دانشکده علوم ریاضی

گروه آمار

پایان‌نامه کارشناسی ارشد

مدل‌بندی پاسخ‌های نرخ و نسبت با رگرسیون مستطیلی بتا

حانیه کیهانی

استاد راهنما

دکتر حسین باغیشنی

شهریور ۱۳۹۴

تقدیم به پدر و مادر عزیز و همسر مهربانم

نخستین سپاس و ستایش از آن پروردگاری است که آثار قدرت او بر چهره روز روشن، تابان است و انوار حکمت او در دل شب تار، درفشان. آفریدگاری که خویشتن را به ما شناساند و درهای علم را بر ما گشود و عمری و فرصتی عطا فرمود تا بدان، بنده ضعیف خویش را در طریق علم و معرفت بیازماید.

حال که در سایه‌سار رحمتش این پایان‌نامه به انجام رسیده است، بر خود لازم می‌دانم از استاد گرامیم جناب آقای دکتر باغیشنی که راهنمایی این پایان‌نامه را برعهده داشتند و بدون کمک‌های ایشان به انجام نمی‌رسید، مراتب سپاس را به‌جای آورم. همچنین از اساتید گران‌قدر داور، جناب آقای دکتر شاهسونی و دکتر آرشی، تشکر و قدردانی نمایم.

کمال سپاس از آن فرشتگان مهربانی است که مهر آسمانی‌شان آرام‌بخش آلام زمینی‌ام است، پدر و مادر بزرگوار و همسر عزیزم.

حانیه کیهانی

شهریور ۱۳۹۴

تعهدنامه

اینجانب **حانیه کیهانی** دانشجوی کارشناسی ارشد رشته **آمار** **دانشکده علوم ریاضی** دانشگاه شاهرود، نویسنده پایان‌نامه با عنوان **مدل‌بندی پاسخ‌های نرخ و نسبت با رگرسیون مستطیلی بتا** ، تحت راهنمایی **دکتر حسین باغیشنی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه شاهرود تعلق دارد، و مقالات مستخرج با نام " دانشگاه شاهرود " یا " Shahrood University " 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.

حانیه کیهانی
شهریور ۱۳۹۴

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 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 بسیاری از مدل‌های رگرسیونی ماهیت متغیر پاسخ به‌صورت نرخ و نسبت می‌باشد. معمولاً برای مدل‌بندی داده‌هایی با دامنه تغییرات $(۰,۱)$ از مدل رگرسیون بتا استفاده می‌شود، زیرا اگر متغیر پاسخ چوله باشد، مدل‌های لجستیک و پروبیت مناسب نیستند. لذا به‌عنوان جایگزینی رگرسیون بتا را برای این نوع از داده‌ها به کار می‌برند. اما مدل رگرسیون بتا نسبت به نقاط پرت تنومند نیست. پس به معرفی یک مدل جدید بر اساس توزیع مستطیلی بتا می‌پردازیم که نسبت به داده‌های دورافتاده تنومند است. در این پایان‌نامه پس از معرفی این توزیع و سپس مدل‌بندی آن، به استنباط بیزی می‌پردازیم و در استنباط بیزی از نمونه‌گیری مونت کارلوی زنجیر مارکوفی استفاده می‌کنیم. با شبیه‌سازی، تأثیر نقاط دور افتاده و کارایی مدل رگرسیون مستطیلی بتا را، بررسی کرده، برآورد ضرایب رگرسیون توزیع پسین پارامترها را به‌دست می‌آوریم. در آخر کاربرد مدل مستطیلی بتا را در قالب مثال کاربردی ارزیابی می‌کنیم.

کلمات کلیدی

تحلیل بیزی، رگرسیون مستطیلی بتا، تنومندی، $MCMC$، نقاط پرت.

پیش گفتار

بسیاری از روش‌های استنباط آماری به وجود داده‌های پرت[۱] حساس هستند. به‌عبارتی، معمولاً، داده‌های پرت منجر به بروز خطاهای افزوده در تحلیل‌های آماری می‌شوند. بنابراین آماردان‌ها، علاقه‌مند به استفاده از مدل‌ها و روش‌هایی هستند که نسبت به داده‌های پرت نیرومند باشند. یعنی وجود داده‌های پرت بر روی نتایج به‌دست آمده، تغییرات قابل ملاحظه‌ای ایجاد نکند. مدل رگرسیون بتا[۲] یک انتخاب متداول برای مدل‌بندی متغیرهای پاسخ پیوسته و محدود به فاصله $(۰,۱)$ می‌باشد. به‌عنوان مثال، پاسخ‌هایی که ماهیت نرخ، کسر، درصد یا نسبت دارند. مانند نرخ بیکاری و درصد یک بیماری خاص در یک منطقه، نسبت آرا برای رئیس‌جمهور وقت برای انتخاب دوباره از این جمله می‌باشند. استفاده از مدل رگرسیون بتا با مطالعه فراری و سریباری (۲۰۰۴) شروع شد. در مدلی که توسط این دو ارائه گردید، پراکندگی به‌عنوان پارامتر مزاحم در نظر گرفته شد. ورکوی‌لن و اسمیتسون (۲۰۰۶) مدل رگرسیون بتا را برای حالتی که پارامتر دقت[۳] مدل نیز تابعی از متغیرهای تبیینی است، تعمیم دادند. استنباط بیزی در مدل‌های رگرسیونی بتا، اولین بار توسط برانسکام و همکاران (۲۰۰۷) ارائه گردید. اما تاکنون، تا جایی که ما اطلاع داریم، مطالعه‌ای در مورد تأثیر داده‌های پرت بر برآورد پارامترهای مدل رگرسیون بتا انجام نشده است.

اخیراً، مدلی تحت عنوان رگرسیون مستطیلی بتا[۴]، که آمیخته‌ای از دو توزیع بتا می‌باشد، به‌عنوان یک مدل تنومند[۵] در این زمینه معرفی شده است. استفاده از مدل رگرسیونی مستطیلی بتا توسط هان (۲۰۰۸) پیشنهاد شد و تحلیل بیزی مدل‌های رگرسیونی مستطیلی بتا توسط بیز و همکاران (۲۰۱۲) ارائه گردید. برای داده‌های نرخ و نسبت، رخداد داده‌های پرت محتمل است. از طرفی، مطالعه‌ای برای تأثیر این نقاط بر روی مدل رگرسیونی بتا صورت نگرفته است. بنابراین بر آن شدیم تا تأثیر داده‌های پرت را در تحلیل داده‌های نرخ و نسبت بررسی کنیم. با توجه به ویژگی مدل رگرسیونی بتا و قدرت الگوریتم‌های نمونه‌گیری مونت کارلوی زنجیر مارکوفی[۶] $(MCMC)$، برای بررسی این تأثیر بر اساس کار بیز و همکاران (۲۰۱۲) علاقه‌مند شدیم از مدل رگرسیونی مستطیلی بتا در چارچوب استنباط بیزی[۷]، استفاده کنیم. با توجه به این مقدمه، ساختار این پایان‌نامه به‌صورت زیر است:

- در فصل اول، تعاریف و مفاهیم مورد نیاز برای ورود به موضوع این پایان‌نامه را مطرح می‌کنیم.
- در فصل دوم، مدل رگرسیون بتا و مستطیلی بتا معرفی شده‌اند.

---

[۱] Outlier Data

[۲] Beta Regression Model

[۳] Precision

[۴] Beta Rectangular Regression

[۵] Robust Model

[۶] Markov Chain Monte Carlo

[۷] Bayesian Inference

- با مطالعه شبیه‌سازی، عملکرد مدل و روش پیشنهادی را در فصل سوم مورد ارزیابی قرار می‌دهیم.
- در فصل چهارم، کاربرد مدل را با دو مثال واقعی شامل داده‌های مؤسسه ورزش استرلیا[۸] (AIS) و داده‌های مقاومت بتن[۹]($stre$) بررسی می‌کنیم.

[۸] Australian Institute of Sport

[۹] Strength

فهرست نشانه‌ها و نمادها

Y متغیر تصادفی
$\mathbf{Y}$ بردار تصادفی
Θ فضای پارامتر
$\pi(\theta)$ توزیع پیشین
$\pi(\theta|y)$ توزیع پسین
$\pi(y \mid \theta)$ تابع درستنمایی
$h_{\mathbf{Y}}(\mathbf{y})$ ثابت نرمالساز
$MCMC$ نمونه‌گیری مونت‌کارلوی زنجیر مارکوفی
GLM مدل تعمیم‌یافته خطی
$f(.)$ توزیع ابزاری
$JAGS$ منبع موتور جستجو زبان $BUGS$
$D(.)$ انحراف
$EAIC$ امید معیار اطلاع آکائیک
$EBIC$ امید معیار اطلاع بیزی
$D(\bar{\theta})$ میانگین انحراف پسین
$D(\bar{\theta})$ انحراف میانگین پسین
p_D پیچیدگی
DIC معیار اطلاع انحراف
MH متروپلیس‌هستینگز
$k(. \mid .)$ هسته انتقال مارکوف
$q(x \mid y)$ توزیع پیشنهادی
$\rho(x, y)$ احتمال پذیرش متروپلیس-هستینگز
$mod(.)$ مد
$D_{KL}(P \parallel Q)$ فاصله کولبک‌لیبلر Q از P
BR مستطیلی بتا
θ پارامتر آمیختگی
ϕ پارامتر پراکندگی
μ میانگین بتا

γ میانگین مستطیلی بتا
$logit(.)$ تابع پیوند لجیت
$log(.)$ تابع پیوند لگاریتمی
x_i, w_i متغیر تبیینی
MSE میانگین توان دوم خطا
Δ اختلال
$\vartheta(\eta^b)$ نمونه‌های زنجیر مارکوف مونت کارلو از توزیع پسین
KL کولبک-لیبلر
SN چوله نرمال
ST چوله-t

فهرست مطالب

لیست تصاویر

لیست جداول

فصل ۱

تعاریف و مفاهیم اولیه

۱.۱ مقدمه

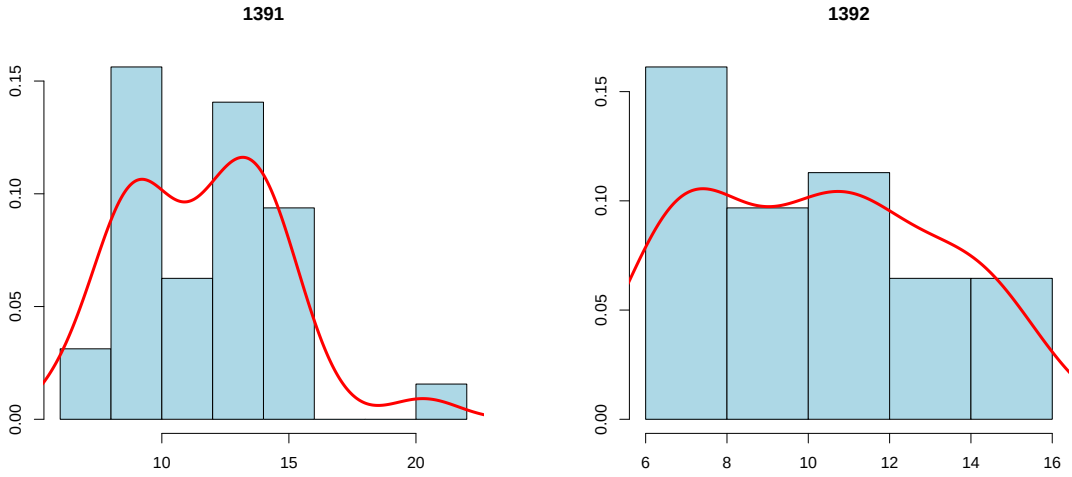
آمار علم جمع‌آوری و سازماندهی داده‌ها و استخراج نتایج است. سازماندهی و توصیف مشخصات عمومی مجموعه داده‌ها، از اهداف آمار توصیفی و چگونگی استخراج نتایج از داده‌ها، از اهداف استنباط آماری است. در این پایان‌نامه، در پی استنباط آماری داده‌ها هستیم و عناوین مرتبط دیگر برای درک مفاهیم اساسی توضیح داده می‌شود.

دامنه این تحلیل‌ها از تحلیل داده‌های یک تا چندمتغیره را در بر می‌گیرد. علم آمار روش‌های مختلفی را برای تعیین رابطه بین متغیرها پیشنهاد می‌کند که شامل انواع مختلف مدل‌های رگرسیونی است. کاربردهای رگرسیون متعدد هستند و تقریباً در هر زمینه‌ای از جمله مهندسی، فیزیک، اقتصاد، مدیریت، علوم زیستی، بیولوژی و علوم اجتماعی دیده می‌شوند. در واقع، تحلیل رگرسیونی را می‌توان پرکاربردترین روش تحلیل داده‌ها در آمار دانست. هدف تحلیل رگرسیون، تعیین بهترین مدلی است که بتواند ارتباط بین متغیر خروجی (متغیر پاسخ) Y با یک یا چند متغیر ورودی یا پیشگو (متغیر تبیینی) X را تعیین کند. در مواردی که متغیر پاسخ توزیع نرمال ندارد یا میانگین پاسخ را نمی‌توان به‌صورت مستقیم، به‌عنوان تابعی از متغیرهای تبیینی، مدل‌سازی نمود، تبدیلی از آن با استفاده از تابع پیوند مورد استفاده قرار می‌گیرد. در این موارد مدلی که بر داده‌ها اعمال می‌شود، عضوی از مدل‌های خطی تعمیم‌یافته [۱] $(GLMs)$ است (جکمن، ۱۹۸۹). مدل رگرسیون خطی حالت خاصی از این مدل‌ها است (مکلا و نلدر، ۱۹۸۹).

در بسیاری از موارد، متغیر پاسخ یک متغیر پیوسته با تغییرات در بازه $(۰, ۱)$ است. این نوع پاسخ‌ها در مواردی رخ می‌دهند که داده‌ها به‌صورت نرخ، کسر، نسبت یا درصد یک پدیده باشند. نرخ بیکاری را می‌توان به‌صورت

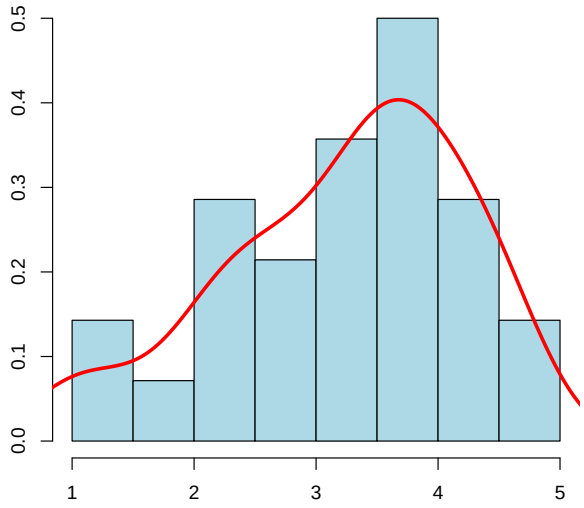
$$\text{نرخ بیکاری} = \frac{\text{جمعیت بیکار}}{\text{جمعیت فعال (بیکار و شاغل)}} \times ۱۰۰$$

[۱] Generalized Linear Models



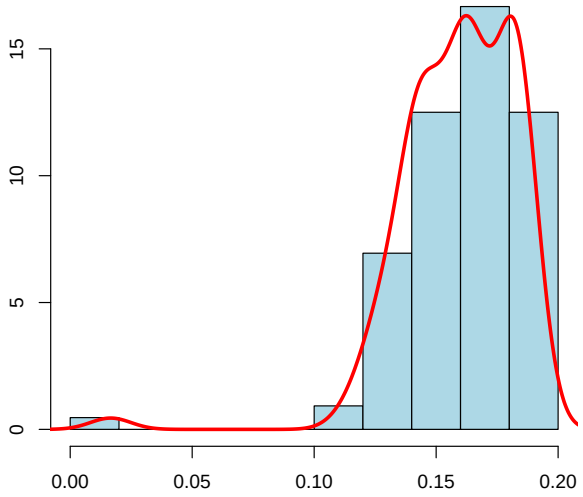
شکل ۱.۱: نرخ بیکاری در استان‌های ایران در سال‌های ۱۳۹۱ و ۱۳۹۲

محاسبه کرد. بر این اساس با توجه به نمودار ۱.۱ نرخ بیکاری در استان‌های مختلف کشور در سال ۱۳۹۱ بین ۱۰ تا ۱۵و در سال ۱۳۹۲ بین ۶ تا ۱۶ بیشترین فراوانی را دارند. همان‌طور که ملاحظه می‌کنید، این داده‌ها از نوع نرخ می‌باشند و در طی دو سال متوالی تقریباً مقادیر خاصی از بازه تغییرات خود را اختیار می‌کنند. نرخ بیکاری در این دو سال کمتر از ۵ و بیشتر از ۲۰ نمی‌باشد.



شکل ۲.۱: درصد سرطان مردان در استان‌های کشور در سال ۱۳۸۸

در صد سرطان مردان در استان‌های مختلف کشور را در شکل ۲.۱ و مقاومت بتن را در شکل ۳.۱ می‌بینید، این نمودارها بر اساس داده‌هایی از نوع درصد و کسر به‌دست آمده‌اند، با توجه به این شکل‌ها



شکل ۳.۱: مقاومت بتن

نیز درمی‌یابیم که درصد سرطان مردان در هیچ استانی در سال ۱۳۸۸ بیشتر از ۵ و کمتر از ۱ درصد نیست. این داده‌ها بیشترین تمرکز خود را معمولاً در بازه $(۲,۵)$ دارند. داده‌های مقاومت بتن نیز در بازه $(۰/۱,۰/۲)$ بیشترین فراوانی را دارند. مدل‌های رگرسیونی معمول برای تحلیل این نوع پاسخ‌ها، رگرسیون پروبیت[۲] و رگرسیون لجستیک[۳] هستند. اما از آن‌جا که داده‌های نرخ و نسبت، معمولاً در زیرفاصله مشخصی از دامنه تغییرات خود متمرکز هستند (به‌عبارتی توزیع این داده‌ها به شدت چوله است)، مدل‌های لجستیک و پروبیت برای مدل‌بندی آن‌ها مناسب نیستند، زیرا مدل رگرسیون لجستیک بر پایه توزیع لجستیک و رگرسیون پروبیت بر اساس توزیع نرمال هستند، این توزیع‌ها متقارن می‌باشند و برای داده‌های چوله مناسب نیستند. در این‌گونه موارد، مدل رگرسیونی بتا پیشنهاد می‌شود که نسبت به داده‌های چوله انعطاف زیادی دارد. (فراری و سریباری، ۲۰۰۴).

برازش مدل رگرسیونی بتا، بر اساس هر دو دیدگاه کلاسیک و بیزی قابل انجام است. روش‌های استنباط کلاسیک، معمولاً، روش‌های مبتنی بر درستنمایی هستند که برای محاسبه تابع درستنمایی در رده $GLMs$ به کار می‌روند. ماکسیمم‌سازی تابع درستنمایی حل عددی توابعی را شامل می‌شود، که می توانند چندمدی و ناهموار و در نتیجه بسیار پیچیده و زمان‌بر باشند. اگرچه روش‌های مختلفی برای بهینه‌سازی توابع پیچیده معرفی شده‌اند، اما هر کدام نیز دارای معایب و کاستی‌هایی می‌باشند. بنابراین برای کناره گرفتن از چنین مشکلاتی، در این پایان‌نامه، دیدگاه بیزی را برای استنباط در مدل‌های مورد نظر انتخاب کردیم. استنباط در دیدگاه بیزی، مبتنی بر توزیع پسین است و می‌تواند انتخابی مناسب برای استنباط مدل‌های پیشنهادی باشد. توزیع پسین در یک مدل بیزی، به روز شده باور و دیدگاه اولیه در مورد پارامتر با توجه به مشاهدات است. دیدگاه بیزی نشان‌دهنده راه و روشی است که اطلاعات خارجی را به شکلی

[۲] Probit Regression

[۳] Logistic

با تحلیل داده‌ها مرتبط کند. اولین مرحله در این روش، مشخص ساختن یک توزیع احتمالی برای فرآیند تحت مطالعه است. از آن‌جا که این توزیع پیش از هیچ تفکری در خصوص داده‌ها آماده می‌شود، به آن توزیع پیشین می‌گویند. ورود داده‌ها (اطلاعات جدید) توزیع پیشین را به توزیع پسین[۴] تبدیل می‌کند. ابزار اساسی برای این تبدیل قضیه بیز[۵] است (گلمن و همکاران، ۲۰۰۳). تحلیل بیزی نیاز به حل انتگرالی دارد که به‌واسطه آن توزیع پسین به‌دست می‌آید. معمولاً نمی‌توان این انتگرال را به‌راحتی به‌دست آورد. در این صورت می‌توان از شبیه‌سازی استفاده نمود. معمولاً نمونه‌گیری مستقیم از توزیع پسین نیز به سادگی صورت نمی‌پذیرد. پس برای حل این مشکل از روش‌های $MCMC$، بهره می‌گیریم. روش‌های $MCMC$ شامل روش‌هایی است که برای انتگرال‌گیری به روش مونت کارلویی، اغلب در مسائل مربوط به استنباط بیزی، مورد استفاده قرار می‌گیرند. $MCMC$ راه مؤثری است که از توزیع پسین نمونه‌هایی را تولید می‌کند (برانسکام و همکاران، ۲۰۰۷). از جمله روش‌های $MCMC$، الگوریتم متروپلیس-هستینگز[۶] (MH) (متروپلیس و همکاران، ۱۹۵۳ و هستینگز، ۱۹۷۰)، نمونه‌گیرگیبز[۷] (گمان و گمان، ۱۹۸۴ و گلفاند و اسمیت، ۱۹۹۰) و الگوریتم متروپلیس-هستینگز تحت گیبز[۸] (رابرت و کسلا، ۲۰۱۰) می‌باشند که به‌عنوان ابزار بسیار مفید برای تحلیل مدل‌های آماری پیچیده پدید آمده‌اند.

