2] Aas, K. & Haff, I. H. (2006). The generalized hyperbolic skew student's t-distribution. *Journal of Financial Econometrics*, 4(2), 275-309.
- [3] Alzaatreh, A., Famoye, F. & Lee, C. (2013). Weibull-Pareto distribution and its applications. *Communications in Statistics-Theory and Methods*, 42(9), 1673-1691.
- [4] Anderson, D. N. (1992). A multivariate Linnik distribution. *Statistics and Probability Letters*, 14(4), 333-336.
- [5] Arellano-Valle, R. B. & Azzalini, A. (2013). The centred parameterization and related quantities of the skew-t distribution. *Journal of Multivariate Analysis*, 113, 73-90.
- [6] Arellano-Valle, R. B., Bolfarine, H. & Lachos, V. H. (2005). Skew-normal linear mixed models. *Journal of Data Science*, 3(4), 415-438.
- [7] Arnold, B. C. & Lin, G. D. (2004). Characterizations of the skew-normal and generalized chi distributions. *Sankhyā: The Indian Journal of Statistics*, 66(4), 593-606.
- [8] Aryal, G. & Nadarajah, S. (2005). On the skew Laplace distribution. *Journal of Information and Optimization Sciences*, 26(1), 205-217.
- [9] Asgharzadeh, A., Esmacili, L., Nadarajah, S. & Shih, S. H. (2013). A generalized skew logistic distribution. *REVSTAT-Statistical Journal*, 11(3), 317-338.
- [10] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2), 171-178.

- [11] Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32(2), 159-188.
- [12] Azzalini, A. & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2), 367-389.
- [13] Azzalini, A. & Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika*, 83(4), 715-726.
- [14] Bakouch, H. S., Khan, M. N., Hussain, T. & Chesneau, C. (2019). A power log-Dagum distribution: estimation and applications. *Journal of Applied Statistics*, 46(5), 874-892.
- [15] Balakrishnan, N., & Lai, C. D. (2009). *Continuous Bivariate Distributions*, New York: Springer Science & Business Media.
- [16] Balakrishnan, N. & Ristić, M. M. (2016). Multivariate families of gamma-generated distributions with finite or infinite support above or below the diagonal. *Journal of Multivariate Analysis*, 143, 194-207.
- [17] Bartolucci, F. & Scaccia, L. (2005). The use of mixtures for dealing with non-normal regression errors. *Computational Statistics & Data Analysis*, 48(4), 821-834.
- [18] Bolfarine, H. & Lachos, V. H. (2007). Skew-probit measurement error models. *Statistical Methodology*, 4(1), 1-12.
- [19] Bolfarine, H., Martínez-Flórez, G. & Salinas, H. S. (2018). Bimodal symmetric-asymmetric power-normal families. *Communications in Statistics-Theory and Methods*, 47(2), 259-276.
- [20] Breslow N. E. & Clayton D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.
- [21] Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations, *IMA Journal of Applied Mathematics*, 6 , 76-90.
- [22] Butler, R. J., McDonald, J. B., Nelson, R. D. & White, S. B. (1990). Robust and partially adaptive estimation of regression models. *The Review of Economics and Statistics*, 72(2), 321-327.

- [23] Buuren S. & Fredriks M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, 20, 1259-1277.
- [24] Cannon, A., Cobb, G., Hartlaub, B., Legler, J., Lock, R., Moore, T., Rossman, A. & Witmer, J. (2019). *Stat2Data: Datasets for Stat2*, R package version 2.0.0.
- [25] Chu, S. (2001). Pricing the C's of Diamond Stones. *Journal of Statistics Education*, 9(2).
- [26] Cole T. J. and Green P. J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine*, 11, 1305-1319.
- [27] Cordeiro, G. M., Alizadeh, M. & Ortega, E. M. (2014). The exponentiated half-logistic family of distributions: Properties and applications. *Journal of Probability and Statistics*, 2014, 21.
- [28] Cordeiro, G. M., Alizadeh, M., Ozel, G., Hosseini, B., Ortega, E. M. M. & Altun, E. (2017). The generalized odd log-logistic family of distributions: properties, regression models and applications. *Journal of Statistical Computation and Simulation*, 87(5), 908-932.
- [29] Cordeiro, G. M. & de Castro, M. (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation*, 81(7), 883-898.
- [30] Cook, R. D., & Weisberg, S. (2009). *An Introduction to Regression Graphics (Vol. 405)*, New York: John Wiley & Sons.
- [31] Cressie, N. (1993). *Statistics for Spatial Data*, Revised Edition, New York: John Wiley & Sons.
- [32] da Silva, N. B., Prates, M. O. & Gonçalves, F. B. (2020). Bayesian linear regression models with flexible error distributions. *Journal of Statistical Computation and Simulation*, 90(14), 2571-2591.
- [33] Di, G. Q., Chen, X. W., Song, K., Zhou, B. & Pei, C. M. (2016). Improvement of Zwicker's psychoacoustic annoyance model aiming at tonal noises. *Applied Acoustics*, 105, 164-170.
- [34] DiCiccio, T. J. & Monti, A. C. (2004). Inferential aspects of the skew exponential power distribution. *Journal of the American Statistical Association*, 99(466), 439-450.

-
- [35] Dodd E. L. (1925). The frequency law of a function of variables with given frequency laws. *Annals of Mathematics*, 27, 12-20.
 - [36] Duarte, G. V., Braga, A., Miquelluti, D. L. & Ozaki, V. A. (2018). Modeling of soybean yield using symmetric, asymmetric and bimodal distributions: implications for crop insurance. *Journal of Applied Statistics*, 45(11), 1920-1937.
 - [37] Douglas A. & Delampady M. (1990). Eastern lake survey-phase I: documentation for the data base and the derived data sets. *SIMS Technical Report 160*.
 - [38] Draper, N. R. & Smith, H. (1998). *Applied regression analysis*, New York: John Wiley & Sons.
 - [39] Eilers P. H. C. & Marx B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 11, 89-121.
 - [40] Eltoft, T., Kim, T. & Lee, T. W. (2006). On the multivariate Laplace distribution. *IEEE Signal Processing Letters*, 13(5), 300-303.
 - [41] Eugene, N., Lee, C. & Famoye, F. (2002). Beta-normal distribution and its applications. *Communications in Statistics-Theory and methods*, 31(4), 497-512.
 - [42] Fallah, B. & Sodoudi, S. (2015). Bimodality and regime behavior in atmosphere-ocean interactions during the recent climate change. *Dynamics of Atmospheres and Oceans*, 70, 1-11.
 - [43] Fahrmeir L. & Lang S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50, 201-220.
 - [44] Fernández, C. & Steel, M. F. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441), 359-371.
 - [45] Ferreira, C. S. & Lachos, V. H. (2016). Nonlinear regression models under skew scale mixtures of normal distributions. *Statistical Methodology*, 33, 131-146.
 - [46] Ferreira, J. T. S. & Steel, M. F. J. (2006). A constructive representation of univariate skewed distributions. *Journal of the American Statistical Association*, 101(474), 823-829.

- [47] Fisher R. A. (1921). 014: On the "Probable Error. of a coefficient of correlation deduced from a small Sample. *Metron*, 1, 3-32.
- [48] Fischer, M. J. & Vaughan, D. (2002). Classes of skew generalized hyperbolic secant distributions. Diskussionspapiere//Friedrich-Alexander-Universitt Erlangen-Nrnberg , Lehrstuhl fr Statistik und konometrie, 2002.
- [49] Fischer M. J. (2013). *Generalized Hyperbolic Secant Distributions: With Applications to Finance*, Heidelberg: Springer Science & Business Media.
- [50] Fischer, M. J. & Vaughan, D. (2010). The beta-hyperbolic secant distribution. *Austrian Journal of Statistics*, 39(3), 245-258.
- [51] Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal*, 13(3), 317-322.
- [52] Gelfand, A. E., Kottas, A. & MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471), 1021-1035.
- [53] Genton, M. G. (Ed.). (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, London: CRC Press.
- [54] Gleaton, J. U., & Lynch, J. D. (2010). Extended generalized log-logistic families of lifetime distributions with an application. *Journal of Probability and Statistics*, 8(1), 1-17.
- [55] Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109), 23-26.
- [56] Graves, S. (2019). *Ecdat: Data sets for econometrics*, R package version 0.3-3.
- [57] Griffin, J. E. & Steel, M. J. (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101(473), 179-194.
- [58] Gupta, R. C., Gupta, P. L. & Gupta, R. D. (1998). Modeling failure time data by Lehman alternatives. *Communications in Statistics-Theory and Methods*, 27(4), 887-904.
- [59] Harandi, S. S. & Alamatsaz, M. H. (2013). Alpha-Skew-Laplace distribution. *Statistics & Probability Letters*, 83(3), 774-782.