روش‌های معمول نمونه‌گیری $MCMC$، از جمله متروپلیس-هستینگز دارای مشکلاتی از جمله تشخیص همگرایی زنجیر و طولانی‌شدن زمان محاسبات می‌باشند. اما نمونه‌گیر گیبز این‌گونه مسائل را به دنباله‌ای از مسائل ساده‌تر تقسیم می‌کند و اجرای آن ساده‌تر است. البته ممکن است این دنباله‌ها زمان زیادی برای همگرایی نیاز داشته باشند ولی با این وجود، این الگوریتم مفید و قابل توجه است.

از سوی دیگر، در مباحث مرتبط با رگرسیون یا طرح آزمایش‌ها، با مفهومی روبرو می‌شویم که به آن داده پرت[۹] می‌گویند. در یک نمونه ممکن است تعدادی از داده‌ها نسبت به سایر مشاهدات انحراف زیادی داشته باشند. این انحراف به اندازه‌ای است که این تفکر را در ذهن افراد ایجاد می‌کند که شاید این داده‌ها از جامعه دیگری آمده باشند (کندال و بوکلاند، ۱۹۵۷).

با نظر در متون آماری به این نتیجه می‌رسیم که در دنیای واقعی، رخدادهایی در ناحیه دم توزیع با فراوانی‌های مختلف اتفاق می‌افتند. همین امر نشان می‌دهد که اغلب توزیع‌ها معمولاً قادر به نشان دادن دنیای واقعی نیستند (مکلا و نلدر، ۱۹۸۹؛ لوی و دوچین، ۲۰۰۴؛ مکفالم، ۲۰۰۳؛ کاتز و همکاران، ۲۰۰۱). برای مثال، در تحلیل داده‌های شمارشی با رگرسیون پواسون معمولاً داده‌ها نسبت به توزیع پواسن پراکندگی بیشتری دارند. زیرا در توزیع پواسن پراکندگی با مقدار متوسط برابر شده اما در داده‌های واقعی شمارشی معمولاً پراکندگی بزرگتر از متوسط است. در این مورد استفاده از توزیع پواسن منجر به متورم شدن آماره‌های آزمون و در نتیجه خطای نوع یک می‌شود (هان، ۲۰۰۸). در این صورت، برازش کارای مدل دچار مشکل می‌گردد. یک راه برای رفع این مشکل استفاده از توزیع‌های منعطف[۱۰] دیگری

[۴] Posterior Distribution
[۵] Bayes Theorem
[۶] Metropolis-Hastings Algoritm
[۷] Gibbs Sampler
[۸] MH-within-Gibbs
[۹] Outlier Data
[۱۰] Flexible

مانند آمیخته‌ای از توزیع t استیودنت و دوجمله‌ای منفی است که اغلب برای اصلاح این مشکلات و بیش‌پراکندگی[۱۱] است و در مشاهده نقاط پرت دقیق‌تر هستند.

در نتیجه برای بهبود شرایط، می‌توان از توزیع‌های آمیخته[۱۲] از جمله آمیخته‌ای از یک توزیع دم باریک مانند نرمال یا پواسن و توزیع دیگری که در محدوده دم‌ها پهن‌تر بوده و انعطاف و نیرومندی بیشتری نسبت به نقاط پرت نشان می‌دهد، استفاده کرد که این امر دقت برازش داده‌ها را بالا می‌برد (گلمن و همکاران، ۲۰۰۳). بررسی‌ها نشان می‌دهند مدل رگرسیون بتا نسبت به داده‌های پرت تنومند[۱۳] نیست. هان (۲۰۰۸) توزیعی با دم سنگین و تکیه‌گاه کراندار معرفی نمود. این توزیع، آمیخته‌ای از توزیع‌های یکنواخت[۱۴] و بتا است. وی به قوه سرایت پدیده‌های دم سنگین در مفاهیم اقتصادی اشاره کرد. این مدل در فصل‌های آتی با نام مدل رگرسیون مستطیلی بتا[۱۵] معرفی می‌گردد (بیز و همکاران، ۲۰۱۲). این مدل رگرسیونی در رده مدل‌های خطی تعمیم‌یافته قرار می‌گیرد. برای درک بهتر مدل مورد بحث، ابتدا مدل‌های رگرسیونی و مدل‌های خطی تعمیم‌یافته را معرفی می‌کنیم.

۲.۱ مدل‌های رگرسیونی

اولین بار موضوع رگرسیون[۱۶] توسط گالتن (۱۸۸۵) مطرح شد. وی قصد داشت رابطه‌ای بین قد والدین و قد پسران پیدا کند. به این صورت که والدین بلندقد تقریباً فرزندان بلندقد دارند و والدین کوتاه‌قد معمولاً فرزندانی با قدکوتاه دارند. وی، این مطلب را به‌صورت یک مدل درآورد. برای نشان دادن رابطه بین دو متغیر، ضریب همبستگی نیز استفاده می‌شود. اما ضریب همبستگی اثر یک متغیر بر متغیر دیگر را نشان نمی‌دهد. از طرفی، گاهی تخمین تغییر در یک متغیر با تغییر متغیر دیگر برای ما مهم است که باز هم ضریب همبستگی نمی‌تواند تخمینی از این تغییرات ارائه بدهد. برای پاسخ به موارد فوق باید به دنبال مدل‌های رگرسیونی رفت (گلمن و همکاران، ۲۰۰۶).

تعریف ۱.۲.۱. رگرسیون روشی برای به مدل درآوردن و بررسی ارتباط بین متغیرها است.

رگرسیون در حالت کلی با معادله زیر مشخص می‌شود:

$$Y = f(X_1, X_2, \cdots, X_k)$$

که در آن متغیری که از سایر متغیرها تأثیرپذیر است، متغیر پاسخ[۱۷] (متغیر وابسته) Y نامیده می‌شود. متغیر یا متغیرهایی که بر متغیر پاسخ اثر می‌گذارند، متغیر تبیینی[۱۸] $\mathbf{X} = (X_1, X_2, \cdots, X_k)^T$ نامیده

[۱۱] Overdispersion
[۱۲] Mixture Distributions
[۱۳] Robust Model
[۱۴] Uniform
[۱۵] Beta Rectangular Regression Model
[۱۶] Regression
[۱۷] Response
[۱۸] Predictor

می‌شود. برای مثال در رگرسیون خطی به دنبال رابطه خطی Y با متغیرهای $X_۱, X_۲, \cdots, X_k$ هستیم، یعنی:

$$Y = \beta_۰ + \beta_۱ X_۱ + \beta_۲ X_۲ + \cdots + \beta_k X_k$$

که در آن $\beta_۰, \beta_۱, \cdots, \beta_k$ ضرایب مدل رگرسیونی نامیده می‌شوند. این ضرایب نامعلوم بوده و باید برآورد شوند. در صورتی‌که y تابعی خطی از xها باشد، مدل رگرسیون خطی[۱۹] است. رگرسیون خطی، ساده‌ترین نوع رگرسیون است.

فرض کنید تنها یک متغیر تبیینی x داریم. هنگامی‌که برای یک مجموعه داده نمودار پراکنش y و x را رسم می‌کنیم، معمولاً داده‌های مشاهده‌شده دقیقاً روی یک خط راست قرار نمی‌گیرند. بنابراین، رابطه بالا بایستی به‌عنوان برآوردکننده خط رگرسیون، اصلاح و تعدیل شود. اگر اختلاف بین مقدار مشاهده‌شده y و خط راست را خطای ϵ قرار دهیم، خطای ϵ به‌عنوان یک خطای آماری تلقی می‌شود. به این معنی که، متغیری است تصادفی که عدم برازش را بیان و اندازه ناتوانی مدل در برازش دقیق داده‌ها را برآورد می‌کند. این خطا ممکن است به لحاظ اثرهای دیگر متغیرها، خطاهای اندازه‌گیری و غیره صورت پذیرد. بنابراین مدل بالا به صورت زیر بهبود می‌یابد:

$$Y = \beta_۰ + \beta_۱ X_۱ + \beta_۲ X_۲ + \cdots + \beta_k X_k + \epsilon = \mu + \epsilon$$

که در آن $\mu = \beta_۰ + \beta_۱ X_۱ + \beta_۲ X_۲ + \cdots + \beta_k X_k$ میانگین شرطی پاسخ به شرط مفروض بودن متغیرهای تبیینی است. در صورتی که در فرمول بالا یک متغیر تبیینی داشته باشیم، مدل رگرسیون خطی ساده و اگر چند متغیر تبیینی در مدل درگیر باشند، مدل رگرسیون خطی چندگانه است.

یکی از اهداف مهم تحلیل رگرسیونی، برآورد پارامترهای مدل است. روند برآورد پارامترها، برازش مدل به داده‌ها نیز نامیده می‌شود. یکی از شیوه‌های برازش مدل، روش میانگین توان‌های دوم خطا[۲۰] است.

مرحله دیگر تحلیل رگرسیونی، کنترل مناسبت مدل نامیده می‌شود، که در آن مناسب بودن مدل مورد مطالعه و کیفیت برازش مشخص می‌شود. طی انجام این مراحل، امکان قابل استفاده بودن و به‌کارگیری مدل رگرسیونی تعیین و در مورد آن تصمیم‌گیری می‌شود. نتیجه رسیدگی مناسبت مدل، این امکان را فراهم می‌کند که بتوان گفت، مدل مناسب است یا این‌که برازش اصلی بایستی اصلاح شود. لذا تحلیل رگرسیونی رویه‌ای همراه با تکرار و بازنگری است که در آن داده‌ها به سمت یک مدل رهنمون می‌شوند و یک برازش مدل به داده‌ها حاصل می‌شود. سپس، کیفیت برازش مورد رسیدگی واقع می‌شود. نتیجه کار منجر به اصلاح مجدد مدل یا برازش و قبول آن منتهی می‌شود.

مدل‌های رگرسیون برای اهداف زیر مورد استفاده قرار می‌گیرند که عبارتند از:

- توصیف داده‌ها

[۱۹] Linear Regression

[۲۰] Mean Square Error

- برآورد پارامترها
- پیشگویی
- کنترل

در مدل‌های رگرسیونی پذیره‌های بنیادی وجود دارند که از طریق خطاها (باقیمانده‌ها) قابل آزمون هستند. پذیره‌های اساسی رگرسیون عبارتند از:

- مؤلفه تصادفی خطا میانگین صفر داشته باشد. یعنی به ازای هر i امید ریاضی ϵ_i برابر با صفر باشد.
- واریانس مؤلفه خطا، ثابت باشد. یعنی به ازای هر i واریانس ϵ_i برابر با σ^2 باشد.
- خطاها ناهمبسته باشند.
- متغیرهای تبیینی X تحت کنترل ما باشند و با خطای ناچیزی اندازه‌گیری شوند.

هیچگاه نباید بدون بررسی پذیره‌های بنیادی، از مدل رگرسیونی استفاده نمود. در این حالت اطلاعات به‌دست آمده از رگرسیون می‌توانند گمراه‌کننده باشند. بنابراین باید در مورد مدل رگرسیونی حاصل استنباط کرد. استنباط آماری مدل را می‌توان در دو دیدگاه کلاسیک و بیزی انجام داد.

۳.۱ مدل‌های خطی تعمیم یافته

تحلیل داده‌های غیرنرمال با استفاده از مدل‌های غیرخطی پیشینه‌ای طولانی دارد. استفاده از رگرسیون پروبیت، برای پاسخ‌های دودویی، مثالی از این نوع است. خانواده مدل‌های خطی تعمیم‌یافته (نلدر و ودربرن، ۱۹۷۲) شامل دامنه گسترده‌ای از مدل‌ها، مانند رگرسیون معمولی و مدل‌های $ANOVA$ برای متغیرهای پاسخ پیوسته هستند و می‌توان آن‌ها را به مدل‌هایی با متغیرهای پاسخ گسسته تعمیم داد. GLM برای داده‌هایی است که معمولاً متغیر پاسخ توزیع نرمال ندارد، (توزیع متغیر پاسخ باید متعلق به خانواده توزیع‌های نمایی[۲۱] یا شبیه به این خانواده باشند) یا میانگین پاسخ به‌صورت مستقیم مدل‌سازی نمی‌شود. بلکه تبدیلی از میانگین پاسخ با استفاده از تابع پیوند[۲۲] مدل‌سازی می‌شود. مدل‌های خطی تعمیم‌یافته دارای سه مؤلفه هستند:

۱. **مؤلفه تصادفی:** توزیع شرطی متغیر پاسخ را مشخص می‌کند. توزیع Y_i معمولاً از خانواده توزیع‌های نمایی، مانند نرمال[۲۳]، دوجمله‌ای[۲۴]، پواسن[۲۵]، گاما[۲۶] یا خانواده نرمال معکوس

[۲۱] Exponential Family Distributions
[۲۲] Link Function
[۲۳] Normal
[۲۴] Binomial
[۲۵] poisson
[۲۶] Gamma

انتخاب می‌شود. انتخاب مؤلفه تصادفی بنا به ساختار متغیر پاسخ، متفاوت است. برای مثال، اگر ساختار متغیر پاسخ به‌صورت شمارشی غیرمنفی باشد، می‌توان توزیع پواسن را به‌عنوان توزیع مؤلفه تصادفی انتخاب نمود. از طرفی اگر مشاهدات پیوسته باشند، بهتر است از توزیع نرمال به‌عنوان توزیع مؤلفه تصادفی استفاده شود (لیندسی و جامس، ۱۹۹۷؛ هتر تورنر، ۲۰۰۸؛هستی و همکاران، ۱۹۹۲).

۲. **مؤلفه منظم:** اگر امید ریاضی شرطی Y با نماد $\mu = E(Y|x)$ نشان داده شود، آنگاه مقدار μ با تغییر سطوح متغیرهای تبیینی، تغییر می‌کند. مؤلفه منظم، نقش متغیرهای تبیینی را مشخص می‌کند. در هنگام ارائه معادله، متغیرهای تبیینی به‌صورت خطی، تحت عنوان پیشگو، در طرف راست مدل قرار می‌گیرند. به‌عبارتی دیگر، مؤلفه منظم بخشی از مدل را شامل می‌شود که نقش متغیرهای تبیینی، $\{x_j\}$، را در معادله رگرسیونی نشان می‌دهد:

$$\eta = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k.$$

ترکیب خطی η، پیشگوی خطی[۲۷] نامیده می‌شود. البته ممکن است بعضی x_jها بر اساس تابعی از بقیه متغیرها، در مدل باشند.

۳. **تابع پیوند:** این مؤلفه نشان می‌دهد که μ چگونه با متغیرهای تبیینی، در پیشگوی خطی، ارتباط دارد. البته، می‌توان میانگین μ را به‌طور مستقیم یا به‌صورت تابعی از آن، مدل‌سازی کرد. بنابراین در مدل داریم:

$$g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

که در آن تابع $g(.)$، تابع پیوند نامیده می‌شود.

تابع پیوند انواع مختلفی دارد، که می‌توان به تابع پیوند همانی[۲۸] $g(\mu) = \mu$، تابع پیوند لگاریتمی[۲۹] $g(\mu) = log(\mu)$، و تابع پیوند لجیت[۳۰] $g(\mu) = log(\frac{\mu}{1-\mu})$ اشاره کرد.

هر توزیع احتمال مؤلفه تصادفی، دارای تابعی خاص از میانگین است که پارامتر طبیعی[۳۱] نامیده می‌شود. یافتن پارامتر طبیعی، برای توزیع‌های خانواده نمایی آسان است. تابع چگالی خانواده نمایی را می‌توان به‌صورت

$$f(y;\theta) = a(y)b(\theta)exp(y\phi(\theta))$$

[۲۷] Linear Predictor

[۲۸] Identity Link Function

[۲۹] Log Link Function

[۳۰] Logit Link

[۳۱] Natural Parameter

ارائه کرد، که در آن به‌ترتیب $a(y)$ و $b(\theta)$ توابعی اختیاری از y و θ هستند که هر دو توابعی مثبت می‌باشند. پارامتر طبیعی در خانواده نمایی، برابر $\phi(\theta)$ است. به‌عنوان مثال، پارامتر طبیعی توزیع نرمال و توزیع پواسن، به‌ترتیب، میانگین و لگاریتم میانگین است. تابع پیوندی که پارامتر طبیعی را به‌صورت $g(\mu)$ در GLM استفاده می‌کند، تابع پیوند کانونی[۳۲] نامیده می‌شود.

رگرسیون معمولی و مدل‌های $ANOVA$ برای متغیرهای پیوسته، حالت‌های خاصی از $GLMs$ هستند. در این مدل‌ها فرض می‌شود مؤلفه تصادفی دارای توزیع نرمال است. میانگین نیز به‌صورت مستقیم، از تابع پیوند همانی به‌دست می‌آید. به‌کارگیری $GLMs$، به منظور یکپارچه کردن روش‌های مختلف رگرسیونی است. از این دیدگاه، می‌توان رگرسیون، $ANOVA$ و مدل‌های داده‌های رسته‌ای را حالت‌های خاصی از مدل‌های خطی تعمیم‌یافته دانست.

۴.۱ استنباط بیزی

در دیدگاه کلاسیک استنباط آماری، پارامترهای مدل ثابت ولی نامعلوم فرض می‌شوند و بر اساس روش‌هایی با استفاده از داده‌ها به‌دنبال شناخت آن‌ها هستیم. در مقابل، دیدگاه بیزی فرض می‌کند که پارامترهای مدل تحقق‌هایی از توزیع‌های تصادفی هستند و به‌عبارتی فرض می‌شود پارامترها خود متغیرهای تصادفی هستند. این دیدگاه، با هدف کمی کردن عدم اطمینان در مورد پارامتر، یک توزیع احتمالی (توزیع پیشین[۳۳]) که برگرفته از اطلاعات یا تجربیات پیشین محقق است، را برای پارامتر مورد بررسی لحاظ می‌کند. دیدگاه بیزی با درآمیختن اطلاعات قبلی محقق در مورد پارامتر نامعلوم (توزیع پیشین $p(\theta)$) و اطلاعات موجود در نمونه (تابع درستنمایی $L(\theta|y)$) به‌وسیله قاعده بیز به استنباط در مورد پارامتر می‌پردازد. اگر پارامتر نامعلوم θ باشد و تابع درستنمایی مدل را به‌صورت $p(Y_1 = y_1, \cdots, Y_n = y_n \mid \theta) = L(\theta|y)$ نشان دهیم، بر اساس قاعده بیز داریم

$$p(\theta|Y_1 = y_1, \cdots, Y_n = y_n) = \frac{p(Y_1 = y_1, \cdots, Y_n = y_n|\theta)p(\theta)}{\int_\Theta \prod_{i=1}^n p(Y_i = y_i|\theta)p(\theta)d\theta}$$

که در آن Θ فضای پارامتر است و می‌تواند یک یا چندبعدی باشد. می‌توان نوشت

$$p(\theta \mid Y_1 = y_1, \cdots, Y_n = y_n) = p(\theta)p(Y_1 = y_1, \cdots, Y_n = y_n \mid \theta)h_{\mathbf{Y}}(\mathbf{y}) \propto$$
$$p(\theta)p(Y_1 = y_1, \cdots, Y_n = y_n \mid \theta) \qquad (۱.۱)$$

که در آن $h_{\mathbf{Y}}(\mathbf{y})$ ثابت نرمال‌ساز[۳۴] نامیده می‌شود و به پارامتر θ وابسته نیست. با نظر به فرمول (۱.۱) ترکیب توزیع پیشین و تابع درستنمایی به تابع توزیع احتمال جدیدی منجر می‌شود که آن را توزیع پسین می‌نامیم و تمامی استنباط‌های بیزی، بر اساس آن صورت می‌گیرد. آن‌چه که از توزیع پسین به‌دست

[۳۲] Canonical Link Function

[۳۳] Prior Distribution

[۳۴] Normalizing constant

می‌آید، مورد ارزیابی قرار می‌گیرد. نتیجه این ارزیابی، رویکرد بیزی را بررسی می‌کند و صحت یا عدم آن را مشخص می‌کند.

۱.۴.۱ مدل‌های رگرسیون بیزی

امروزه کاربرد روش‌های مدل‌سازی و رگرسیون در رشته‌های مختلف علوم به‌طور گسترده رواج یافته است. این روش‌ها برای افزایش کیفیت پیش‌بینی، با استفاده از داده‌های نمونه و مجموعه اطلاعات به‌دست آمده از آزمایش‌ها و تحقیقات مختلف به‌کار می‌روند. بدیهی است استفاده مناسب از این روش‌ها برای برطرف نمودن قسمت عمده‌ای از عدم قطعیت‌ها و پیش‌بینی نتایج، در شرایطی که به‌دست آوردن اطلاعات سخت و گاهی ناممکن است، بسیار مفید می‌باشد. غالباً روش‌های شبکه‌های عصبی، منطق فازی، رگرسیون و سایر روش‌های آماری در این‌گونه مدل‌سازی‌ها مورد استفاده قرار می‌گیرند. در بسیاری از موارد لزوم تخمین دقیق نتایج و به‌دست آوردن روابط بین پارامترها و متغیرهای مؤثر در نتایج، طیف وسیعی از روش‌های رگرسیونی را به‌وجود آورده است. یک زیررده پرکاربرد در این زمینه، مدل‌های رگرسیون بیزی است. رگرسیون بیزی همان مدل رگرسیون در استنباط آماری است که در حیطه استنباط بیزی صورت می‌گیرد (توماس و مینکا، ۲۰۰۱). به‌کارگیری این روش، در تفسیر داده‌ها و پیش‌بینی نتایج و برآورد پارامترهای بسیاری از مسائل، دارای اهمیت بسیاری است. این امر از انجام آزمایش‌های پرهزینه و گاهی غیرممکن جلوگیری می‌کند.

معادله‌های پیشگو ابزار مهمی در استنباط‌های آماری به‌شمار می‌آیند. اما معمولاً داده‌های مورد استفاده در این معادله‌ها، ناقص هستند. داده‌ها اغلب، در اندازه نمونه، ناقص‌اند، خطا دارند یا کامل نمی‌باشند. در چنین شرایطی، مدل‌های رگرسیونی کارایی مناسبی ندارند. بهترین راه برای حل این مشکل، استفاده از دیدگاه بیزی برای مدل‌بندی داده‌ها است. در رگرسیون بیزی محققان آماری از داده‌های مکمل ضعیف استفاده می‌کنند. آن‌ها این عملکرد را با وارد کردن توزیع پیشین انجام می‌دهند. از این رو، مدلی که ارائه می‌دهند نسبت به رگرسیون کلاسیک برتر است.

در ادامه، حالت خاصی از رگرسیون بیزی، یعنی رگرسیون خطی بیزی را معرفی می‌کنیم.

- رگرسیون خطی چندمتغیره را با پذیره‌های خطاهای نرمال و ناهمبستگی بین خطا و متغیرهای تبیینی در نظر بگیرید.
- اگر خطاها مستقل و نرمال توزیع شوند، متغیر پاسخ نیز توزیع نرمال خواهد داشت.
- توزیع پیشین ناآگاهی‌بخش[۳۵] را برای بردار پارامترها در نظر می‌گیریم و سپس توزیع‌های شرطی هر یک از پارامترها را به‌دست می‌آوریم (والتر و همکاران، ۲۰۰۹).
- حتی اگر توزیع پیشین مستقل فرض شود، پارامترها در توزیع پسین همبسته خواهند بود. در صورتی که پارامترها همبستگی زیادی داشته باشند، از الگوریتم‌های $MCMC$ استفاده می‌شود. سپس متغیرهای تبیینی و پاسخ را استاندارد کرده و مدل را به آن‌ها برازش می‌دهیم و ضرایب

[۳۵] Non Informative Prior

را برآورد می‌کنیم. واریانس خطاها در مقیاس اصلی از ضرب واریانس استانداردشده در انحراف معیار اصلی به‌دست می‌آید.

رگرسیون خطی

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

را در نظر بگیرید که در آن متغیرهای تبیینی $\mathbf{x}_i$، $i = ۱, ۲, \cdots, n$، دارای بعد k، $\boldsymbol{\beta}$ بردار k بعدی پارامترها و ϵ_i جمله خطا و دارای توزیع نرمال است. به‌عبارتی دقیق‌تر

$$\epsilon_i \sim N(۰, \sigma^۲), \quad i = ۱, ۲, \cdots, n$$

بنابراین تابع درستنمایی به‌صورت زیر خواهد بود:

$$p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}, \sigma^۲) \propto (\sigma^۲)^{-\frac{n}{۲}} exp(-\frac{۱}{۲\sigma^۲}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})).$$

که در آن، $\mathbf{y}$ بردار مشاهدات پاسخ هستند. همان‌طور که می‌دانیم، برآورد ضرایب رگرسیون به روش کمترین توان‌های دوم خطا و همچنین روش درستنمایی ماکسیمم به‌صورت زیر است:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-۱}\mathbf{X}^T\mathbf{y}$$

که در آن $\mathbf{X}$ ماتریس $n \times k$ است. در استنباط بیزی توزیع پیشین و درستنمایی بر اساس قضیه بیز ترکیب می‌شوند و توزیع پسین پارامترهای $\boldsymbol{\beta}, \sigma$ را به وجود می‌آورند. توزیع پیشین می‌تواند صورت‌های متفاوتی داشته باشد. برای استنباط پسین، می‌توانیم توزیع پیشین را توأم در نظر بگیریم. توزیع پیشین توأم $p(\boldsymbol{\beta}, \sigma^۲)$ است. تابع لگاریتم درستنمایی پارامتر $\boldsymbol{\beta}$ دارای صورت توان دوم است. این تابع را بر اساس $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ و تغییر متغیر

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \quad (۲.۱)$$

بازنویسی می‌کنیم. با استفاده از (۲.۱) تابع درستنمایی به‌صورت

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^۲) \propto (\sigma^۲)^{\frac{\nu}{۲}} exp(-\frac{\nu S^۲}{۲\sigma^۲})(\sigma^۲)^{-\frac{(n-\nu)}{۲}} \times exp(-\frac{۱}{۲\sigma^۲}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})) \quad (۳.۱)$$

تبدیل می‌شود، که در آن $\nu = n - k$، $\nu S^۲ = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ و k تعداد ضرایب رگرسیون است. توزیع پیشین را به‌صورت

$$p(\boldsymbol{\beta}, \sigma^۲) \propto p(\sigma^۲)p(\boldsymbol{\beta} \mid \sigma^۲)$$

در نظر می‌گیریم که در آن $p(\sigma^۲)$ توزیع گامای معکوس

$$p(\sigma^۲) \propto (\sigma^۲)^{(-\frac{\nu_۰}{۲}+۱)} exp(-\frac{\nu_۰ S_۰^۲}{۲\sigma^۲}) \qquad (۴.۱)$$

است به‌طوری که $\nu_۰$ و $S_۰^۲$ مقادیر معلوم پارامترهای توزیع هستند. همچنین فرض می‌کنیم چگالی شرطی پیشین $p(\boldsymbol{\beta} \mid \sigma^۲)$ نرمال $N(\mu_۰, \sigma^۲\Lambda_۰)$ است. بنابراین

$$p(\boldsymbol{\beta}|\sigma^۲) \propto \sigma^{۲-\frac{k}{۲}} exp(-\frac{۱}{۲\sigma^۲}(\boldsymbol{\beta}-\mu_۰)^T\Lambda_۰(\boldsymbol{\beta}-\mu_۰)). \qquad (۵.۱)$$

بر اساس روابط (۳.۱) تا (۵.۱)، توزیع پسین به‌صورت زیر به‌دست می‌آید:

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^۲|\mathbf{y}\mathbf{X}) &\propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma^۲\mathbf{X}) p(\boldsymbol{\beta}|\sigma^۲) p(\sigma^۲) \\ &\propto \sigma^{۲-\frac{n}{۲}} exp(-\frac{۱}{۲\sigma^۲}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}))(\sigma^۲)^{(-\frac{\nu_۰}{۲}+۱)} exp(-\frac{\nu_۰ S_۰^۲}{۲\sigma^۲}) \\ &\times \sigma^{۲-\frac{k}{۲}} exp(-\frac{۱}{۲\sigma^۲}(\boldsymbol{\beta}-\mu_۰)^T\Lambda_۰(\boldsymbol{\beta}-\mu_۰)) \end{aligned}$$

از شکل این توزیع پسین معلوم است که از هیچ توزیع معروف و مشخصی پیروی نمی‌کند و در نتیجه استخراج نتایج به‌صورت بسته امکان‌پذیر نیست. نکته قابل توجه این است که این توزیع پسین مدل رگرسیون خطی است که مدل پیچیده‌ای محسوب نمی‌شود. بنابراین قابل حدس است که برای مدل‌های رگرسیونی پیچیده‌تر، مانند مدل‌های موردنظر در این پایان‌نامه، توزیع پسین شکل پیچیده‌تری خواهد داشت. بنابراین باید از روش‌های تقریب برای تقریب توزیع پسین استفاده کرد. یک رده کلی از روش‌های تقریب توزیع‌های پسین، روش‌های مبتنی بر نمونه‌گیری مانند روش‌های رد و پذیرش[۳۶] (رابرت و کسلا، ۲۰۱۰)، نمونه‌گیری نقاط مهم[۳۷] و نمونه‌گیری $MCMC$ می‌باشد. با استفاده از این روش‌ها، برای انجام استنباط‌های بیزی بر اساس توزیع پسین، دیگر نیازی به بهینه‌سازی توابع پیچیده و محاسبه گشتاورهای توزیع‌های پیچیده (انتگرال‌گیری‌های پیچیده) نیستیم. در ادامه، برخی از این روش‌های نمونه‌گیری را به اختصار معرفی می‌کنیم.

۵.۱ روش‌های تقریب توزیع پسین مبتنی بر نمونه‌گیری

دو گروه اصلی مسائل عددی در استنباط آماری، مسائل بهینه‌سازی و انتگرال‌گیری هستند. مثال‌های عددی نشان می‌دهند که مثلاً محاسبه انتگرال همیشه از روش مستقیم قابل انجام نیست. یک راه استفاده از شبیه‌سازی[۳۸] است که به روش‌های مونت کارلو معروف هستند. در روش‌های مونت کارلو از تابع مورد

[۳۶] Accept-Reject Methods
[۳۷] Importance Sampling
[۳۸] Simulation

نظر $p(y)$ شبیه‌سازی‌های مستقل $(y^{(1)}, y^{(2)}, \cdots, y^{(T)})$ را تولید می‌کنیم و از آن برای منظور برآورد انتگرال استفاده می‌کنیم (کسلا، ۱۹۹۶). در واقع مسئله اصلی محاسبه (برآورد) انتگرال

$$\mathbf{I} = \int g(y)dy \tag{۶.۱}$$

است. می‌توان تابع $g(y)$ را به‌صورت زیر بازنویسی کرد:

$$g(y) = h(y)p(y)$$

که در آن $p(y)$ یک تابع چگالی است. در این صورت محاسبه انتگرال (۶.۱) معادل محاسبه امید ریاضی $h(y)$ تحت توزیع با تابع چگالی $p(y)$ است. یعنی

$$\int g(y)dy = \int h(y)p(y)dy = E_p[h(y)].$$

از طریق شبیه‌سازی از توزیع $p(y)$ و با تکیه بر قانون قوی اعداد بزرگ[۳۹] (گات، ۲۰۰۵) به‌سادگی می‌توان انتگرال مورد نظر را به‌صورت زیر برآورد کرد:

$$\hat{\mathbf{I}} = \frac{1}{T}\sum_{t=1}^{T} h(y^{(t)}) \xrightarrow{a.s} \mathbf{I}.$$

بنابراین تمامی مسائل استنباطی به تولید نمونه از توزیع منتهی می‌شود. حال در دیدگاه بیزی $p(\cdot)$ همان توزیع پسین $p(\theta|y)$ است. در اغلب موارد تولید مستقیم نمونه از $p(\theta|y)$ امکان‌پذیر نیست و باید از روش‌های غیرمستقیم تولید نمونه مانند الگوریتم‌های $MCMC$ استفاده کنیم (چن و همکاران، ۲۰۰۰). روش‌های نمونه‌گیری نقاط مهم و الگوریتم‌های متروپلیس-هستینگز و نمونه‌گیر گیبز از جمله روش‌های تولید نمونه غیرمستقیم هستند که در ادامه به معرفی آن‌ها می‌پردازیم.

۱.۵.۱ نمونه‌گیری نقاط مهم

در این قسمت، به‌طور مختصر به معرفی نمونه‌گیری نقاط مهم می‌پردازیم. این روش، به توابعی تکیه دارد که به برخی از نقاط در نمونه‌گیری اهمیت بیشتری نسبت به بقیه می‌دهد. از این رو به آن، نمونه‌گیری نقاط مهم می‌گویند. در صورتی که تولید نمونه مستقیم از توزیع امکان‌پذیر نباشد، نمونه‌گیری نقاط مهم جایگزین آن می‌شود. اساس این روش مبتنی بر استفاده از تابع چگالی دیگری، مانند $f(\cdot)$، است. در این حالت، صورت نمایش انتگرال به‌صورت زیر قابل بازنویسی است:

$$E_p[h(Y)] = \int_{\chi} (h(y)\frac{p(y)}{f(y)})f(y)dy = E_f[h(y)\frac{p(y)}{f(y)}].$$

این صورت نمایش، اجازه استفاده از توزیع‌های دیگری جز $p(\cdot)$ را می‌دهد. به $p(\cdot)$ توزیع اصلی و به $f(\cdot)$ توزیع ابزاری[۴۰] می‌گویند. توزیع ابزاری باید طوری انتخاب شود که تولید نمونه از آن ساده باشد.

[۳۹] Strong Law of Large Number

[۴۰] Instrumental Distribution

همچنین، از نمونه تولیدشده توسط توزیع ابزاری، می‌توان به دفعات استفاده کرد. یعنی نه‌تنها برای توابع $h(\cdot)$ متفاوت بلکه برای توزیع‌های $p(\cdot)$ گوناگون، نمونه مشابه قابل استفاده است.

برای محاسبه انتگرال موجود به این روش، می‌توان به‌صورت زیر عمل کرد:

۱. نمونه $y^{(1)}, y^{(2)}, \cdots, y^{(T)}$ را از توزیع ابزاری تولید می‌کنیم.

۲. از تقریب زیر استفاده می‌کنیم:

$$\frac{1}{T}\sum_{t=1}^{T} h(y^t)\frac{p(y^t)}{f(y^t)}$$

استفاده از سایر توزیع‌ها به جای $p(\cdot)$ می‌تواند باعث کاهش واریانس برآوردگرهای حاصل شود. بنابراین در پی تابع ابزاری هستیم که واریانس برآوردگر نمونه‌گیری نقاط مهم را کمترین مقدار کند. اما این انتخاب بهینه امکان‌پذیر نیست. زیرا اطلاع از آن نیازمند دانستن مقدار انتگرالی است که به دنبال محاسبه آن هستیم. نمونه‌گیری نقاط مهم در ابعاد بالا عملکرد خیلی خوبی ندارد. پس برای حل انتگرال (۶.۱) به دنبال روشی هستیم که در ابعاد بالا نیز عملکرد مناسبی داشته باشد.

۲.۵.۱ الگوریتم متروپلیس-هستینگز

اساس کار الگوریتم‌های $MCMC$ برای تولید نمونه از توزیع $p(\cdot)$ به این صورت است که ابتدا هسته مارکوف[۴۱] $K(\cdot|\cdot)$ را با چگالی مانای $p(\cdot)$ می‌سازیم. آنگاه زنجیر مارکوف $Y^{(t)}$ را با استفاده از این هسته تولید می‌نماییم، که توزیع حدی زنجیر $Y^{(t)}$ چگالی $p(\cdot)$ است. ساختن هسته مارکوف $K(\cdot|\cdot)$ که به چگالی دلخواه $p(\cdot)$ نسبت داده شود، ممکن است کار دشواری باشد. جالب است بدانید روش‌هایی وجود دارند که می‌توان به کمک آن‌ها هسته‌هایی که عمومیت بیشتری دارند را برای هر چگالی $p(\cdot)$ به‌دست آورد. الگوریتم متروپلیس –هستینگز نمونه‌ای از این روش‌ها است. در این روش، چگالی هدف $p(\cdot)$ را فرض کرده و با چگالی شرطی $q(x|y)$ که شبیه‌سازی از آن بسیار ساده‌تر است، کار را انجام می‌دهیم. به‌علاوه، چگالی $q(\cdot|\cdot)$ اختیاری لحاظ می‌گردد. در حالی‌که تنها شرطی که باید به آن توجه نماییم، این است که نسبت $\frac{p(x)}{q(x|y)}$ مشخص و ثابت باشد. این ثابت مستقل از y است. این الگوریتم را می‌توان به‌صورت زیر تعریف نمود:

۱. با مقدار اولیه $y^{(0)}$ شروع کنید.

۲. برای مقدار کنونی $y^{(t)}$، $t = 0, 1, 2, \cdots$ یک نقطه پیشنهادی را از توزیع پیشنهادی $q(x|y^{(t)})$ تولید کنید:

$$X_t \sim q(x|y^{(t)})$$

[۴۱] Markov Kernel

۳. نسبت چگالی نقطه پیشنهادی و مقادیر کنونی را محاسبه کنید:

$$\frac{p(x)}{p(y^{(t)})}\frac{q(y^{(t)}|x)}{q(x|y^{(t)})}$$

سپس:

$$Y^{(t+1)} = \begin{cases} X_t & \text{با احتمال} \quad \rho(X_t, y^{(t)}) \\ y^{(t)} & \text{با احتمال} \quad 1 - \rho(X_t, y^{(t)}) \end{cases}$$

که در آن

$$\rho(x, y) = min\{\frac{p(x)}{p(y)}\frac{q(y \mid x)}{q(x \mid y)}, 1\}.$$

این مراحل یک زنجیر مارکوف را تولید می‌کند.

احتمال $\rho(x, y)$ احتمال پذیرش متروپلیس-هستینگز نامیده می‌شود. نرخ پذیرش میانگین، احتمال پذیرش روی تکرارها است. یعنی

$$\bar{\rho} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T}\rho(X_t, Y^{(t)}) = \int \rho(x, y)p(y)q(x \mid y)dydx.$$

در متون مربوط به الگوریتم‌های $MCMC$ نشان داده شده است (رابرت و کسلا، ۲۰۱۰) که توزیع مانای زنجیر تولیدشده $\{Y^{(t)}\}$ در بالا همان $p(\cdot)$ است. عملکرد الگوریتم به انتخاب $q(\cdot|\cdot)$ وابسته است. خروجی‌های $MCMC$ و نمونه‌گیری نقاط مهم در ظاهر مشابه به نظر می‌رسند. در حالی که باید به این نکته توجه داشته باشیم که نمونه مونت کارلوی زنجیر مارکوفی همبسته هستند. این مطلب باعث می‌شود که کیفیت نمونه‌ای که از $MCMC$ به‌دست می‌آید، از نمونه‌گیری نقاط مهم کمتر باشد. پس برای این که به کیفیت مشابه دست یابیم، باید در شبیه‌سازی‌های $MCMC$ حجم نمونه بیشتری تولید کنیم.

اگر فرض متقارن بودن توزیع پیشنهادی را نیز اضافه کنیم، یعنی $q(x|y) = q(y|x)$ آن‌گاه

$$\rho(x_t, y^{(t)}) = min\{1, \frac{p(x_t)}{p(y^{(t)})}\}.$$

در نتیجه الگوریتم مستقل از $q(\cdot|\cdot)$ خواهد بود. در صورتی که نسبت $\frac{p(x_t)}{q(x_t|y^{(t)})}$ از نسبت $\frac{p(y^{(t)})}{q(y^{(t)}|x_t)}$ بزرگتر باشد، همواره مقدار پیشنهادی x_t را می‌پذیریم. حتی در صورتی که کوچکتر باشد (در حالی که اختلاف ناچیز باشد)، نیز مقدار x_t معمولاً پذیرفته می‌شود. دو حالت خاص الگوریتم متروپلیس-هستینگز مستقل[۴۲] (IMH) و الگوریتم متروپلیس-هستینگز گام زدن تصادفی[۴۳] $(RWMH)$ را در ادامه ذکر می‌کنیم.

[۴۲] Independent Metropolis-Hastings Algorithm

[۴۳] Random Walk Metropolis-Hastigs Algorithm

الگوریتم متروپلیس-هستینگز مستقل

در حالتی که مقدار پیشنهادی جدید از مقدار حال زنجیر یعنی $Y^{(t)}$، مستقل باشد حالت خاصی از الگوریتم اتفاق می‌افتد که در آن احتمال پذیرش متروپلیس -هستینگز به‌صورت زیر است:

$$\rho(y^{(t)}, x_t) = min\{۱, \frac{p(x_t)}{p(y^{(t)})}\frac{q(y^{(t)})}{q(x_t)}\}.$$

بنابراین، مقدار پیشنهادی از یک توزیع مورد علاقه که مستقل از مقدار کنونی است نتیجه می‌شود. به الگوریتم‌های حاصل از این توزیع پیشنهادی، الگوریتم MH مستقل می‌گویند. برای اطلاع از ویژگی‌های این نوع الگوریتم به رابرت و کسلا (۲۰۱۰) مراجعه کنید.

الگوریتم متروپلیس-هستینگز گام زدن تصادفی

در این الگوریتم مقدار جدید پیشنهادی بر طبق مقدار شبیه‌سازی‌شده قبلی تولید می‌شود، که در آن یک جستجوی محلی[۴۴] حول یک همسایگی از مقدار جاری زنجیر مارکوف صورت می‌گیرد. اجرای این ایده با شبیه‌سازی X_t طبق $X_t = Y^{(t)} + \epsilon_t$ است که در آن ϵ_t خطای تصادفی است و مستقل از $Y^{(t)}$ با چگالی $g(\cdot)$ می‌باشد. در این حالت توزیع پیشنهادی[۴۵] عبارت است از:

$$q(x|y^{(t)}) = g(x - y^{(t)})$$

در صورتی که توزیع $g(\cdot)$ حول صفر متقارن باشد (مثلاً در توزیع نرمال یا نرمال چندمتغیره، با میانگین صفر یا توزیع یکنواخت استاندارد)، زنجیر مارکوف با یک فرآیند گام زدن تصادفی متناظر می‌شود. با داشتن $y^{(t)}$ الگوریتم به‌صورت زیر است:

۱. برای مقدار کنونی یک نقطه پیشنهادی را از توزیع $g(\cdot)$، تولید کنید

$$X_t \sim g(x - y^{(t)})$$

۲. نسبت چگالی نقطه پیشنهادی و مقادیر کنونی را محسابه کنید:

$$\frac{p(X_t)}{p(y^{(t)})}$$

۳. سپس

$$Y^{(t+۱)} = \begin{cases} X_t & \text{با احتمال} \quad min\{\frac{p(X_t)}{p(y^{(t)})}, ۱\} \\ y^{(t)} & \text{سایر نواحی} \end{cases}$$

[۴۴] Local Exploration

[۴۵] Proposal Distribution

متذکر می‌شویم که احتمال پذیرش به $g(\cdot)$ وابسته نیست. این بدین معنی است که، برای جفت $(x_t, y^{(t)})$ احتمال پذیرش $y^{(t)}$ در حالتی که از نرمال یا کوشی تولید شود، مشابه است. اما باید توجه داشت که تغییر $g(\cdot)$ بازه‌های متفاوتی را برای $Y^{(t)}$ به وجود می‌آورد و در نتیجه نرخ‌های پذیرش متفاوت خواهند شد. پس انتخاب $g(\cdot)$ بر عملکرد الگوریتم مؤثر است.