-
- [60] Hastie T. & Tibshirani R. (1990). Generalized additive models. *Statistical Science*, 1, 297-310.
- [61] Hosseini, F., Eidsvik, J. & Mohammadzadeh, M. (2011). Approximate Bayesian inference in spatial GLMM with skew normal latent variables. *Computational Statistics & Data Analysis*, 55(4), 1791-1806.
- [62] Hosseini, F. & Mohammadzadeh, M. (2012). Bayesian prediction for spatial generalised linear mixed models with closed skew normal latent variables. *Australian & New Zealand Journal of Statistics*, 54(1), 43-62.
- [63] Huang, W. J., Su, N. C. & Teng, H. Y. (2019). On some study of skew-t distributions. *Communications in Statistics-Theory and Methods*, 48(19), 4712-4729.
- [64] Huber, P. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1), 73-101.
- [65] Joe, H., Seshadri, V. & Arnold, B. C. (2012). Multivariate inverse Gaussian and skew-normal densities. *Statistics & Probability Letters*, 82(12), 2244-2251.
- [66] Jones, M. C. & Faddy, M. J. (2003). A skew extension of the t-distribution, with applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1), 159-174.
- [67] Karimi, O., Omre, H. & Mohammadzadeh, M. (2010). Bayesian closed-skew Gaussian inversion of seismic AVO data for elastic material properties. *Geophysics*, 75(1), 185-198.
- [68] Kim, H. J. (2008). Moments of truncated Student-t distribution. *Journal of the Korean Statistical Society*, 37(1), 81-87.
- [69] Koenker, R. & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 46(1), 33-50.
- [70] Koenker, R. & Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448), 1296-1310.
- [71] Lange, K. L., Little, R. J. & Taylor, J. M. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408), 881-896.

- [72] Lee Y. & Nelder J. (2001). Hierarchical generalized linear models: a synthesis of generalized linear models random-effect models and structured dispersions. *Biometrika*, 88, 987-1006.
- [73] Lin X. & Zhang D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61, 381-400.
- [74] Ma, C., Li, Q., Liu, Q., Wang, D., Gao, J., Tang, H. & Sun, Y. (2016). Sound quality evaluation of noise of hub permanent-magnet synchronous motors for electric vehicles. *IEEE Transactions on Industrial Electronics*, 63(9), 5663-5673.
- [75] Ma, Y. & Genton, M. G. (2004). Flexible class of skew-symmetric distributions. *Scandinavian Journal of Statistics*, 31(3), 459-468.
- [76] Malik, H. J. & Abraham, B. (1973). Multivariate logistic distributions. *The Annals of Statistics*, 1(3), 588-590.
- [77] Maronna, R. A., Martin, R. D. & Yohai, V. J. (2019). *Robust Statistics: Theory and Methods (with R)*, New York: John Wiley & Sons.
- [78] Nadarajah, S., Cancho, V. G. & Ortega, E. M. (2013). The geometric exponential Poisson distribution. *Statistical Methods & Applications*, 22(3), 355-380.
- [79] Nadarajah, S. & Kotz, S. (2006). Skew distributions generated from different families. *Acta Applicandae Mathematica*, 91(1), 1.
- [80] Nadarajah, S., Nassiri, V. & Mohammadpour, A. (2014). Truncated-exponential skew-symmetric distributions. *Statistics*, 48(4), 872-895.
- [81] Nekoukhou, V. & Alamatsaz, M. H. (2012). A family of skew-symmetric-Laplace distributions. *Statistical Papers*, 53(3), 685-696.
- [82] Nelder J. & Wedderburn R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135, 370-384.
- [83] Omei, E., Van Gulck, S. & Vesilo, R. (2018). Semi-heavy tails. *Lithuanian Mathematical Journal*, 58(4), 480-499.
- [84] Osorio, F. (2019). *heavy: Robust estimation using heavy-tailed distributions*, R package version 0.38.196.