۳.۵.۱ الگوریتم نمونه‌گیر گیبز

در سال‌های اخیر، آمارشناسان به‌طور عمده‌ای روش‌های الگوریتم $MCMC$ را به کار برده‌اند. الگوریتم نمونه‌گیر گیبز، یکی از بهترین روش‌های شناخته‌شده است. روش نمونه‌گیر گیبز یک ابزار شبیه‌سازی برای یافتن نمونه‌هایی از چگالی‌های توأم پیچیده می‌باشد. نمونه‌گیری گیبز در به‌دست آوردن توزیع‌های پسین کناری[۴۶]، در مدل‌های بیزی می‌باشد (کسلا و همکاران، ۱۹۹۲). یکی از محدودیت‌ها در کاربرد گسترده روش‌های بیزی، یافتن توزیع پسین کناری است، که اغلب نیازمند انتگرال‌گیری از تابع‌هایی با ابعاد بالا بوده و این نیز مستلزم محاسبات بسیار پیچیده می‌باشد. معمولاً وقتی بردار پارامتر چندبعدی است، حتی روش‌های عددی انتگرال‌گیری مستقیم هم ممکن است برای حل چنین انتگرال‌هایی مفید نباشند. در این صورت، استفاده از روش نمونه‌گیر گیبز می‌تواند به‌عنوان یک روش جایگزین مؤثر واقع شود.

نمونه گیر گیبز دومرحله‌ای

نمونه‌گیر گیبز دومرحله‌ای یک زنجیر مارکوف را از توزیع توأم، به‌صورت زیر ایجاد می‌کند. اگر فرض کنیم متغیرهای تصادفی X و Y دارای تابع چگالی توأم $p(x,y)$ با چگالی‌های شرطی $p(x|y)$ و $p(y|x)$ هستند، نمونه‌گیر دومرحله‌ای گیبز مقادیر $(X^{(t)}, Y^{(t)}), t = ۱, ۲, \cdots$ را به‌صورت زیر تولید می‌کند:
با فرض $X_۰ = x_۰$ برای $t = ۰, ۱, ۲, \cdots$ تولید می‌کنیم:

۱. $Y^{(t+۱)} \sim p(y|x_{(t)})$

۲. $X_{(t+۱)} \sim p(x|y^{(t+۱)})$

در صورتی که شبیه‌سازی در هر دو مرحله ساده و تولید نمونه از آن‌ها ساده باشد، اجرای الگوریتم آسان است. یا به‌عبارت دیگر، در صورتی که چگالی‌های شرطی فرم بسته‌ای داشته باشند، شبیه‌سازی به سادگی انجام می‌شود. بنابراین، در صورتی که نمونه‌گیری مستقیم قابل انجام نباشد، تقریب‌های مونت کارلو می‌تواند مورد استفاده قرار بگیرد. به راحتی در می‌یابیم که اگر $(X_{(t)}, Y^{(t)})$ از توزیع $p(x,y)$ باشد، در این صورت $(X_{(t+۱)}, Y^{(t+۱)})$ نیز از توزیع $p(x,y)$ خواهد بود. به عبارت دیگر توزیع مانای این زنجیر، $p(x,y)$ است.

[۴۶] Marginal Posterior Distribution

نمونه‌گیر گیبز چندمرحله‌ای

نمونه‌گیر گیبز چندمرحله‌ای تعمیمی از نمونه‌گیر گیبز دومرحله‌ای است. فرض کنید که برای $m > ۱$، متغیر تصادفی $\mathbf{Y}$ به‌صورت $\mathbf{Y} = (Y_۱, \cdots, Y_m)$ نوشته شود. مؤلفه‌های Y_i ممکن است تک بعدی یا چند بعدی باشند. می‌توان برای $i = ۱, ۲, \cdots, m$ به‌صورت زیر شبیه‌سازی کرد:

$$Y_i \mid y_۱, y_۲, \cdots, y_{i-۱}, y_{i+۱}, \cdots, y_m \sim p_i(y_i \mid y_۱, y_۲, \cdots, y_{i-۱}, y_{i+۱}, \cdots, y_m)$$

الگوریتم نمونه‌گیر گیبز تحت انتقال از $Y^{(t)}$ به $Y^{(t+۱)}$ با تکرار $t = ۱, ۲, \cdots$ و با فرض کردن $\mathbf{y}^{(t)} = (y_۱^{(t)}, y_۲^{(t)}, \cdots, y_m^{(t)})$ طی گام‌های زیر تولید می‌شود:

۱. $Y_۱^{(t+۱)} \sim p_۱(y_۱ \mid y_۲^{(t)}, \cdots, y_m^{(t)})$

۲. $Y_۲^{(t+۱)} \sim p_۲(y_۲ \mid y_۱^{(t+۱)}, y_۳^{(t)}, \cdots, y_m^{(t)})$

$\vdots$

m. $Y_m^{(t+۱)} \sim p_m(y_m \mid y_۱^{(t+۱)}, \cdots, y_{m-۱}^{(t+۱)})$

چگالی‌های $p_۱, \cdots, p_m$ چگالی شرطی کامل[۴۷] نامیده می‌شوند. یک خصوصیت نمونه‌گیر گیبز این است که تنها چگالی‌ها برای شبیه‌سازی مورد استفاده قرار می‌گیرند. بنابراین، حتی مسئله‌ای با ابعاد بالا را می‌توان به چند مسئله با بعدهای کوچکتر تقسیم کرد که این مزیت روش نمونه‌گیری گیبز به حساب می‌آید.

۴.۵.۱ الگوریتم متروپلیس-هستینگز تحت گیبز

نمونه‌گیر گیبز را در صورتی که برخی از چگالی‌های شرطی کامل را نتوان به‌سادگی شبیه‌سازی کرد، به آسانی با تنظیماتی می‌توان تعمیم داد. اگر تحت یک مجموعه از چگالی‌های شرطی کامل $p_۱, \cdots, p_m$ برخی از p_iها غیرمتداول باشند، نتایج نمونه‌گیر گیبز در صورتی که راهکار متروپلیس تحت گیبز را به کار بگیریم، به مخاطره نمی‌افتد. مثلاً فرض کنید شبیه‌سازی از

$$Y_i^{(t+۱)} \sim p_i(y_i | y_۱^{(t+۱)}, \cdots, y_{i-۱}^{(t+۱)}, y_{i+۱}^{(t)}, \cdots, y_m^{(t)})$$

مشکل باشد. در این حالت می‌توان برای تولید نمونه از آن (در یک مرحله از مراحل نمونه‌گیری گیبز) از الگوریتم MH استفاده کرد. در این حالت این گام با یک الگوریتم MH که توزیع مانای آن $p(y_i|y_۱^{(t+۱)}, \cdots, y_{i-۱}^{(t+۱)}, y_{i+۱}^{(t)}, \cdots, y_m^{(t)})$ است، جایگزین می‌شود. این نوع الگوریتم ترکیبی از دو الگوریتم گیبز و MH است که به آن الگوریتم متروپلیس-هستینگز تحت گیبز گویند. در حالتی که شبیه‌سازی بیشتر از یک چگالی شرطی کامل مشکل باشد، می‌توان به همین ترتیب مراحل مختلفی از نمونه‌گیر گیبز را با الگوریتم‌های MH ترکیب کرد.

[۴۷] Full Conditional Density

نرم‌افزارهای $JAGS$[۴۸]، $WinBUGS$[۴۹] و $OpenBUGS$ (اندرو و فینلی، ۲۰۱۳) به‌طور خودکار تولید نمونه از توزیع‌های پسین را با روش نمونه‌گیری گیبز یا MH تحت گیبز انجام می‌دهند و چگالی‌های شرطی کامل را شبیه‌سازی می‌کنند.

۵.۵.۱ معیار همگرایی گلمن-روبین

جهت تشخیص همگرایی توسط شبیه‌سازی، گلمن و روبین در سال ۱۹۹۲ روشی را در دنباله‌های متعدد با نقاط شروع مختلف پیشنهاد کردند. با داشتن چگالی هدف $p(\theta)$، این معیار بر اساس هفت مرحله تشکیل می‌شود.

- مرحله اول، شامل تولید $m \geq ۲$ دنباله با طول $۲n$ و با مقادیر آغازی متفاوت می‌باشد. در این روش، n تکرار اول هر دنباله حذف شده و با n تکرار آخر همگرایی زنجیر بررسی می‌شود.

- مرحله دوم، برای هر پارامتر مورد نظر، مقادیر زیر محاسبه می‌شوند:

$$C = \frac{n}{m-۱}\sum_{i=۱}^{m}(\bar{\theta}_{i.} - \bar{\theta}_{..})^۲,$$
$$W = \frac{۱}{m}\sum_{i=۱}^{m} s_i^۲,$$

که در آن $\bar{\theta}_{i.}$ میانگین هر دنباله، $\bar{\theta}_{..}$ میانگین تمام m دنباله و $s_i^۲ = \frac{۱}{n-۱}\sum_{j=۱}^{n}(\theta_{ij} - \bar{\theta}_{i.})$ واریانس درون هر دنباله می‌باشند. توجه داشته باشید که $\frac{C}{n}$ واریانس بین میانگین‌های m دنباله و W، متوسط m واریانس درون دنباله‌ای است.

- مرحله سوم، میانگین $\hat{\mu}$ را برابر میانگین نمونه‌ای mn مقادیر شبیه‌سازی‌شده θ، یعنی $\hat{\mu} = \bar{\theta}_{..}$ قرار می‌دهیم.

- مرحله چهارم، واریانس توزیع هدف به عنوان متوسط وزنی W و C محاسبه می‌شود. یعنی $\hat{\sigma}^۲ = \frac{n-۱}{n}W + \frac{۱}{n}C$. از این رو، اگر $n \to \infty$، آنگاه $E(W) = \sigma^۲$.

- مرحله پنجم، توزیع $N(\hat{\mu}, \hat{\sigma}^۲)$ به عنوان برآورد توزیع هدف در نظر گرفته می‌شود که نتیجه آن تقریب توزیع t با میانگین $\hat{\mu}$، مقیاس $\sqrt{\hat{V}} = \sqrt{\hat{\sigma}^۲ + C/mn}$ و درجه آزادی $df = ۲\hat{V}^۲/\hat{var}(\hat{V})$ برای θ خواهد بود.

- مرحله ششم، مرحله نظارت بر همگرایی توسط برآورد عاملی است که عامل کاهش مقیاس بالقوه[۵۰]

[۴۸] Just Another Gibbs Sampler

[۴۹] Win Bayesian inference Using Gibbs Sampler

[۵۰] Potential Scale Reduction Factor

$(PSRF)$ نام دارد و به صورت زیر برآورد می‌شود:

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}\frac{df}{df-۲}}$$
$$= \sqrt{(\frac{n-۱}{n} + \frac{m+۱}{mn}\frac{C}{W})\frac{df}{df-۲}}.$$

این عامل زمانی که $n \to \infty$، به سمت یک کاهش می‌یابد.

- در مرحله آخر می‌توان گفت اگر $\hat{R}$ عددی نزدیک به یک را نشان دهد، همگرایی به توزیع مانا حاصل شده است. در غیر این‌صورت، نیازمند اجرای زنجیرهای طولانی‌تر، یا به عبارتی تولید نمونه‌های بیشتر خواهیم بود.

۶.۱ معیارهای برازش بیزی

گستردگی روش‌های $MCMC$ امکان برازش رده بزرگی از مدل‌ها را با کاوش پیچیدگی‌های داده فراهم ساخته است (گیلکس و همکاران، ۱۹۹۶). این توانایی، طبیعتاً پایه‌های مقایسه مدل‌ها را، با کمک تعریف یک رده از مدل‌های فشرده که اطلاعات را در داده‌ها توصیف می‌کند، فراهم می‌کند. در چارچوب مدل‌بندی کلاسیک، مقایسه مدل با تعریف اندازه برازش، که آماره انحراف[۵۱] می‌باشد و همچنین پیچیدگی[۵۲]، که تعداد پارامترهای آزاد در مدل است، همراه می‌شود (وهتری و اجانن، ۲۰۱۲؛ واتاناب، ۲۰۱۳). هر چه پیچیدگی افزایش یابد، برازش بهتری از مدل خواهیم داشت. مدل‌ها تحت این دو کمیت مقایسه می‌شوند و در ابتدا با کار آکاییک در سال ۱۹۷۳ آغاز شد. برای مقایسه مدل‌ها، پیشنهاد شد که می‌توان اندازه زیان مورد انتظار را، روی یک مجموعه داده تکراری، به‌دست آورد. هر چه اندازه زیان کوچکتر باشد، مدل بهتر است. برای اطلاع بیشتر به افرون (۱۹۸۶)، ریپلی (۱۹۹۶) و برنهام و اندرسون (۱۹۹۸) مراجعه کنید.

برای مقایسه مدل، معیار اطلاع بیزی را نیز می‌توان به کار گرفت. ولی باید به این نکته توجه داشت که تعداد پارامترها در مدل باید مشخص و از تعداد مشاهدات کمتر باشد (کاس و رفتری، ۱۹۹۵). اما، در مدل‌های پیچیده ممکن است تعداد پارامترها از تعداد مشاهدات بیشتر باشد و واضح است که نمی‌توان این روش‌ها را به‌صورت مستقیم به کار برد (گلفاند و دی، ۱۹۹۴).

اگر مدلی را برای داده‌های مشاهده‌شده y با چگالی $p(y|\theta)$ در نظر بگیریم، در این صورت $D(\theta) = -۲log\{p(y \mid \theta)\}$ انحراف نامیده می‌شود. انحراف، تابعی از θ است. برای انتخاب مدل، در دیدگاه کلاسیک، در مدل‌های لانه‌ای از آزمون فرضیه و در مدل‌های غیرلانه‌ای از معیار اطلاع آکاییک[۵۳] (AIC)

[۵۱] Deviance Statistics

[۵۲] Complexity

[۵۳] Akike Information Criterion

$$AIC = -۲log\{p(y|\hat{\theta})\} + ۲p$$

استفاده می‌شود که در آن $\hat{\theta}$ برآورد ماکسیمم درستنمایی و p تعداد پارامترها در مدل است. می‌توان پیش‌بینی‌های بهتری از مدل ارزیابی‌شده، با استفاده از انحراف داشت. عموماً داده‌های مستقل و معتبر نداریم و در صورتی که پارامترها برآورد شوند، تاوانی برای استفاده دوباره از داده‌ها نیاز است. تاوان برای معیار اطلاع آکایی برابر $۲p$ است. همچنین، AIC را در مدل‌هایی با پیشین آگاهی‌بخش[۵۴]، مانند مدل‌های مرتبه‌ای نمی‌توان به کار برد. همچنین، تعداد پارامتر آزاد عموماً مشخص نیست. اسچوارز (۱۹۸۷) معیار اطلاع بیزی[۵۵] (BIC) را به‌صورت زیر تعریف می‌کند:

$$BIC = -۲log\{p(y \mid \hat{\theta})\} + plog(n)$$

برای nهای بزرگ، جریمه BIC بزرگ‌تر از AIC است. این دو معیار، عملکردهای متفاوتی برای مقایسه مدل دارند. در سال ۱۹۹۶ میانگین انحراف پسین $\bar{D}(\theta) = E_{\theta|y}[D(\theta)]$ به‌عنوان اندازه‌ای از برازش، توسط اسپیگلهالتر و همکاران پیشنهاد شد. بنابراین، در استنباط بیزی بر اساس معیار AIC معیار جدیدی بر که به معیار اطلاع آکاییک مورد انتظار[۵۶] $(EAIC)$ معروف است و به‌صورت زیر بر حسب امید پسین محاسبه می‌گردد:

$$EAIC = \bar{D}(\theta) + ۲p.$$

مشابه $EAIC$ برای معیار اطلاع بیزی نیز $EBIC$[۵۷] به‌صورت زیر معرفی شد

$$EBIC = \bar{D}(\theta) + plog(n)$$

این معیار نسبت به معیارهای قبلی کاملاً متفاوت عمل می‌کند (اسپیگلهالتر و همکاران، ۲۰۰۲).

به‌طور کلی، جریمه‌ای که برای پیچیدگی در مدل باید به کار گرفته شود، مشخص نیست. در سال ۱۹۹۷ مقاله‌ای $\frac{1}{2}V_{\theta|y}\{D(\theta)\}$ را به‌عنوان جریمه در نظر گرفت، که در آن $V_{\theta|y}(.)$ واریانس تحت توزیع پسین است. معیار معرفی‌شده با این جریمه، در سال ۱۹۹۸ به‌صورت معیار اطلاع انحراف[۵۸] (DIC) تغییر یافت که به‌صورت زیر تعریف می‌شود (اسپیگلهالتر و همکاران، ۲۰۰۲):

$$DIC = D(\bar{\theta}) + ۲p_D = \bar{D}(\theta) + p_D \qquad (۷.۱)$$

که در آن

$$p_D = \bar{D}(\theta) - D(\bar{\theta})$$

[۵۴] Informative Prior

[۵۵] Bayesian Information Criterion

[۵۶] Expected Akike Information Criterion

[۵۷] Expected Bayesian Information Criterion

[۵۸] Deviance Information Criterion

$\bar{\theta} = E(\theta|y)$ و $D(\cdot)$ همان آماره انحراف است. در معیار DIC مقدار پیچیدگی بر اساس p_D تعریف می‌شود و سنجش برازش مدل نیز بر عهده $D(\bar{\theta})$ است.

در مدل‌هایی با اطلاعات پیشین کم‌آگاهی‌بخش DIC، تقریباً برابر معیار آکاییک است. زیرا، استنباط بیزی بر مبنای توزیع پسین است و توزیع پسین با حاصل‌ضرب توزیع پیشین و درستنمایی متناسب است. در صورتی که توزیع پیشین کم‌آگاهی‌بخش باشد، توزیع پسین با درستنمایی متناسب می‌شود و اندازه برازش برای معیار آکاییک و DIC مقادیر مشابهی به‌دست می‌آید.

در مقایسه مدل‌ها، مدلی که کمترین مقدار را در این معیارها به خود اختصاص دهد، به‌عنوان مدل مناسب برگزیده می‌شود (برانهام و اندرسون، ۱۹۹۸؛ کلوکس و همکاران، ۲۰۰۶).

در مدل‌های با پارامترهای زیاد، انجام الگوریتم‌های $MCMC$ کاری بسیار دشوار و تقریباً غیرممکن است. از این دشواری با پیدایش نرم‌افزارهایی مانند $JAGS$ و $WinBUGS$ کاسته شده است. در این پایان‌نامه، تمام استنباط‌ها بر اساس نرم‌افزار $JAGS$ انجام شده‌اند. در ادامه به معرفی مختصری از آن می‌پردازیم.

۷.۱ $JAGS$

استنباط بیزی، استنباطی بر پایه توزیع پسین بوده و در بسیاری از موارد صورت بسته‌ای برای آن وجود ندارد یا تقریب آن بسیار مشکل است. در نتیجه، از توزیع پسین با الگوریتم $MCMC$ نمونه تولید می‌کنیم. با نظر به این که $MCMC$ شامل نمونه‌گیر گیبز و متروپلیس-هستینگز است و با توجه به این که نمونه‌گیر گیبز قادر به ایجاد توزیع‌های پسین توأم و توزیع‌های پسین شرطی است، استفاده از نمونه‌گیر گیبز ترجیح داده می‌شود. اما، زمانی که بعد پارامتر بالا باشد، محاسبه این توزیع‌های شرطی تقریباً ممکن نیست. اما $BUGS$ برنامه‌ای برای تحلیل مدل‌های بیزی و حل مسائلی است که حل تحلیلی دقیق یا تقریبی برای آن‌ها وجود ندارد. شکل اولیه این برنامه $classicBUGS$ است که در سال‌های ۱۹۹۲-۱۹۹۶ ارائه شد. این برنامه یک رابطه خطی دستوری ساده بود و برای تحلیل زنجیرهای مارکوف ظرفیت محدودی داشت. در سال ۱۹۹۷، $WinBUGS$ ارائه شد. این نسخه جامع‌تری نسبت به $classicBUGS$ است. اما قابلیت اجرا را بر روی بسته‌های نرم‌افزاری به غیر از ویندوز ندارد. نسخه دیگری از $BUGS$ در سال ۲۰۰۲ به نام $OpenBUGS$ نوشته شد که بر خلاف نسخه قبلی، قابلیت نصب بر روی سیستم‌های عامل $Unix/Linux$ را داشت.