- [85] Papritz, A. & Schwierz, C. (2020) *georob: Robust geostatistical analysis of spatial data*, R package version 0.3-13.
- [86] Perks W. (1932). On some experiments in the graduation of mortality statistics. *Journal of the Institute of Actuaries*, 63, 12-57.
- [87] Ribeiro Jr, P. J. & Diggle, P. J. (2020) *geoR: analysis of geostatistical data*, R package version 1.8-1.
- [88] Rigby R. A. & Stasinopoulos D. M. (1996). A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, 6, 57-65.
- [89] Rigby R. A., Stasinopoulos M. D. and Voudouris V. (2013). Discussion: A comparison of GAMLSS with quantile regression. *Statistical Modelling*, 13, 335-348.
- [90] Rigby R. A. & Stasinopoulos M. D. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54, 507-554.
- [91] Ristić, M. M. & Nadarajah, S. (2014). A new lifetime distribution. *Journal of Statistical Computation and Simulation*, 84(1), 135-150.
- [92] Rubio, F. J. & Genton, M. G. (2016). Bayesian linear regression with skew-symmetric error distributions with applications to survival analysis. *Statistics in Medicine*, 35(14), 2441-2454.
- [93] Rue H. & Held L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. London: Chapman and Hall/CRC.
- [94] Sacks J., Welch W. J., Mitchell T. J. & Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical science*, 4, 409-423.
- [95] Sahu, S. K., Dey, D. K. & Branco, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics*, 31(2), 129-150.
- [96] Salazar, E., Ferreira, M. A. & Migon, H. S. (2012). Objective Bayesian analysis for exponential power regression models. *Sankhya B*, 74(1), 107-125.
- [97] Sastry, D. V. S. & Bhati, D. (2016). A new skew logistic distribution: Properties and applications. *Brazilian Journal of Probability and Statistics*, 30(2), 248-271.

- [98] Sato K. (1999). *Lévy Processes and Infinitely Divisible Distributions*, New York: Cambridge University Press.
- [99] Shah, S., Hazarika, P. J. & Chakraborty, S. (2019). The Balakrishnan Alpha skew Laplace distribution: properties and its applications. *arXiv preprint arXiv:1910.01084*.
- [100] Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111), 647-656.
- [101] Soffritti, G. & Galimberti, G. (2011). Multivariate linear regression with non-normal errors: a solution based on mixture models. *Statistics and Computing*, 21(4), 523-536.
- [102] Stasinopoulos, M. D. & Rigby, R. A. (2020) *gamlss.dist: Distributions for generalized additive models for location scale and shape*, R package version 5.1-7.
- [103] Stasinopoulos, M. D. & Rigby, B. (2020). *gamlss: Generalised additive models for location scale and shape*, R package version 5.2-0.
- [104] Stasinopoulos M. D., Rigby R. A., Heller G. Z., Voudouris V. & De Bastiani F. (2017). *Flexible Regression and Smoothing using GAMLSS in R*, London: Chapman and Hall/CRC.
- [105] Susanti, Y. & Pratiwi, H. (2014). M estimation, S estimation, and MM estimation in robust regression. *International Journal of Pure and Applied Mathematics*, 91(3), 349-360.
- [106] Sweeting, T. J. (1980). Uniform asymptotic normality of the maximum likelihood estimator, *The Annals of Statistics*. 1375-1381.
- [107] Takeuchi, I., Bengio, Y. & Kanamori, T. (2002). Robust regression with asymmetric heavy-tail noise distributions. *Neural Computation*, 14(10), 2469-2496.
- [108] Taylor, J. & Verbyla, A. (2004). Joint modelling of location and scale parameters of the t distribution. *Statistical Modelling*, 4(2), 91-112.
- [109] Team, R. C. (2020). R: A language and environment for statistical computing.
- [110] Vaughan, D. C. (2002). The generalized secant hyperbolic distribution and its properties. *Communications in Statistics-Theory and Methods*, 31(2), 219-238.

- [111] Vilca, F., Balakrishnan, N. & Zeller, C. B. (2014). Multivariate skew-normal generalized hyperbolic distribution and its properties. *Journal of Multivariate Analysis*, 128, 73-85.
- [112] Wood S.N. (2006). *Generalized Additive Models, An Introduction with R*, New York: Chapman and Hall.
- [113] Wang, Y., & Ke, C. (2020). *assist: A suite of R functions implementing spline smoothing techniques*, R package version 3.1.7.
- [114] Xu, Z. M., Xia, X. J., He, Y. S. & Zhang, Z. F. (2014). Analysis and evaluation of car engine starting sound quality. *Journal of Vibration and Shock*, 11, 142-147.
- [115] Yang, T., Gallagher, C. M. & McMahan, C. S. (2006). A robust regression methodology via M-estimation. *Computer Methods and Programs in Biomedicine*, 82(1), 31.
- [116] Yeh, H. C. (2010). Multivariate semi-logistic distributions. *Journal of Multivariate Analysis*, 101(4), 893-908.
- [117] Yilmaz, A. (2016). The flexible skew Laplace distribution. *Communications in Statistics-Theory and Methods*, 45(23), 7053-7059.
- [118] Zeller, C. B., Cabral, C. R. & Lachos, V. H. (2016). Robust mixture regression modeling based on scale mixtures of skew-normal distributions. *Test*, 25(2), 375-396.
- [119] Zhang, E., Zhang, Q., Xiao, J., Hou, L. & Guo, T. (2018). Acoustic comfort evaluation modeling and improvement test of a forklift based on rank score comparison and multiple linear regression. *Applied Acoustics*, 135, 29-36.