یک موتور منبع آزاد برای زبان $BUGS$ که در $C++$ نوشته شده، $JAGS$ است. در اجرای برنامه $JAGS$ ابتدا داده‌ها را وارد نموده و پس از فرمول‌بندی مدل از زنجیرها نمونه‌گیری می‌شود. برای شبیه‌سازی $MCMC$ باید چند مقدار آغازین را به مدل ارائه دهیم. سپس مدل را برای شبیه‌سازی آماده می‌کنیم. خروجی $MCMC$ در بخش اول شامل دوره سوزاندن[۵۹] و باقیمانده اجرا، خروجی است که برای همگرا شدن به توزیع پسین توأم مشاهده می‌شود. به منظور کنترل همگرایی و برای تولید نمونه‌هایی با وابستگی کمتر، به‌طور معمول به چندین زنجیر نیاز داریم. هر یک از زنجیرها شامل دنباله‌ای از

[۵۹] Burn-in

نمونه‌های تولیدشده از توزیع پسین هستند. بعد از تولید نمونه، همگرایی مدل را بررسی می‌کنیم (لان و همکاران، ۲۰۰۰؛ اندرو و فینلی، ۲۰۱۳). با توجه به مراحلی که عنوان شد، دستورالعمل کلی اجرای یک برنامه در $JAGS$ به‌صورت زیر خواهد بود:

```
data <- list(x=x, y=y, n=length(x))
inits <- list( list( var1, var2, var3, var4 ),
               list( var1, var2, var3, var4 ),
               list( var1, var2, var3, var4 ) )
cat( " model {
##Likelihood
...
##Priors
...
}", fill=TRUE, file="model.jag " )
mod <- jags.model("model.jag", data=data,
                               inits=inits,
                               n.chains=3
                               n.adapt=1000)
out <- coda.samples(mod,
                               var=c("var1", "var2", "var3", "var4")
                               n.iter=10000,
                               thin=1)
summary(out)
```

۸.۱ مفاهیم

اگر داده‌ها محدود به فاصله $(۰, ۱)$ و پیوسته باشند، مدلی مناسب برای برازش، مدل رگرسیون بتا است. در ادامه، توزیع بتا را تعریف کرده و در فصل‌های بعد مدل رگرسیون بتا و در پی آن مدل رگرسیون مستطیلی بتا را عنوان می‌کنیم. در پی آنچه تا کنون ذکر شد، به تعریف داده پرت و تنومندی مدل می‌پردازیم. در مباحث مرتبط با رگرسیون یا طرح آزمایش‌ها، با مفهومی روبرو می‌شویم، که به آن داده پرت می‌گویند. در حقیقت، در یک نمونه ممکن است تعدادی از داده‌ها نسبت به سایر مشاهدات انحراف زیادی داشته باشند و این انحراف به اندازه‌ای است که این تفکر را در ذهن افراد ایجاد می‌کند که شاید این داده‌ها از جامعه دیگری آمده باشند. بررسی‌ها نشان می‌دهد که مدل رگرسیون بتا نسبت به داده‌های پرت نیرومند نیست. پس باید مدل جدیدی معرفی گردد که متغیر پاسخ در آن در بازه $(۰, ۱)$ تعریف گردد. مدلی که معرفی می‌شود باید نسبت به داده‌های پرت عملکرد نیرومندی داشته باشد. این مدل در

فصل‌های آتی با نام مدل رگرسیون مستطیلی بتا معرفی می‌گردد. همان‌طور که می‌دانیم پس از معرفی هر مدل باید به تحلیل آن مدل بپردازیم و صحت مدل پیشنهادی را تأیید نماییم. جهت پرداختن به این مدل‌ها نیازمند تعریف توزیع بتا هستیم.

۱.۸.۱ توزیع بتا

توزیع بتا[۶۰] توزیع احتمالی پیوسته‌ای است که، بر بازه $[a,b]$ تعریف می‌شود. تابع چگالی احتمال آن به‌صورت

$$p_Y(y)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(y-a)^{\alpha-1}(b-y)^{\beta-1}(b-a)^{\alpha+\beta-1},\quad a<y<b,\quad \alpha,\beta>0$$

است، که در آن α و β پارامترهای شکل[۶۱] هستند (جان و کوپر، ۲۰۰۷). البته با استفاده از تغییر متغیر $\frac{y-a}{y-b}$ می‌توان توزیع بتا را به‌صورت بتای استاندارد نوشت که توزیع آشنایی می‌باشد. این توزیع، حالت خاصی از توزیع دریکله است. این توزیع، برای مدل‌هایی استفاده می‌شود که درباره، احتمال موفقیت یک آزمایش است. نسبت‌ها[۶۲] و درصدها[۶۳]، در این دسته جای دارند. تابع چگالی احتمال بتای استاندارد به شکل زیر است:

$$p_Y(y)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1},\quad 0<y<1$$

که در آن تکیه‌گاه پارامترها، مقادیر مثبت هستند. از این پس، بتای استاندارد را بتا می نامیم. در صورتی که $\alpha=\beta$ باشد، توزیع بتا متقارن است. هر چه α و β بزرگتر باشند شکل توزیع کشیده‌تر و واریانس آن کمتر خواهد شد (بومن و شنتن، ۲۰۰۷). اگر $\alpha=\beta=1$ باشد، آنگاه یک توزیع یکنواخت در فاصله $[0,1]$ خواهیم داشت. اگر $\alpha\neq\beta$ باشد توزیع به‌صورت چوله درمی‌آید. در صورتی که پارامتر اول از پارامتر دوم بزرگتر باشد، چوله به سمت چپ خواهد بود. توزیع بتا، یک توزیع انعطاف‌پذیر است. زیرا تابع چگالی احتمال آن به ازای مقادیر مختلف پارامترها، می‌تواند شکل‌های متفاوتی داشته باشد. میانگین و واریانس این توزیع به‌صورت زیر هستند:

$$E(Y)=\frac{\alpha}{\alpha+\beta},\quad Var(Y)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

مد توزیع، زمانی که α و β بزرگتر از یک هستند، موجود و به‌صورت

$$mod(y)=\frac{(\alpha-1)}{(\alpha+\beta-2)}$$

است. کاربردهای توزیع بتا توسط باری (۱۹۹۹) بحث شده است.

---

[۶۰] Beta Distribution

[۶۱] Shape Parameter

[۶۲] Proportion

[۶۳] Precentage

۲.۸.۱ داده‌های پرت

داده‌های پرت به دلایل زیادی مشاهده می‌شوند. اما، معمولاً نشانگر این هستند که یا برخی از اندازه‌گیری‌ها دارای اشتباه‌اند، یا داده‌ها دارای توزیع احتمالی با دم سنگین هستند.

تعریف ۱.۸.۱. داده پرت مشاهده‌ای است که به‌طور غیرعادی یا اتفاقی، از وضعیت عمومی داده‌های تحت آزمایش و نسبت به قاعده‌ای که بر اساس آن آزمایش می‌شوند، انحراف داشته باشد. فرگوسن (۱۹۶۱) اعتقاد دارد در یک نمونه با اندازه اصلاح‌شده از یک جامعه، معمولاً یک یا دو مشاهده به‌طور شگفت‌آوری از گروه اصلی فاصله دارند، که از روند بقیه داده‌ها پیروی نمی‌کنند.

میلر (۱۹۸۱) نیز بیان کرد که داده پرت یک مشاهده یا میانگین ساده است که از روند بقیه داده‌ها پیروی نمی‌کند.

برخی مشاهدات نسبت به سایرین پراکندگی بیشتری دارند. همین امر موجب بروز خطا می‌گردد. در حالی که در مباحث نظری این خطاها در نظر گرفته نمی‌شوند. اگر تشخیص دهیم که تعدادی از داده‌ها پرت هستند، گرچه به راه تشخیصی مطمئن نباشیم، تحلیل داده‌ها به ما کمک می‌کند. با توجه به تعریف داده‌های پرت، تشخیص آن‌ها با روش‌های آزمون فرضیه ممکن نمی‌باشد، زیرا که داده‌های پرت ممکن است از توزیعی همانند توزیع بقیه مشاهدات، ولی با یک یا چند پارامتر متفاوت آمده باشند، یا این که از توزیع متفاوتی پیروی کنند. در صورتی که بدانیم داده‌ها هم‌توزیع هستند، می‌توان با روش‌هایی داده‌های پرت را تشخیص داد.

ساده‌ترین راهی که برای تشخیص داده‌های پرت به نظر می‌رسد، این است که قرارداد کنیم انحراف بیش از سه انحراف استاندارد را باید داده پرت در نظر گرفت. باید به این نکته توجه داشت، این روش برای داده‌های چوله مناسب نیست. برای داده‌هایی با توزیع‌های چوله، نقاط برشی مطرح می‌گردند. هنگامی که توزیع‌های چند متغیری داریم، شاخص‌هایی از فاصله باقیمانده، استانداردشده یا استیودنت شده، مربوط به رگرسیون مطرح می‌شود که می‌توانند در شناسایی نقاط پرت مفید باشند.

۳.۸.۱ تنومندی مدل

بحث تنومندی و روش‌های رگرسیون تنومند زمانی‌که مشاهدات غیرمعمولند، یا توزیع‌هایی که متقارن نیستند، مطرح می‌گردد (مارکاتو، ۱۹۹۸ و ۱۹۹۶). هنگامی‌که رگرسیون مؤثر کرانداری یا $M-$برآورد مطرح باشد به‌طور خاص روی تنومندی تأکید می‌شود. تحت هابر (۱۹۸۱) تنومندی، عدم حساسیت به انحرافات کوچک از پذیره‌های مدل مورد نظر تعریف می‌شود. به‌صورت خاص به نیرومندی توزیعی علاقه‌مندیم و تأثیر توزیع‌های نامتقارن یا نقاط پرت در برآورد رگرسیون مدنظر ما است. همچنین تنومندی در توصیف خطاهای استاندارد که واریانس خطا ثابت نیست، مطرح می‌گردد. نیرومندی به شکل توزیع، که به‌صورت نظری فرض شده است، ترجیح داده می‌شود.

استنباط‌های آماری بر اساس مشاهدات و پذیره‌های پیشین درباره توزیع‌های تحت مطالعه و روابط بین متغیرها است. همچنین این پذیره‌ها با خطا همراه هستند. یک مدل اگر خصوصیات زیر را داشته باشد، نیرومند است:

- کارایی نسبی و ناارریبی
- انحرافات کوچک از پذیره‌های مدل، اساساً عملکرد مدل را ناقص نمی‌کند.
- گاهی انحرافات بزرگتر کاملاً مدل را نامعتبر نمی‌کند.

۴.۸.۱ آمیخته‌های گسسته

متغیر تصادفی Y دارای توزیع آمیخته گسسته است، اگر توزیع آن، برای دنباله‌ای از متغیرهای تصادفی $Y_۱, Y_۲, \cdots$ به‌صورت مجموع وزنی

$$p(Y \leqslant y) = \sum p_i p(Y_i \leqslant y)$$

باشد، به‌طوری که $\sum p_i = ۱$. به ثابت‌های p_i ضرایب یا وزن‌های آمیختگی گویند.

آمیخته گسسته دومؤلفه‌ای

متغیر تصادفی Y دارای توزیع آمیخته گسسته دومؤلفه‌ای است، اگر توزیع آن، برای دو متغیر تصادفی $Y_۱$ و $Y_۲$ به‌صورت مجموع وزنی

$$p(Y \leqslant y) = p_۱ p(Y_۱ \leqslant y) + p_۲ p(Y_۲ \leqslant y)$$

باشد به‌طوری که $p_۱ + p_۲ = ۱$. نمایش توزیع‌های آمیخته را با تابع چگالی نیز می‌توان نمایش داد.

۵.۸.۱ اندازه واگرایی کولبک-لیبلر

در احتمال، واگرایی کولبک-لیبلر [۶۴] اندازه‌ای نامتقارن بین توزیع‌های احتمال P و Q است. این اندازه ویژگی‌های یک متر را دارا نیست. بنابراین اگر به‌دنبال متری برای سنجش فاصله بین دو توزیع باشیم، از آن نمی‌توان استفاده کرد. به‌عبارتی بر اساس این معیار فاصله P از Q برابر فاصله Q از P نیست. واگرایی کولبک-لیبلر حالت خاصی از یک رده گسترده‌تر از واگرایی‌ها است، که Pواگرایی نامیده می‌شوند. واگرایی کولبک-لیبلر در ابتدا توسط سالمون کولبک و ریچارد لیبلر در سال ۱۹۵۱ تحت عنوان واگرایی جهت‌دار بین دو توزیع معرفی شد. واگرایی جهت‌دار از واگرایی برگمن استنتاج شده بود (کولبک و لیبلر، ۱۹۵۱).

واگرایی کولبک-لیبلر برای توزیع‌های گسسته و پیوسته قابل تعریف است. اما در این‌جا تنها به تعریف آن در حالت پیوسته می‌پردازیم.

[۶۴] Kullback-Leibler

تعریف ۲.۸.۱. برای توزیع‌های پیوسته P و Q، معیار واگرایی کولبک-لیبلر Q از P به‌صورت زیر تعریف می‌شود:

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(y) ln \frac{p(y)}{q(y)} dy$$

که در آن $p(\cdot)$ و $q(\cdot)$ توابع چگالی P و Q می‌باشند. به عبارت دیگر، این همگرایی امید لگاریتم اختلاف بین احتمال‌های P و Q است، وقتی که امید بر اساس احتمالات P باشد.

واگرایی کولبک-لیبلر بیزی

در آماره‌های بیزی، واگرایی کولبک-لیبلر می‌تواند اندازه‌ای از اطلاع هدف، برای تبدیل توزیع پیشین به توزیع پسین باشد. این فاصله می‌تواند تبدیلی از توزیع احتمال Y از $p(y)$ به توزیع احتمال پسین $p(y \mid x)$ با استفاده از قضیه بیز باشد:

$$p(y \mid x) = \frac{p(x \mid y) p(y)}{p(x)}.$$

در واگرایی کولبک-لیبلر داریم

$$D_{KL}(p(y|x) \mid p(y)) = \int p(y|x) ln \frac{p(y|x)}{p(y)}.$$

در واقع هدف کلی در این‌جا ماکسیمم‌سازی واگرایی کولبک-لیبلر بین توزیع پیشین و پسین است (کنه و همکاران، ۲۰۰۱؛ اندرسون و همکاران، ۲۰۰۱).

معمولاً چون به دنبال معرفی پیشینی هستیم که تأثیر کمتری در استنباط داشته باشد از منظر این معیار، پیشینی مطلوب است که اندازه KL آن با پسین حاصل زیاد باشد.

فصل ۲

مدلی تنومند برای پاسخ‌های پیوسته محدود به (۰, ۱)

در موارد متعددی برای تحلیل پاسخ‌های نرخ و نسبت، مانند نرخ رشد، علاقه‌مند به کشف عوامل مؤثر بر متغیر پاسخ هستیم. معمولاً با مدل‌های رگرسیونی می‌توان این عوامل را بررسی کرد. تفسیر ضرایب رگرسیون، به مقیاس داده‌ها حساس است. اگر ضرایب را در میان مجموعه‌هایی از داده تغییر مقیاس دهیم، مقایسه آن‌ها گیج‌کننده می‌شود (کینگ، ۱۹۸۶). زیرا با تغییر بازه تغییرات یک پیشگو، ضریب آن تغییر مقیاس می‌دهد، حتی اگر مدل رگرسیونی تغییر نکرده باشد. بلالک (۱۹۶۱) نیز به چالش‌های مقایسه ضرایب و این‌که تغییر مقیاس در رگرسیون لازم است یا نه، اشاره دارد. همچنین گرینلند و همکاران (۱۹۸۶) در مورد چالش‌هایی که در مورد ضرایب رگرسیون استانداردشده مطرح است، بحث کرده‌اند.

اما فرآیند استانداردسازی می‌تواند برای فهم رگرسیون ابزاری مفید باشد. در این صورت متغیر پاسخ در بازه (۰, ۱) قرار می‌گیرد. برای تحلیل آن‌ها می‌توان از مدل‌های رگرسیونی لجیت و پروبیت استفاده کرد. اما از آن‌جا که این نوع پاسخ‌ها معمولاً به‌شدت چوله هستند، استفاده از این مدل‌ها برای این پاسخ‌ها مناسب نیستند. رگرسیون بتا که در این فصل مورد بحث واقع شده، روشی مناسب برای مدل‌بندی پاسخ‌هایی با دامنه تغییرات (۰, ۱) است.

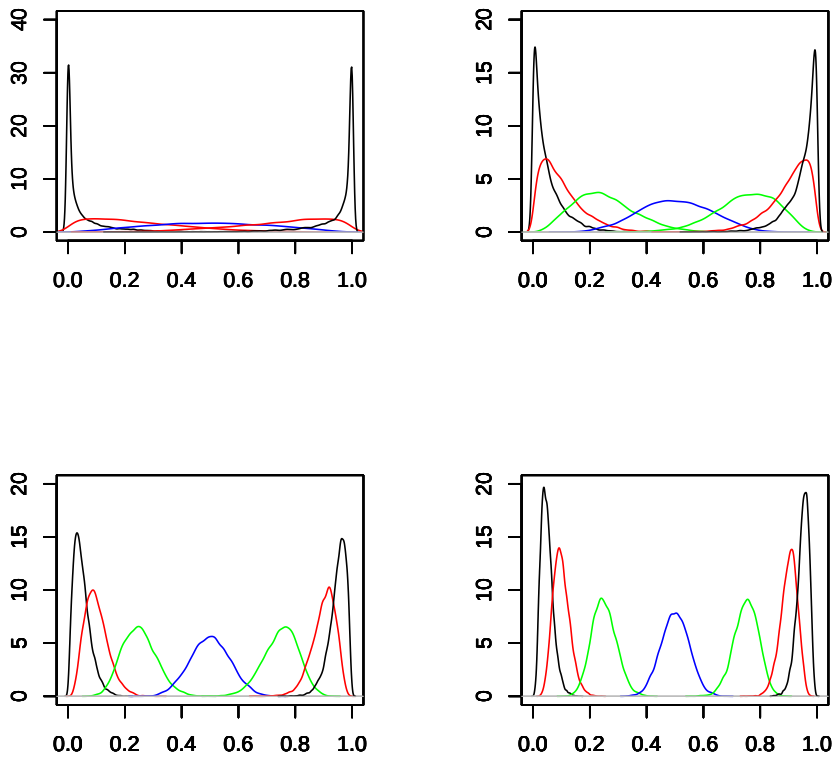
بسیاری از روش‌های استنباط آماری به وجود داده‌های پرت حساس هستند، از این رو علاقه‌مند به استفاده از روش‌هایی هستیم که نسبت به داده‌های پرت منعطف باشند. مدل رگرسیون مستطیلی بتا برای داده‌های پیوسته در بازه (۰, ۱) که در این فصل معرفی می‌شود، مدلی است که نسبت به نقاط پرت تنومند است.

۱.۲ توزیع بتا

برای بیان رگرسیون بتا به تعریف چگالی بتا نیاز است. در صورتی که تابع چگالی متغیر تصادفی Y را به صورت

$$p(y;\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{(\alpha-1)}(1-y)^{(\beta-1)},\quad 0<y<1,\quad \alpha,\beta>0$$

در نظر بگیریم، آنگاه Y دارای توزیع بتا با پارامترهای α و β است و آن را با نماد $Y\sim Beta(\alpha,\beta)$ نمایش می‌دهیم. توزیع بتا توزیع معروفی است که دارای دو پارامتر شکل است. این توزیع تابع چگالی منعطفی دارد و می‌توان شکل‌های مختلفی از آن را با در نظر گرفتن مقادیر مختلف α و β داشت. همان‌طور که در شکل ۱.۲ می‌بینید تابع چگالی بتا به ازای مقادیر مختلفی از پارامترهای خود رسم شده



شکل ۱.۲: نمودارهای توابع چگالی بتا به ازای مقادیر متفاوت α و β

است. شکل این تابع چگالی می‌تواند کشیده یا پخ، متقارن یا چوله باشد. در نتیجه برای داده‌های مورد نظر در این پایان‌نامه که بسیار چوله هستند، مناسب می‌باشد.

همان‌طور که در هان (۲۰۰۸) و گارسیا (۲۰۱۱) آمده است، با در نظر گرفتن واریانس مشخص، توزیع بتا در ناحیه دم برای پارامترهای متفاوت انعطاف‌پذیری زیادی از خود نشان می‌دهد. با در نظر

گرفتن این مطلب می‌توانیم از آن برای مدل‌سازی نسبت‌ها استفاده کنیم.

در مدل رگرسیونی زمانی که می‌خواهیم تأثیر متغیرهای تبیینی را روی متغیر پاسخ ببینیم، باید میانگین پاسخ را مدل کنیم. بنابراین، برای تعریف مدل رگرسیونی بر حسب توزیع بتا نیازمند استفاده از شکل متفاوتی از توزیع بتا هستیم که پارامترهای آن طبق میانگین توزیع و پارامتر دیگری معروف به دقت باشد. این نمایش سبب می‌شود تا بتوان میانگین توزیع را به متغیرهای تبیینی مرتبط کرد. بر این اساس، توزیع بتا را با جایگزاری عبارات $\mu = \frac{\alpha}{\alpha+\beta}$ و $\phi = \alpha + \beta$ بازپارامتری [۱] می‌کنیم که می‌توان $\alpha = \mu\phi$ و $\beta = (۱-\mu)\phi$ را از این عبارات به‌دست آورد. بازنویسی توزیع بتا توسط فراری و سریباری (۲۰۰۴) ارائه شد، به‌طوری که تابع چگالی بازنویسی‌شده به‌صورت زیر است:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((۱-\mu)\phi)} y^{(\mu\phi-۱)}(۱-y)^{((۱-\mu)\phi-۱)}, \quad ۰ < y < ۱ \qquad (۱.۲)$$

که در آن $۰ < \mu < ۱$ و $\phi > ۰$. متغیر تصادفی Y دارای توزیع بتا بازنویسی‌شده است که آن را با نماد $Y \sim Beta(\mu,\phi)$ نشان می‌دهیم و تابع چگالی آن را با نماد $b(y \mid \mu,\phi)$ و از این پس به آن توزیع بتا می‌گوییم. با استفاده از تعریف مستقیم امید ریاضی، به راحتی می‌توان بررسی کرد که میانگین توزیع بتا با تابع چگالی (۱.۲) همان μ است:

$$\begin{aligned}
E(Y|\mu,\phi) &= \int_۰^۱ \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((۱-\mu)\phi)} y^{\mu\phi-۱}(۱-y)^{(۱-\mu)\phi-۱} y dy \\
&= \frac{\Gamma(\mu\phi+۱)\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi+۱)} \underbrace{\int_۰^۱ \frac{\Gamma(\phi+۱)}{\Gamma(\mu\phi+۱)\Gamma((۱-\mu)\phi)} y^{\mu\phi}(۱-y)^{(۱-\mu)\phi-۱} dy}_{=۱} \\
&= \frac{\mu\phi\Gamma(\mu\phi)\Gamma(\phi)}{\Gamma(\mu\phi)\phi\Gamma(\phi)} \\
&= \frac{\mu\phi}{\phi} \\
&= \mu. \qquad (۲.۲)
\end{aligned}$$

توزیع بتا با مقادیر متفاوتی که برای پارامترهای (μ,ϕ) در نظر گرفته شده، بازپارامتری می‌شود. در این صورت، شکل‌های متفاوتی از توابع چگالی توزیع بتا خواهیم داشت. با تغییر مقادیر پارامتر μ شکل‌های متقارن ($\mu = \frac{۱}{۲}$) و شکل‌های نامتقارن ($\mu \neq \frac{۱}{۲}$) از توزیع بتا خواهیم داشت.

برای محاسبه واریانس توزیع بازپارامتری‌شده از تعریف مستقیم واریانس استفاده می‌کنیم:

$$Var(Y \mid \mu,\phi) = E(Y^۲ \mid \mu,\phi) - E^۲(Y \mid \mu,\phi). \qquad (۳.۲)$$

ابتدا $E(Y^۲ \mid \mu,\phi)$ را محاسبه می‌نماییم:

[۱] Reparametrization

$$E(Y^{۲} \mid \mu, \phi) = \int_{۰}^{۱} \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((۱-\mu)\phi)} y^{\mu\phi-۱}(۱-y)^{(۱-\mu)\phi-۱} y^{۲} dy.$$

با ضرب و تقسیم عبارت $\Gamma(\mu\phi+۲)\Gamma(\phi+۲)$:حاصل انتگرال بالا به صورت زیر به‌دست می‌آید

$$\begin{aligned}
&= \frac{\Gamma(\mu\phi+۲)\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi+۲)} \underbrace{\int_{۰}^{۱} \frac{\Gamma(\phi+۲)}{\Gamma(\mu\phi+۲)\Gamma((۱-\mu)\phi)} y^{\mu\phi+۱}(۱-y)^{(۱-\mu)\phi-۱} dy}_{=۱} \\
&= \frac{(\mu\phi+۱)(\mu\phi)\Gamma(\mu\phi)\Gamma(\phi)}{\Gamma(\mu\phi)(\phi+۱)\phi\Gamma(\phi)} \\
&= \frac{\mu(\mu\phi+۱)}{\phi+۱} \\
&= \frac{\mu^{۲}\phi+\mu}{\phi+۱}. \qquad (۴.۲)
\end{aligned}$$

آن‌گاه با استفاده از (۲.۲)، (۳.۲) و (۴.۲) داریم

$$\begin{aligned}
Var(Y|\mu,\phi) =& E(Y^{۲} \mid \mu, \phi) - E^{۲}(Y \mid \mu, \phi) \\
&= \frac{\mu^{۲}\phi+\mu}{\phi+۱} - \mu^{۲} \\
&= \frac{\mu^{۲}\phi+\mu-\mu^{۲}\phi-\mu^{۲}}{\phi+۱} \\
&= \frac{\mu(۱-\mu)}{\phi+۱}. \qquad (۵.۲)
\end{aligned}$$

با توجه به رابطه (۵.۲) با در نظر گرفتن μ ثابت، هر چه ϕ افزایش پیدا کند، واریانس بتا کاهش می‌یابد. پس ϕ را می‌توان به‌عنوان پارامتری برای ارزیابی دقت مدل مورد استفاده قرار داد. در این فصل به دنبال معرفی یک توزیع جدید هستیم. این توزیع آمیخته‌ای از یک توزیع بتا و توزیع یکنواخت استاندارد است. دامنه تعریف این توزیع در بازه (۰, ۱) قرار می‌گیرد. سپس یک مدل رگرسیونی بر اساس این توزیع تعریف می‌کنیم. پارامترهای متغیر پاسخ در این مدل رگرسیونی "دقت" و "میانگین" هستند. قبل از آن ابتدا مدل رگرسیون بتا را معرفی می‌کنیم:

۲.۲ مدل رگرسیون بتا

فراری و سربیاری (۲۰۰۴) برای اولین بار مدلی را پیشنهاد کردند که در آن متغیر پاسخ با توزیع بتا و تابع چگالی (۱.۲) در چارچوب مدل‌های خطی تعمیم‌یافته با استفاده از میانگین پاسخ، μ، و یک پارامتر مثبت به نام پارامتر دقت، ϕ، مدل‌سازی شده است.

در چارچوب مدل‌های خطی تعمیم‌یافته و بر مبنای توزیع بازپارامتری‌شده بتا (۱.۲)، تابع چگالی متغیر پاسخ y را می‌توان با نماد$Y \sim Beta(\mu\phi, (۱-\mu)\phi)$ نشان داد.

اگر $Y_1, \cdots, Y_n$، n متغیر تصادفی مستقل با میانگین $\mu_1, \cdots, \mu_n$ باشند، می‌نویسیم $Y_i \sim Beta(\mu\phi_i, (1-\mu)\phi_i)$. بنابراین مدل رگرسیون بتا، تبدیلی برای میانگین μ_i و پارامتر دقت ϕ_i انجام می‌دهد که می‌توان آن را به‌صورت

$$g_1(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}, \qquad g_2(\phi_i) = \mathbf{w}_i^T\boldsymbol{\delta} \tag{۶.۲}$$

نوشت که به‌ترتیب $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^T$ و $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^T$ بردارهای ضرایب رگرسیونی نامعلوم و $\mathbf{x}_i = (x_1, \ldots, x_n)^T$ و $\mathbf{w}_i = (w_1, \ldots, w_n)^T$ بردارهای متغیرهای تبیینی معلوم برای iامین عضو و $g_1(\cdot)$ و $g_2(\cdot)$ توابع پیوند یکنوا و مشتق‌پذیر می‌باشند.

رده GLM نیازمند تعریف تابع پیوند برای میانگین پاسخ و پارامتر دقت می‌باشد. اغلب، تابع پیوند مناسب برای میانگین پاسخ که بتواند پیشگوی خطی را به بازه $(0, 1)$ بنگارد، تابع پیوند لجیت انتخاب می‌شود. به‌عبارتی

$$logit(\mu_i) = \ln(\frac{\mu_i}{1-\mu_i}) = \mathbf{x}_i^T\boldsymbol{\beta}.$$

به‌طور مشابه، یک تابع پیوند مناسب برای پارامتر دقت، که یک پارامتر اکیداً مثبت است، تابع پیوند لگاریتمی، $g_2(\phi_i) = log(\phi_i)$، می‌باشد. این پارامتر می‌تواند ثابت در نظر گرفته شود (فراری و سریباری، ۲۰۰۴) یا در یک ترکیب رگرسیونی مشابه میانگین، مانند $log(\phi_i) = \mathbf{w}_i^T\boldsymbol{\delta}$ مدل‌سازی شود (اسمیتسون و ورکوی‌لن، ۲۰۰۶).

۳.۲ توزیع مستطیلی بتا

با نظر در متون آماری، درمی‌یابیم که در نواحی دم توزیع‌ها پیشامدها با تنوع فراوانی اتفاق می‌افتند. به همین سبب توزیع‌ها نمایانگر دنیای واقعی نخواهند بود. برای بهبود این شرایط، توزیع‌های مستطیلی مطرح می‌گردند (مارکتوا، ۲۰۰۰). معمولاً، آمیخته‌ای از یک توزیع دم باریک و توزیع دیگری که در ناحیه خاص دم‌ها پهن‌تر بوده و انعطاف‌پذیری بیشتری نسبت به نقاط پرت دارد، در نظر گرفته می‌شود (گلمن و همکاران، ۲۰۰۳). در همین راستا، اگر بخواهیم انعطاف‌پذیری توزیع بتا را افزایش دهیم و یک مدل رگرسیونی با مقادیر مختلف دقت و درستنمایی بزرگتر داشته باشیم که پیشامدهای زیادی را در نواحی دمی ایجاد می‌کند، با استفاده از توزیع مستطیلی بتا، امکان‌پذیر است. این توزیع، اولین بار توسط هان در سال ۲۰۰۸ مطرح شد.

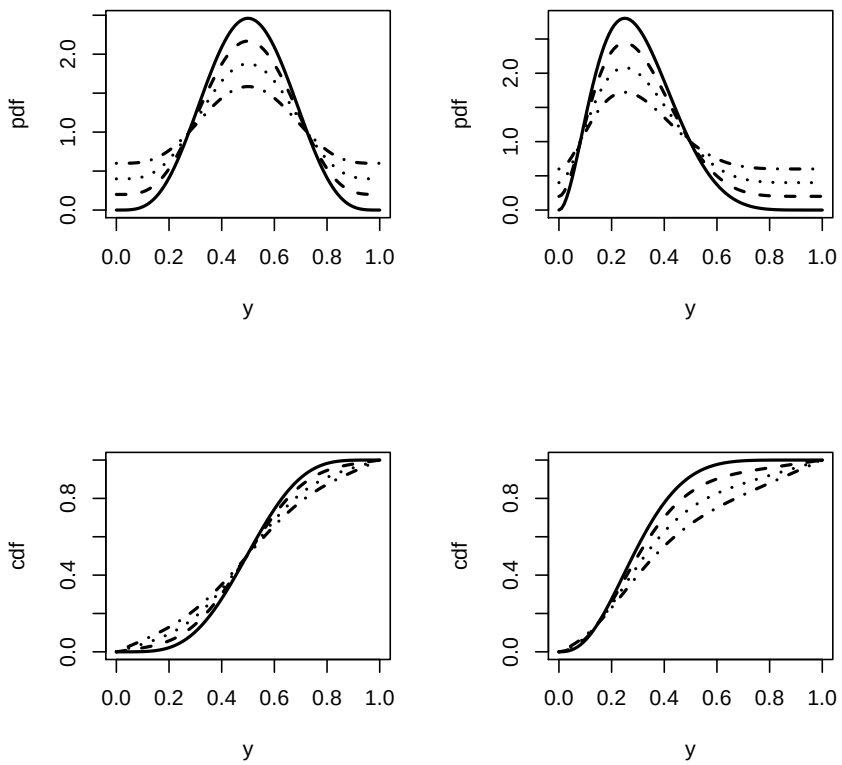
تابع چگالی توزیع مستطیلی بتا به‌صورت زیر است:

$$g(y \mid \mu, \phi, \theta) = \theta + (1-\theta)b(y \mid \mu, \phi) \tag{۷.۲}$$

که در آن $0 \leq \theta \leq 1$ پارامتر آمیختگی[۲] است. توزیع مستطیلی بتا آمیخته‌ای از دو توزیع بتا است

[۲] Mixture Parameter

که یکی از آن‌ها توزیع $Beta(\frac{۱}{۲}, ۲)$ است. در واقع، این توزیع همان توزیع یکنواخت استاندارد است. اگر متغیر تصادفی Y دارای تابع چگالی آمیخته (۷.۲) باشد، آن را با نماد $Y \sim BR(\mu, \phi, \theta)$ نشان می‌دهیم. با توجه به شکل تابع چگالی آمیخته، واضح است زمانی که $\theta = ۱$ باشد، توزیع یکنواخت نتیجه می‌شود و هنگامی‌که $\theta = ۰$، توزیع مستطیلی بتا، همان توزیع بتا است. شکل ۲.۲ منحنی توابع چگالی و توزیع تجمعی مستطیلی بتا را به ازای مقادیر مختلفی از μ و θ نمایش می‌دهد.



شکل ۲.۲: نمودارهای سمت چپ مربوط به توزیع مستطیلی بتا برای مقادیر $\mu = ۰/۵$ و $\phi = ۱۰$ است و نمودارهای سمت راست برای $\mu = ۰/۳$ و $\phi = ۱۰$ است. پارامتر آمیختگی برای نمودار خط $\theta = ۰$ و برای خط چین $\theta = ۰/۲$ وبرای نقطه‌چین $\theta = ۰/۴$ و برای نقطه خط‌چین $\theta = ۰/۶$ است. نمودارهای بالا توابع چگالی و نمودارهای پایین توابع توزیع هستند.

با توجه به شکل ۲.۲درمی‌یابیم که هرچه پارامتر آمیختگی بزرگتر می‌شود، رفته رفته شکل نمودار از توزیع بتا به توزیع یکنواخت می‌گراید. همان‌طور که در شکل ملاحظه می‌کنید، شکل توزیع بتا که با خط نمایش داده شده در ناحیه دم بسیار باریک بوده و با مشاهده داده در این قسمت توزیع با احتمال بالایی، نقطه پرت است. این در حالی است که با افزایش پارامتر آمیختگی در توزیع مستطیلی بتا، نقاط پرت در دم بتا با توزیع یکنواخت استاندارد جایگزین می‌شود. توزیع مستطیلی بتا در ناحیه دم نسبت به توزیع بتا پهن بوده و با احتمال کمتری مشاهدات در ناحیه دم توزیع مستطیلی بتا نقاط پرت هستند. پس، توزیع مستطیلی بتا نسبت به بتا انعطاف زیادی نسبت به نقاط پرت دارد.

۱.۳.۲ گشتاورهای اول و دوم توزیع مستطیلی بتا

حال برای بررسی بیشتر توزیع مستطیلی بتا، به معرفی میانگین آن با استفاده از تعریف امید ریاضی می‌پردازیم:

$$E(Y \mid \mu,\phi,\theta) = \int \left[\theta + (۱-\theta)\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((۱-\mu)\phi)} y^{\mu\phi-۱}(۱-y)^{(۱-\mu)\phi-۱}\right] y dy$$
$$= \int \theta y dy + \int (۱-\theta)\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((۱-\mu)\phi)} y^{\mu\phi-۱}(۱-y)^{(۱-\mu)\phi-۱} y dy.$$

با ضرب و تقسیم عبارت $\Gamma(\mu\phi+۱)\Gamma(\phi+۱)$ در انتگرال دوم، نتیجه می‌شود

$$= \int \theta y dy + (۱-\theta)\underbrace{\frac{\Gamma(\mu\phi+۱)\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi+۱)}}_{=\mu} \underbrace{\int \frac{\Gamma(\phi+۱)}{\Gamma(\mu\phi+۱)\Gamma((۱-\mu)\phi)} y^{\mu\phi}(۱-y)^{(۱-\mu)\phi-۱} dy}_{=۱}$$
$$= \frac{\theta}{۲} + (۱-\theta)\mu. \tag{۸.۲}$$

واریانس توزیع مستطیلی بتا در پی به‌دست آمدن $E(Y^۲ \mid \mu,\phi,\theta)$ حاصل می‌شود

$$E(Y^۲ \mid \mu,\phi,\theta) = \int \left[\theta + (۱-\theta)\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((۱-\mu)\phi)} y^{\mu\phi-۱}(۱-y)^{(۱-\mu)\phi-۱}\right] y^۲ dy$$

مشابه (۸.۲)، با پخش انتگرال بر روی عمل جمع و ضرب و تقسیم عبارت $\Gamma(\mu\phi+۲)\Gamma(\phi+۲)$ در انتگرال دوم، داریم

$$= \theta \int y^۲ dy + (۱-\theta)\underbrace{\frac{\Gamma(\mu\phi+۲)\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi+۲)}}_{\frac{\mu(\mu\phi+۱)}{\phi+۱}} \underbrace{\int \frac{\Gamma(\phi+۲)}{\Gamma(\mu\phi+۲)\Gamma((۱-\mu)\phi)} y^{\mu\phi+۱}(۱-y)^{(۱-\mu)\phi-۱} dy}_{=۱}$$
$$= \frac{\theta}{۳} + (۱-\theta)\frac{\mu(\mu\phi+۱)}{\phi+۱}. \tag{۹.۲}$$

بنابراین بنا به روابط (۵.۲)، (۸.۲) و (۹.۲) نتیجه می‌شود

$$Var(y \mid \mu,\phi,\theta) = \frac{\theta}{۳} + (۱-\theta)\frac{\mu(\mu\phi+۱)}{\phi+۱} - \frac{\theta^۲}{۴} - (۱-\theta)^۲\mu^۲ - \theta(۱-\theta)\mu$$
$$= \frac{(۱-\theta)}{(۱+\phi)}\left[\mu^۲\phi + \mu - \mu^۲(۱-\theta)(۱+\phi) - \theta\mu(۱+\phi)\right] + \theta(\frac{۱}{۳} - \frac{\theta}{۴})$$
$$= \frac{(۱-\theta)}{(۱+\phi)}\left[-\mu^۲(-\phi+۱+\phi-\theta-\theta\phi) + \mu(۱-\theta(۱+\phi))\right] + \frac{\theta}{۱۲}(۴-۳\theta)$$
$$= \frac{(۱-\theta)}{(۱+\phi)}\left[-\mu^۲(۱-\theta(۱+\phi)) + \mu(۱-\theta(۱+\phi)\right] + \frac{\theta}{۱۲}(۴-۳\theta)$$
$$= \frac{(۱-\theta)\mu(۱-\mu)}{(۱+\phi)}(۱-\theta(۱+\phi)) + \frac{\theta}{۱۲}(۴-۳\theta).$$

با توجه به فرمول‌های میانگین و واریانس، در صورتی که $\theta = ۰$ باشد، میانگین و واریانس توزیع آمیخته با میانگین و واریانسی که برای توزیع بتا ذکر کردیم، برابر می‌شود. در صورتی که $\theta = ۱$ باشد، میانگین و واریانس توزیع آمیخته با توزیع یکنواخت استاندارد برابر می‌شود.

۴.۲ بازنویسی توزیع مستطیلی بتا

همان‌طور که در فصل قبل ذکر شد، در تحلیل رگرسیون عموماً میانگین پاسخ مدل می‌شود. میانگین توزیع مستطیلی بتا تابعی از پارامتر آمیختگی θ و میانگین توزیع بتا μ است. اگر $\gamma = \frac{\theta}{۲} + (۱ - \theta)\mu$ را در نظر بگیریم، فضای پارامتر θ به ازای مقدار γ، به صورت زیر است:

$$۰ < \theta < ۱ - |\, ۲\gamma - ۱ \,|.$$

زیرا، با توجه به تعریف γ و این‌که $۰ \leq \theta \leq ۱$ است، همواره

$$\theta < ۲\gamma$$

است. پس می‌توان نوشت

$$\begin{aligned} &\theta - ۱ < ۲\gamma - ۱ \\ &-(۱ - \theta) < ۲\gamma - ۱. \end{aligned} \tag{۱۰.۲}$$

از سوی دیگر، همواره

$$\theta < ۱$$

است، پس داریم

$$\begin{aligned} &\theta < \frac{۲(۱ - \mu)}{۲(۱ - \mu)} \\ &\Rightarrow ۲\theta - ۲\theta\mu < ۲ - ۲\mu \\ &\Rightarrow \theta < ۲ - ۲\mu + ۲\theta\mu - \theta \\ &۲\gamma - ۱ < ۱ - \theta \end{aligned} \tag{۱۱.۲}$$

بنابراین بنا به روابط (۱۰.۲) و (۱۱.۲) نتیجه می‌شود

$$\begin{aligned} &|۲\gamma - ۱| < ۱ - \theta \\ &\Rightarrow \theta < ۱ - |۲\gamma - ۱| \end{aligned}$$

به منظور دستیابی به ساختار رگرسیونی مناسب‌تر این میانگین، توزیع مستطیلی بتا را بازپارامتری می‌کنیم. قرار می‌دهیم

$$\gamma = \frac{\theta}{۲} + (۱ - \theta)\mu, \quad \alpha = \frac{\theta}{۱ - (۱ - \theta) \mid ۲\mu - ۱ \mid}.$$

این بازنویسی جدید بر اساس میانگین توزیع و پارامتر دقت است. در این مورد فضای پارامتر برای دو پارامتر γ و α مجموعه $\{۰ \leq \gamma \leq ۱, ۰ \leq \alpha \leq ۱\}$ است. تحت این بازنویسی با ضرب و تقسیم عبارت $۱ - (۱ - \theta) \mid ۲\mu - ۱ \mid$ و بنا به (۷.۲)، تابع چگالی توزیع مستطیلی بتا به‌صورت زیر به‌دست می‌آید:

$$g(y \mid \gamma, \phi, \alpha) = \left(\frac{\theta}{۱ - (۱ - \theta) \mid ۲\mu - ۱ \mid}\right)(۱ - (۱ - \theta) \mid ۲\mu - ۱ \mid) +$$
$$۱ - \left[\left(\frac{\theta}{۱ - (۱ - \theta) \mid ۲\mu - ۱ \mid}\right)(۱ - (۱ - \theta) \mid ۲\mu - ۱ \mid)\right] b(y \mid \mu, \phi) \qquad (۱۲.۲)$$

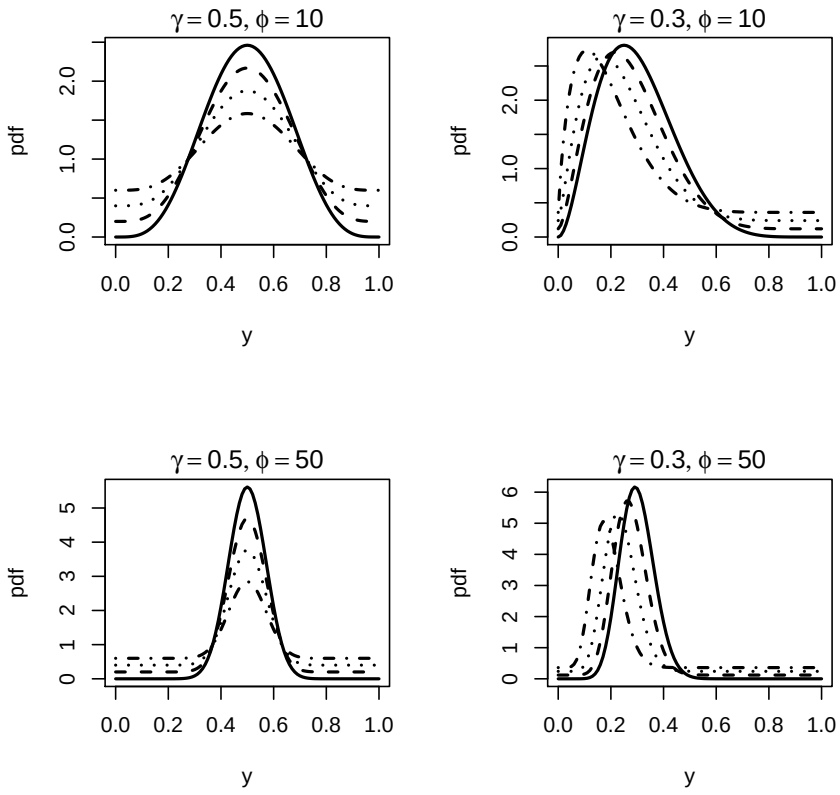
در ادامه، برای بازنویسی توزیع مستطیلی بتا و با ساده کردن عبارت (۱۲.۲) بر اساس (۷.۲) داریم

$$g(y \mid \gamma, \phi, \alpha) = \left(\frac{\theta}{۱ - (۱ - \theta) \mid ۲\mu - ۱ \mid}\right)(۱ - (۱ - \theta) \mid ۲\mu - ۱ \mid) +$$
$$۱ - \left[\left(\frac{\theta}{۱ - (۱ - \theta) \mid ۲\mu - ۱ \mid}\right)(۱ - (۱ - \theta) \mid ۲\mu - ۱ \mid)\right] b(y \mid \mu \times \frac{(۱ - \theta) - \frac{\theta}{۲} + \frac{\theta}{۲}}{۱ - \theta}, \phi)$$
$$= \alpha(۱ - \mid ۲\underbrace{((۱ - \theta)\mu + \frac{\theta}{۲})}_{\gamma} - ۱ \mid) + (۱ - \alpha(۱ - \mid ۲\gamma - ۱ \mid))b(y \mid \frac{\gamma - \frac{\theta}{۲}}{۱ - \theta}, \phi)$$
$$= \alpha(۱ - \mid ۲\gamma - ۱ \mid) + [۱ - \alpha(۱ - \mid ۲\gamma - ۱ \mid)]b(y \mid \frac{\gamma - ۰/۵\alpha(۱ - \mid ۲\gamma - ۱ \mid)}{۱ - \alpha(۱ - \mid ۲\gamma - ۱ \mid)}, \phi).$$

این توزیع جدید را با نماد $Y \sim BRr(\gamma, \phi, \alpha)$ نشان می‌دهیم و از این پس به آن توزیع مستطیلی بتا می‌گوییم. همان‌طور که در شکل توزیع بازنویسی‌شده ملاحظه می‌کنیم، پارامتر آمیختگی ترکیبی از میانگین توزیع مستطیلی بتا پیش از بازنویسی است. همچنین، میانگین توزیع بتایی که با توزیع یکنواخت آمیخته شده، ترکیبی از میانگین توزیع مستطیلی بتا است. نمودارهایی از توزیع BRr برای مقادیر مختلف γ, ϕ و α ارائه می‌دهیم، در شکل ۳.۲، α پارامتر شکل و ϕ پارامتر دقت توزیع است. به‌عبارتی هرچه ϕ افزایش یابد، پراکندگی کمتر می‌شود. در صورتی که $\alpha = ۰$ باشد، توزیع بتا حاصل می‌شود. هرچه α افزایش یابد، توزیع در دم پهن‌تر می‌شود.