واژه‌نامه فارسی به انگلیسی

Bias	اریبی
Bimodal	دومدی
Bimodal–Unimodal Laplace	لاپلاس دومدی–یکمدی
Bimodal–Unimodal normal	نرمال دومدی–یکمدی
Bimodal–Unimodal student t	t استیودنت دومدی–یکمدی
Dispersion	پراکنش
Mixture	آمیخته
M–estimator	برآوردگرهای M
Generalized Additive Models	مدل‌های جمعی تعمیم‌یافته
Hyperbolic secant	هایپربولیک سکانت
Mean squared errors	میانگین توان‌های دوم خطا
Multivariate Bimodal–Unimodal normal	نرمال دومدی–یکمدی چندمتغیره
Robust	تنومند
Score vector	بردار امتیاز
Skew hyperbolic secant	هایپربولیک سکانت چوله
Skew Laplace	لاپلاس چوله
Skew Normal	نرمال چوله
Skew Student t	t استیودنت چوله
Spatial data	داده‌های فضایی
Standard deviation	خطای استاندارد
Stochastic representation	نمایش تصادفی
Truncated	بریده‌شده
Unimodal	یکمدی

واژه‌نامه انگلیسی به فارسی

Mixture.....	آمیخته
Bias.....	اریبی
M-estimator.....	برآوردگرهای M
Score vector.....	بردار امتیاز
Truncated.....	بریده‌شده
Dispersion.....	پراکنش
Robust.....	تنومند
Standard deviation.....	خطای استاندارد
Spatial data.....	داده‌های فضایی
Bimodal.....	دومدی
Skew laplace.....	لاپلاس چوله
Bimodal-Unimodal Laplace.....	لاپلاس دومدی-یکمدی
Generalized Additive Models.....	مدل‌های جمعی تعمیم‌یافته
Mean squared errors.....	میانگین توان‌های دوم خطا
Stochastic representation.....	نمایش تصادفی
Bimodal-Unimodal normal.....	نرمال دومدی-یکمدی
Multivariate Bimodal-Unimodal normal.....	نرمال دومدی-یکمدی چندمتغیره
Skew Normal.....	نرمال چوله
Hyperbolic secant.....	هایپربولیک سکانت
Skew hyperbolic secant.....	هایپربولیک سکانت چوله
Unimodal.....	یکمدی
Skew Student t.....	t استیودنت چوله
Bimodal-Unimodal student t.....	t استیودنت دومدی-یکمدی

Abstract

In practice, to analyze real data, assuming normality is not valid in many cases, and the models that can consider the possible real structure of the data should be used. For this purpose, we need flexible distributions for the analysis of non-normal data such as a thick tail, semi-heavy tail, and heavy tail skewed and bimodality. To this end, in this thesis, we introduce some probability distributions that are suitable for analyzing data with the mentioned properties. We first propose two approaches to develop symmetric distributions to the Bimodal–Unimodal symmetric and skewed distributions. Then we describe and evaluate their applications in the robust and spatial linear regression models. In the second part of the thesis, we introduce a new location–scale semi-parametric regression model constructed based on a semi-heavy tailed hyperbolic secant distribution. We develop skewed generalizations of the hyperbolic secant distribution and construct a linear regression model based on them. We also calculate and evaluate the asymptotic behavior of ML estimators of the introduced regression model. Finally, we introduce a multivariate version of the hyperbolic secant distribution.

Keywords: Bimodal–Unimodal distributions, Outliers, Skewed, Location–Scale regression, Spatial regression, Robust linear models, Modal regression.



Faculty Of Mathematical Sciences

PhD Thesis in Probability

Bimodal-Unimodal Distributions and Its Applications

By: Jamil Ownuk

Supervisors:

Dr. Ahmad Nezakati
Dr. Hossein Baghishani

February 2021