۵.۲ مدل رگرسیون مستطیلی بتا

فرض کنید $\mathbf{y} = (y_۱, y_۲, \cdots, y_n)^T$ برداری از پاسخ‌های مشاهده‌شده باشد که مقادیر خود را در بازه $(۰, ۱)$ اختیار می‌کنند. فرض می‌کنیم پاسخ‌های $\mathbf{y}_i$، $i = ۱, ۲, \cdots, n$ از توزیع $BRr(\gamma_i, \phi_i, \alpha)$ پیروی



شکل ۳.۲: مثال‌هایی برای تابع چگالی مستطیلی بتا برای مقادیر مختلف γ, ϕ و α: $\alpha = ۰$ برای خط، $\alpha = ۰/۲$ برای خط‌چین، $\alpha = ۰/۴$ برای نقطه‌چین و $\alpha = ۰/۶$ برای نقطه خط‌چین است.

می‌کنند. مشابه مدل رگرسیونی بتا (۶.۲) دو پیشگوی خطی مدل به‌صورت زیر به میانگین و پارامتر دقت پیوند داده می‌شوند:

$$F_۱^{-۱}(\gamma_i) = \mathbf{x_i}^T \boldsymbol{\beta}$$
$$F_۲^{-۱}(\phi_i) = -\mathbf{w_i}^T \boldsymbol{\delta}$$

که در آن $\boldsymbol{\beta} = (\beta_۱, \beta_۲, \cdots, \beta_k)^T \in \mathbb{R}^k$ و $\boldsymbol{\delta} = (\delta_۱, \delta_۲, \cdots, \delta_l)^T \in \mathbb{R}^l$ بردارهای ضرایب رگرسیونی نامعلوم و $\mathbf{x_i} = (x_{i۱}, x_{i۲}, \cdots, x_{ik})^T$ و $\mathbf{w_i} = (w_{i۱}, w_{i۲}, \cdots, w_{il})^T$ بردارهای متغیرهای تبیینی مشاهده‌شده به‌ترتیب k و l بعدی هستند. توابع $F_۱^{-۱}(\cdot) : (۰, ۱) \to R$ و $F_۲^{-۱}(\cdot) : R^+ \to R$ توابع پیوند یکنوا و مشتق‌پذیر هستند. به‌طورکلی اگر $F_۱(\cdot)$ می‌تواند هر تابع توزیع تجمعی پیوسته باشد. معکوس این تابع توزیع تجمعی، تابع پیوند برای میانگین پاسخ γ_i است که می‌تواند هر پیشگوی خطی را به بازه $(۰, ۱)$ بنگارد. مشابه مدل بتا (۶.۲) مناسب‌ترین تابع پیوند در این شرایط، تابع پیوند لجیت است. به عبارت دیگر

$$logit(\gamma_i) = ln(\frac{\gamma_i}{۱ - \gamma_i}) = \mathbf{x_i}^T \boldsymbol{\beta}.$$

همچنین $F_{۲}^{-۱}(\cdot)$ تابع پیوندی است که پارامتر دقت با دامنه مثبت را به متغیرهای تبیینی $\mathbf{w_i}$ پیوند می‌دهد. تابع پیوند معمول برای این پارامتر نیز تابع لگاریتمی است. بنابراین

$$log(\phi_i) = -\mathbf{w_i}^T \boldsymbol{\delta}.$$

علت استفاده از علامت منفی برای پارامترهای $\boldsymbol{\delta}$ این است که اگر δ_j مثبت باشد، واریانس کوچک می‌گردد که بنا بر مفهوم پارامت دقت ϕ_i می‌تواند گمراه‌کننده باشد. معمولاً، پراکندگی مدل بیشتر از دقت است، پس در جهت منفی از آن استفاده می‌کنیم (اسمیتسون و ورکوی‌لن، ۲۰۰۶).

اگر $\alpha = ۰$ و ϕ_i ثابت باشند، همان مدل رگرسیون بتا به‌دست می‌آید. چنانچه $\alpha = ۰$ و ϕ_i ثابت نباشد، مدل رگرسیونی بتا با پارامتر دقت متغیر نتیجه می‌شود (پاألینو، ۲۰۰۱؛ اسمیتسون و ورکوی‌لین، ۲۰۰۶؛ سیماس و همکاران، ۲۰۱۰).

تابع درستنمایی برای نمونه nتایی و طبق (۷.۲) و (۱۲.۲) به‌صورت زیر تعریف می‌شود:

$$L(\boldsymbol{\psi} \mid \mathbf{y}) = \prod_{i=۱}^{n} g(y_i \mid \mu_i, \phi_i, \theta_i) = \prod_{i=۱}^{n} g(y_i \mid \gamma_i, \phi_i, \alpha) \qquad (۱۳.۲)$$

که در آن $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\delta}, \alpha)$ و همچنین، با توجه به شکل تابع چگالی (۱۲.۲)، $\mu_i = \frac{\gamma_i - \frac{\theta_i}{۲}}{۱-\theta_i}$ و $\theta_i = \alpha(۱- \mid ۲\gamma - ۱ \mid)$ است.

۶.۲ تحلیل بیزی رگرسیون مستطیلی بتا

در روشی مشابه آنچه در فراری و سریباری (۲۰۰۴) آمده، برای برازش مدل در صورتی که حجم نمونه کوچک نباشد، می‌توان برآوردگرهای ماکسیمم درستنمایی پارامترهای مدل را به‌دست آورد. اگر اندازه نمونه کوچک باشد، روش ماکسیمم درستنمایی مناسب نیست. در این صورت، اگر اطلاعات کافی از پارامترها در دست باشد، همان‌طور که در (کانگدان، ۲۰۱۴) آمده، تحلیل بیزی روشی مناسب برای برازش مدل پیشنهادی خواهد بود. در این فصل به تحلیل بیزی مدل رگرسیونی بتا می‌پردازیم.

استنباط بیزی با تحلیل توزیع پسین صورت می‌گیرد. تحت مدل پیشنهادی، $\mathbf{y_i} = (y_{i۱}, y_{i۲}, \cdots, y_{in})^T$ بردار پاسخ مشاهده‌شده برای نمونه واحد i است و هر مؤلفه این بردار y_i، که بازه تعریف آن $(۰, ۱)$ است. ساختار مدل رگرسیونی به‌صورت

$$logit(\gamma_i) = \mathbf{x_i}^T \boldsymbol{\beta} = \boldsymbol{\eta_i}, \quad log(\phi_i) = -\mathbf{w_i}^T \boldsymbol{\delta} = \boldsymbol{\tau_i}, \quad i = ۱, ۲, \cdots, n$$

بنابراین

$$\begin{aligned} \gamma_i &= \frac{e^{\boldsymbol{\eta_i}}}{۱ + e^{\boldsymbol{\eta_i}}} = \frac{e^{\mathbf{x_i}^T \boldsymbol{\beta}}}{۱ + e^{\mathbf{x_i}^T \boldsymbol{\beta}}}. \\ \phi_i &= e^{\boldsymbol{\tau_i}} = e^{-\mathbf{w_i}^T \boldsymbol{\delta}}. \end{aligned} \qquad (۱۴.۲)$$

در نتیجه

$$Y_i \mid \boldsymbol{\beta}, \boldsymbol{\delta}, \alpha \sim BRr(\gamma_i, \phi_i, \alpha).$$

با توجه به تابع درستنمایی مستطیلی بتا، توزیع پسین به‌صورت زیر است:

$$p(\boldsymbol{\psi} \mid \mathbf{y}) \propto L(\boldsymbol{\psi} \mid \mathbf{y}) p(\boldsymbol{\psi}) \tag{۱۵.۲}$$

که در آن توزیع پیشین برای پارامترهای $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\delta}, \alpha)$ را با در نظر گرفتن شرط استقلال مؤلفه‌های این بردار، به‌صورت زیر تعریف می‌کنیم:

$$p(\boldsymbol{\psi}) = p(\boldsymbol{\beta}) p(\boldsymbol{\delta}) p(\alpha). \tag{۱۶.۲}$$

تحلیل بیزی نیازمند تعیین توزیع پیشین است. در حالت کلی، هر توزیع پیشینی می‌تواند استفاده شود. چنانچه اطلاعات اولیه‌ای در دست باشد، می‌توان از توزیع پیشین آگاهی‌بخش استفاده کرد. در غیر این صورت، از توزیع پیشین ناآگاهی‌بخش استفاده می‌کنیم. برای اطمینان از سره بودن توزیع پسین ناآگاهی‌بخش نیازمند معرفی توزیع‌های پیشین سره هستیم. این اطمینان منجر به اطمینان دیگری می‌شود که آن هم همگرایی الگوریتم‌های $MCMC$ به توزیع هدف (یعنی توزیع پسین) است (ربرت و کسلا، ۱۹۹۹). برای ضرایب رگرسیونی میانگین و پراکندگی به‌ترتیب توزیع‌های پیشین $\boldsymbol{\beta} \sim N_k(\mathbf{a}, \mathbf{B})$ و $\boldsymbol{\delta} \sim N_l(\mathbf{c}, \mathbf{D})$ را لحاظ کرده و برای ضریب آمیختگی α توزیع پیشین ناآگاهی‌بخش $\alpha \sim U(۰, ۱)$ را در نظر می‌گیریم. با توجه به روابط (۱۲.۲) تا(۱۶.۲) توزیع پسین به‌صورت زیر به‌دست می‌آید:

$$\begin{aligned} p(\boldsymbol{\psi} \mid \mathbf{y}) \propto & \prod_{i=۱}^{n} [\alpha(۱- \mid ۲\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱+e^{\mathbf{x_i}^T\boldsymbol{\beta}}} - ۱ \mid) + (۱ - \alpha(۱- \mid ۲\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱+e^{\mathbf{x_i}^T\boldsymbol{\beta}}} - ۱ \mid))] \\ & \times b(y \mid \frac{\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱+e^{\mathbf{x_i}^T\boldsymbol{\beta}}} - ۰/۵\alpha(۱ - \alpha(۱- \mid ۲\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱+e^{\mathbf{x_i}^T\boldsymbol{\beta}}} - ۱ \mid))}{۱ - \alpha(۱- \mid ۲\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱+e^{\mathbf{x_i}^T\boldsymbol{\beta}}} - ۱ \mid)}, e^{-\mathbf{w_i}^T\boldsymbol{\delta}}) \\ & \times \phi_k(\boldsymbol{\beta} \mid \mathbf{a,B}) \times \phi_l(\boldsymbol{\delta} \mid \mathbf{c,D}) \times ۱ \end{aligned} \tag{۱۷.۲}$$

که در آن $\phi_k(\cdot)$ و $\phi_l(\cdot)$ توزیع‌های نرمال چندمتغیره k و l بعدی هستند. توجه کنید، در صورتی که پارامتر پراکندگی مدل‌بندی نشود، یعنی ϕ را ثابت در نظر بگیریم، بردار پارامترها به‌صورت $\boldsymbol{\psi} = (\boldsymbol{\beta}, \phi, \alpha)$ در می‌آید. با توجه به کار برانسکام و همکاران (۲۰۰۶)، توزیع پیشین پارامتر دقت $\phi \sim Gamma(c, c)$ را در نظر می‌گیریم.

۷.۲ برازش مدل با استفاده از نمونه‌گیری $MCMC$

تحت مدل (۱۷.۲) تابع چگالی پسین عبارتست از

$$
\begin{aligned}
p(\boldsymbol{\psi} \mid \mathbf{y}) &= L(\boldsymbol{\psi} \mid \mathbf{y})p(\boldsymbol{\eta}) = L(\boldsymbol{\beta}, \boldsymbol{\delta}, \alpha \mid \mathbf{y})p(\boldsymbol{\beta})p(\boldsymbol{\delta})p(\alpha) \\
&\propto \prod_{i=۱}^{n} g(\mathbf{y_i} \mid (\boldsymbol{\beta}, \boldsymbol{\delta}, \alpha)p(\boldsymbol{\beta})p(\boldsymbol{\delta})p(\alpha) \\
&\propto \prod_{i=۱}^{n} [\alpha(۱- \mid ۲\gamma_i - ۱ \mid) + (۱ - \alpha(۱- \mid ۲\gamma_i - ۱ \mid)) \\
&\times \left(\frac{\Gamma(\frac{\gamma_i - ۰/۵\alpha(۱-|۲\gamma_i-۱|)}{۱-\alpha(۱-|۲\gamma_i-۱|)} + \phi_i)}{\Gamma(\frac{\gamma_i - ۰/۵\alpha(۱-|۲\gamma_i-۱|)}{۱-\alpha(۱-|۲\gamma_i-۱|)})\Gamma(\phi_i)} \right) \mathbf{y_i}^{(\frac{\gamma_i - ۰/۵\alpha(۱-|۲\gamma_i-۱|)}{۱-\alpha(۱-|۲\gamma_i-۱|)} - ۱)} (۱ - \mathbf{y_i})^{(\phi_i - ۱)}] \\
&\times (\frac{۱}{\mid ۲\pi\mathbf{B} \mid^{\frac{۱}{۲}}} exp\{-\frac{۱}{۲}(\boldsymbol{\beta} - \mathbf{a})^T \mathbf{B}^{-۱}(\boldsymbol{\beta} - \mathbf{a})\}) \\
&\times (\frac{۱}{\mid ۲\pi\mathbf{D} \mid^{\frac{۱}{۲}}} exp\{-\frac{۱}{۲}(\boldsymbol{\delta} - \mathbf{c})^T \mathbf{D}^{-۱}(\boldsymbol{\delta} - \mathbf{c})\}) \times ۱
\end{aligned}
\tag{۱۸.۲}
$$

این توزیع پسین صورت بسته و سرراستی ندارد. یکی از روش‌های معمول برای تقریب این‌گونه توزیع‌های پیچیده استفاده از الگوریتم‌های $MCMC$ است. چنانچه بخواهیم از نمونه‌گیر گیبز برای تولید نمونه از پسین (۱۸.۲) استفاده کنیم، باید توزیع‌های شرطی کامل

$$p(\boldsymbol{\beta}|\boldsymbol{\delta}, \alpha, \mathbf{y}), p(\alpha \mid \boldsymbol{\delta}, \boldsymbol{\beta}, \mathbf{y}), p(\boldsymbol{\delta}|\boldsymbol{\beta}, \alpha, \mathbf{y})$$

را محاسبه و از آن‌ها تولید نمونه کنیم. اما این کار به راحتی امکان‌پذیر نیست. زیرا توزیع‌های شرطی کامل پارامترها دارای شکل بسته و مشخص نیستند. به‌عنوان مثال توزیع شرطی کامل $\boldsymbol{\beta}|\alpha, \boldsymbol{\delta}, \mathbf{y}$ به‌صورت زیر خواهد بود:

$$p(\boldsymbol{\beta}|\alpha, \boldsymbol{\delta}, \mathbf{y}) = \frac{p(\boldsymbol{\beta}, \alpha, \boldsymbol{\delta}, \mathbf{y})}{p(\alpha, \boldsymbol{\delta}, \mathbf{y})} = \frac{p(\mathbf{y} \mid \alpha, \boldsymbol{\delta}, \boldsymbol{\beta})p(\alpha)p(\boldsymbol{\delta})p(\boldsymbol{\beta})}{p(\alpha)p(\boldsymbol{\delta})p(\mathbf{y} \mid \alpha, \boldsymbol{\delta}, \boldsymbol{\beta})} \tag{۱۹.۲}$$

برای ساده کردن کسر (۱۹.۲)، ابتدا مخرج آن را ساده می‌کنیم

$$
\begin{aligned}
p(\mathbf{y}|\alpha, \boldsymbol{\delta}, \boldsymbol{\beta}) &= \int_{\mathbb{R}^k} \prod_{i=۱}^{n} [\alpha(۱- \mid ۲\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱ + e^{\mathbf{x_i}^T\boldsymbol{\beta}}} - ۱ \mid) + (۱ - \alpha(۱- \mid ۲\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱ + e^{\mathbf{x_i}^T\boldsymbol{\beta}}} - ۱ \mid))] \\
&\times b(\mathbf{y} \mid \frac{\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱+e^{\mathbf{x_i}^T\boldsymbol{\beta}}} - ۰/۵\alpha(۱ - \alpha(۱- \mid ۲\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱+e^{\mathbf{x_i}^T\boldsymbol{\beta}}} - ۱ \mid)}{۱ - \alpha(۱- \mid ۲\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱+e^{\mathbf{x_i}^T\boldsymbol{\beta}}} - ۱ \mid}, e^{-\mathbf{w_i}^T\boldsymbol{\delta}}) d\boldsymbol{\beta} = \mathbf{I}
\end{aligned}
\tag{۲۰.۲}
$$

با استفاده از (۱۶.۲)، (۱۹.۲) و (۲۰.۲) داریم

$$
\begin{aligned}
&p(\boldsymbol{\beta} \mid \alpha, \boldsymbol{\delta}, \mathbf{y}) \\
&\propto \frac{\prod_{i=۱}^{n} [\alpha(۱- \mid ۲\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱+e^{\mathbf{x_i}^T\boldsymbol{\beta}}} - ۱ \mid) + (۱ - \alpha(۱- \mid ۲\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱+e^{\mathbf{x_i}^T\boldsymbol{\beta}}} - ۱ \mid))]}{\mathbf{I}} \\
&\times b(\mathbf{y} \mid \frac{\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱+e^{\mathbf{x_i}^T\boldsymbol{\beta}}} - ۰/۵\alpha(۱-\alpha(۱-|۲\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱+e^{\mathbf{x_i}^T\boldsymbol{\beta}}}-۱|)}{۱-\alpha(۱-|۲\frac{e^{\mathbf{x_i}^T\boldsymbol{\beta}}}{۱+e^{\mathbf{x_i}^T\boldsymbol{\beta}}}-۱|}, e^{-\mathbf{w_i}^T\boldsymbol{\delta}}) (\frac{۱}{|۲\pi\mathbf{B}|^{\frac{۱}{۲}}} exp\{-\tfrac{۱}{۲}(\boldsymbol{\beta} - \mathbf{a})^T\mathbf{B}^{-۱}(\boldsymbol{\beta} - \mathbf{a})\})
\end{aligned}
$$

بنابراین ملاحظه می‌کنید که صورت تابع چگالی شرطی کامل $\boldsymbol{\beta} \mid \alpha, \boldsymbol{\delta}, \mathbf{y}$ خیلی پیچیده است و نمی‌توان از آن به راحتی نمونه تولید کرد و باید از روش‌های تولید نمونه مانند $MCMC$ بهره برد. باید از الگوریتم‌های $MCMC$ تحت گیبز استفاده کرد. در این پایان‌نامه، تولید نمونه با این روش‌ها توسط نرم‌افزار $JAGS$، که به‌طور خودکار توزیع‌های شرطی کامل را محاسبه می‌کند، انجام شده است.

فصل ۳

تنومندی مدل رگرسیون مستطیلی بتا

در این فصل، عملکرد مدل‌های رگرسیونی بتا و مستطیلی بتا را زمانی‌که نقاط پرت وجود داشته باشند، با استفاده از دو شبیه‌سازی، مقایسه می‌کنیم. در شبیه‌سازی اول، مدل‌ها را بدون در نظر گرفتن متغیر تبیینی برازش می‌دهیم. فرآیند تولید داده‌های پرت بر اساس آلوده کردن پاسخ‌های تولیدشده از توزیع بتا با متغیر دیگر انجام شده است. به این صورت که درصدی از پاسخ‌های ترکیب‌شده بر اساس مدل رگرسیون بتا، با مقادیر تولیدشده در دم بالایی توزیع یکنواخت جایگزین شده‌اند. در مطالعه شبیه‌سازی دوم، مدل‌ها را در حضور متغیرهای تبیینی برازش می‌دهیم. سپس، مدل مناسب را بر اساس معیارهای DIC، MSE، $EAIC$ و $EBIC$ و همچنین اریبی انتخاب می‌کنیم. مطالعه شبیه‌سازی نشان می‌دهد که مدل رگرسیونی مستطیلی بتا در مقایسه با مدل رگرسیونی بتا نسبت به وجود نقاط پرت حساس نیست و از تنومندی لازم برخوردار می‌باشد.

نمونه‌هایی که با برازش مدل‌های بتا و مستطیلی بتا در این فصل تولید می‌شوند، با سوزاندن ۱۰۰۰۰ از ۲۰۰۰۰ نمونه زنجیر مارکوفی تولیدی با گام ۱۰ به‌دست آمده‌اند. به‌عبارت دقیق‌تر، گام l یعنی پس از سوزاندن زنجیر مارکوف نمونه‌های باقیمانده را به دسته‌های lتایی تقسیم نموده و از هر دسته، اولین مقدار را برمی‌گزینیم. در این‌صورت اندازه نمونه حاصل ($\frac{\text{اندازه نمونه زنجیر مارکوف}}{l}$) است. این کار به این دلیل انجام می‌شود که نمونه‌های زنجیر مارکوف به هم وابسته‌اند و براساس یک هسته سلسله‌وار به‌دست می‌آیند، با این کار تا حدودی این وابستگی را کاهش می‌دهیم و نمونه بهتری تولید می‌کنیم.

علاوه بر این، در این فصل مدل‌های بتا و مستطیلی بتا را با متغیر پراکندگی ثابت برازش می‌دهیم و توزیع پیشین آن $gamma(c,c)$ با مقادیر $c = ۰/۱, ۰/۰۱, ۰/۰۰۱$ خواهد بود (گوش و همکاران، ۲۰۰۹).

۱.۳ مطالعه شبیه‌سازی بدون متغیر تبیینی

در این مطالعه شبیه‌سازی، مقادیر ϕ را برابر ۱۰ و ۳۰، حجم نمونه را برابر $n = ۵۰, ۱۰۰, ۲۰۰$ و همچنین درصد نقاط پرت را $r = ۲, ۵, ۸$ درصد در نظر گرفتیم. مراحل انجام مطالعه به‌صورت زیر

می‌باشند:

۱. ابتدا از توزیع بتا با پارامترهای $\mu = ۰/۲$ و پارامتر دقت ϕ، نمونه‌ای با اندازه n شبیه‌سازی کردیم.

۲. نمونه‌ای بدون جایگذاری از n با اندازه $n \times \frac{r}{۱۰۰}$ انتخاب کردیم.

۳. نمونه حاصل را جایگزین مقادیر تولیدشده از توزیع یکنواخت در فاصله $(q, ۱)$ کردیم که در آن q چندک ۰/۹۹۹ توزیع بتا در مرحله اول است.

جدول ۱.۳: مقادیر اریبی و MSE برای دو مدل رگرسیونی بتا و مستطیلی بتا و درصد انتخاب مدل مستطیلی بتا بر اساس DIC در مطالعه شبیه‌سازی اول بر اساس ۱۰۰ مجموعه داده

سناریوها				بتا		مستطیلی بتا		$\%DIC$
	ϕ	٪ پرتی	n	اریبی	MSE	اریبی	MSE	
۱	۱۰	۲	۵۰	۰/۰۲۱۸	۰/۰۰۰۹	۰/۰۲۰۸	۰/۰۰۰۸	۵۰
۲	۱۰	۲	۱۰۰	۰/۰۲۰۴	۰/۰۰۰۵	۰/۰۱۷۴	۰/۰۰۰۴	۷۱
۳	۱۰	۲	۲۰۰	۰/۰۲۰۵	۰/۰۰۰۵	۰/۰۱۴۳	۰/۰۰۰۳	۹۲
۴	۱۰	۵	۵۰	۰/۰۳۵۵	۰/۰۰۱۲	۰/۰۲۹۶	۰/۰۰۱۱	۶۵
۵	۱۰	۵	۱۰۰	۰/۰۴۳۹	۰/۰۰۲۲	۰/۰۳۰۴	۰/۰۰۱۱	۸۷
۶	۱۰	۵	۲۰۰	۰/۰۴۷۷	۰/۰۰۲۴	۰/۰۳۱۰	۰/۰۰۱۱	۱۰۰
۷	۱۰	۸	۵۰	۰/۰۶۸۳	۰/۰۰۵۴	۰/۰۵۰۵	۰/۰۰۲۹	۸۰
۸	۱۰	۸	۱۰۰	۰/۰۷۲۲	۰/۰۰۵۵	۰/۰۴۹۳	۰/۰۰۲۶	۹۶
۹	۱۰	۸	۲۰۰	۰/۰۷۳۷	۰/۰۰۵۶	۰/۰۴۸۱	۰/۰۰۲۴	۱۰۰
۱۰	۳۰	۲	۵۰	۰/۰۱۷۳	۰/۰۰۰۶	۰/۰۱۶۵	۰/۰۰۰۴	۶۲
۱۱	۳۰	۲	۱۰۰	۰/۰۱۶۵	۰/۰۰۰۴	۰/۰۱۳۱	۰/۰۰۰۲	۹۰
۱۲	۳۰	۲	۲۰۰	۰/۰۱۷۳	۰/۰۰۰۴	۰/۰۱۱۶	۰/۰۰۰۲	۱۰۰
۱۳	۳۰	۵	۵۰	۰/۰۳۲۱	۰/۰۰۱۳	۰/۰۲۳۷	۰/۰۰۰۷	۸۲
۱۴	۳۰	۵	۱۰۰	۰/۰۴۱۳	۰/۰۰۱۹	۰/۰۲۶۳	۰/۰۰۰۷	۹۸
۱۵	۳۰	۵	۲۰۰	۰/۰۴۰۹	۰/۰۰۱۸	۰/۰۲۴۵	۰/۰۰۰۶	۱۰۰
۱۶	۳۰	۸	۵۰	۰/۰۵۸۵	۰/۰۰۳۸	۰/۰۳۹۷	۰/۰۰۱۷	۹۲
۱۷	۳۰	۸	۱۰۰	۰/۰۶۱۵	۰/۰۰۳۹	۰/۰۳۹۱	۰/۰۰۱۶	۱۰۰
۱۸	۳۰	۸	۲۰۰	۰/۰۶۳۵	۰/۰۰۴۲	۰/۰۳۷۴	۰/۰۰۱۴	۱۰۰

نتایج این شبیه‌سازی در جدول ۱.۳ آمده است. همان‌طور که در جدول ملاحظه می‌کنید، با توجه به مقادیری که برای ϕ ، n و r فرض کردیم، ۱۸ سناریو از داده آلوده‌شده بتا داریم. در هر سناریو، ۱۰۰ مجموعه از داده‌های بتا تولید شدند. برای مجموعه داده تولیدشده، مدل‌های رگرسیونی بتا و مستطیلی بتا را برازش دادیم (بخش‌های ۲.۲ و ۵.۲ را ببینید). نمونه‌های تولیدشده از الگوریتم $MCMC$ در $JAGS$

به دو قسمت تقسیم می‌شوند. قسمت اول، شامل یک دوره سوزاندن و قسمت دوم، خروجی است که برای استنباط در مورد مدل‌ها استفاده می‌شود. در واقع نمونه‌های قسمت دوم برای تقریب توزیع پسین به کار می‌روند. به منظور کنترل همگرایی، به‌طور معمول به تولید چندین زنجیر موازی از نمونه‌ها نیاز داریم. در این قسمت، ۳ زنجیر را برای کنترل همگرایی در نظر گرفتیم.

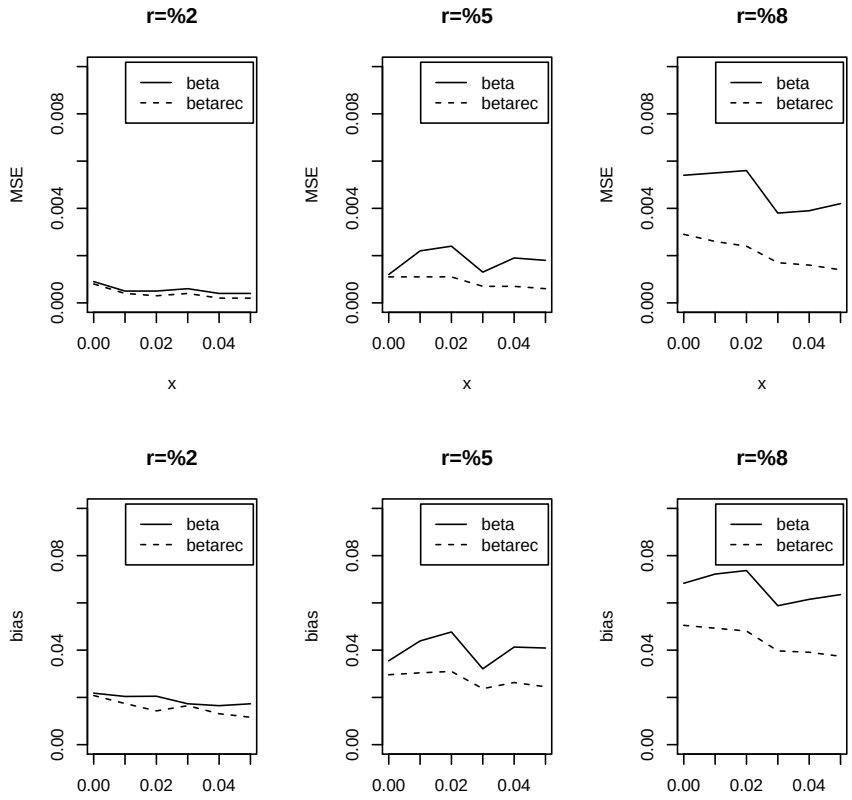
در جدول ۱.۳، MSE و اریبی برآورد پارامتر برای هر سناریو به‌دست آمده‌اند. بر اساس معیار DIC، مدل مستطیلی بتا نسبت به مدل بتا در تمامی سناریوها مدل برتر است، به این معنی که برای ۱۰۰ مجموعه داده شبیه‌سازی شده در هر سناریو، بر حسب DIC کمتر، تعداد دفعاتی که مدل مستطیلی بتا برازش بهتری دارد بیشتر است، به جز سناریوی اول که شانس انتخاب مدل‌های بتا و مستطیلی بتا برابرند. در سناریوهایی که نقاط پرت مشابهی دارند (برای مثال، سناریوهای ۱ تا ۳)، با افزایش حجم نمونه با احتمال بیشتری مدل مستطیلی بتا برگزیده می‌شود. اگر دسته‌ای از سناریوها را در نظر بگیرید که درصد نقاط پرت و اندازه نمونه مشابهی دارند، برای مقادیر پراکندگی کمتر (یعنی ϕ بزرگتر) با احتمال بیشتری مدل مستطیلی بتا برتر است.

همان‌طور که در جدول ملاحظه می‌کنید، با افزایش حجم نمونه، برای درصد نقاط پرت یکسان، MSE کاهش می‌یابد. در صورتی که حجم نمونه را ثابت فرض کنیم، با افزایش درصد نقاط پرت، MSE نیز در هر دو مدل افزایش می‌یابد. اما با حجم نمونه ثابت و همچنین در یک سطح ثابت از درصد نقاط پرت، مدل مستطیلی بتا MSE کمتری از مدل بتا دارد. به‌عبارتی، با افزایش نقاط پرت در مدل‌ها، خطا در آن‌ها افزایش می‌یابد.

با افزایش حجم نمونه و درصد نقاط پرت، مدل مستطیلی بتا اریبی کمتری نسبت به بتا دارد. این در حالی است که با حجم نمونه ثابت و افزایش درصد نقاط پرت، اریبی در هر دو مدل افزایش می‌یابد. همچنین، با حجم نمونه و نقاط پرت مشابه، به ازای مقدار بالاتر ϕ، در هر دو مدل مورد نظر، اریبی کاهش می‌یابد (به‌عبارتی، کاهش پراکندگی، اریبی را کاهش می‌دهد).

نتایج مطالعه شبیه‌سازی مذکور نشان می‌دهد که در هر سناریو، برای داده‌های آلوده‌شده، برآورد پارامترهای مدل مستطیلی بتا نسبت به مدل بتا از نظر دقت (MSE و اریبی کمتر) بهبود یافته است. این جدول مؤید آن است که، با معیار DIC در اکثر موارد مدل مستطیلی بتا انتخاب بهتری نسبت به مدل بتا است. بدیهی است که با افزایش حجم نمونه و درصد نقاط پرت این نتیجه آشکارتر می‌گردد.

شکل ۱.۳ منحنی مقادیر MSE و اریبی پارامتر برآوردشده را برای اندازه‌های نمونه مختلف به ازای درصد نقاط پرت متفاوت، نمایش می‌دهد. همان‌طور که مشهود است با نقاط پرت کمتر، اریبی و MSE نیز مقادیر کمتری دارند. هر چه تعداد نقاط پرت افزایش یابد، اریبی و MSE برای هر دو مدل مقادیر بزرگتری اختیار می‌کنند. در تمامی حالات، نمودارهای مربوط به مدل مستطیلی بتا زیر نمودار بتا قرار دارد. هرچه نقاط پرت بیشتر می‌شوند، فاصله نمودار بتا از مستطیلی بتا افزایش می‌یابد. این نتایج تنومندی مدل رگرسیون مستطیلی بتا را نسبت مدل بتا در حضور داده‌های پرت نشان می‌دهد و هر چه درصد نقاط پرت بیشتر باشد، مدل رگرسیون مستطیلی بتا برای به کار گرفتن بیشتر توصیه می‌شود.



شکل ۱.۳: اریبی و $\bar{MSE}$ برای تعداد نمونه‌های مختلف نسبت به نقاط پرت متفاوت تحت مدل‌های مستطیلی بتا و بتا

۲.۳ مطالعه شبیه‌سازی در حضور متغیرهای تبیینی

در این بخش، مطالعه شبیه‌سازی شامل دو قسمت می‌شود: تحلیل حساسیت مدل‌ها برای داده‌های آلوده‌نشده و آلوده‌شده تولیدشده از مدل بتا.

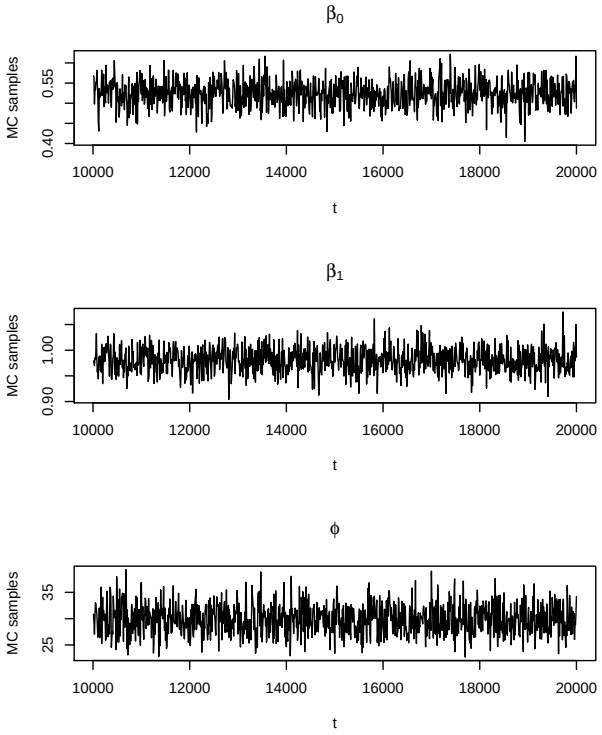
۱.۲.۳ تحلیل حساسیت مدل‌ها برای داده‌های آلوده‌نشده

در این قسمت، حساسیت مدل‌های بتا و مستطیلی بتا را با وجود نقاط پرت، بر اساس معیارهای متفاوت، مقایسه می‌کنیم. به منظور شبیه‌سازی، مقادیر اولیه مشخصی را برای پارامترها، در نظر گرفتیم. تعداد نمونه را برابر ۲۰۰ قرار داده و از تابع پیوند لجیت برای میانگین استفاده کردیم و متغیر تبیینی $\mathbf{x} = (x_1, \cdots, x_n)^T$ را از توزیع یکنواخت $(-3, 3)$ تولید کردیم.

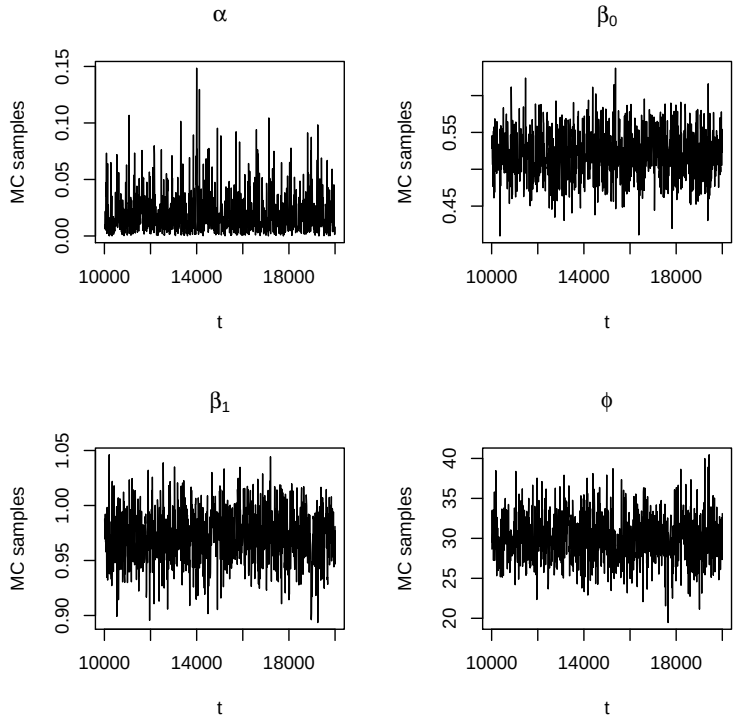
مقادیر واقعی پارامترهای مدل رگرسیونی بخش ۲.۲ را برابر $\beta_0 = 0/5$، $\beta_1 = 1$ و $\phi = 30$ در نظر می‌گیریم. با این روش مجموعه داده‌ای که از مدل بتا به‌دست می‌آید، آلوده‌نشده است.

شکل‌های ۲.۳ و ۳.۳ نمودارهای اثر[1] پارامترهای مدل بتا و مستطیلی بتا را نمایش می‌دهند که فقط از زنجیر اول به‌دست آمده‌اند. همان‌طور که می‌دانید نمودارهای اثر، اثر پارامترها را در نمونه‌های

[1] Trace plots



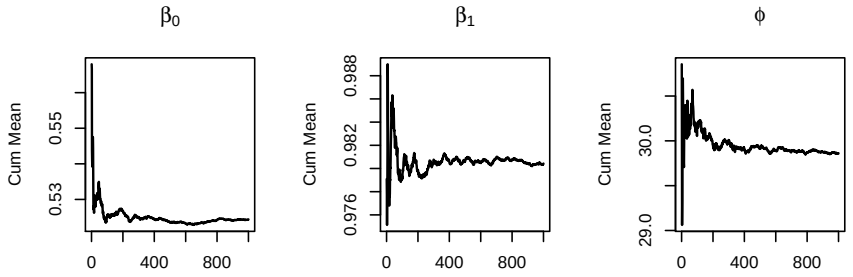
شکل ۲.۳: نمودارهای اثر پارامترهای مدل بتا به ازای داده‌های آلوده‌نشده



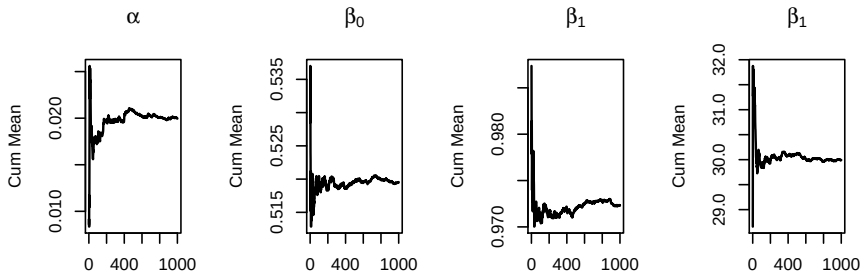
شکل ۳.۳: نمودارهای اثر پارامترهای تحلیل حساسیت مدل مستطیلی بتا به ازای داده‌های آلوده‌نشده

تولیدی بررسی می‌کنند و در صورتی که زنجیر حاصل منظم باشد و پرش چندانی بین نمونه‌های نزدیک هم به وجود نیاید، می‌توان گفت که نمودار نسبت به پارامتر مورد نظر از آمیختگی خوبی برخوردار است.

با توجه به این نمودارها، همه پارامترهای مدل بتا و مستطیلی بتا از آمیختگی خوبی برخوردار هستند.



(الف)



(ب)

شکل ۴.۳: نمودارهای میانگین تجمعی پارامترهای مدل‌های بتا (الف) و مستطیلی بتا (ب) به ازای داده‌های آلوده‌نشده

نمودار ۴.۳ میانگین تجمعی پارامترهای مدل بتا و مستطیلی بتا را نمایش می‌دهد. این نمودار یک آزمون بصری برای بررسی در دو مدل بتا و مستطیلی بتا است. این نمودار میانگین پسین تجمعی را در مقابل تعداد تکرارها رسم می‌کند. بنابراین با توجه به شکل‌های ۲.۳ تا ۴.۳ می‌توان نتیجه گرفت که زنجیر دارای ویژگی‌های خوب آمیختگی است.

برای بررسی همگرایی زنجیرها می‌توان از آزمون‌های رسمی برای بررسی همگرایی زنجیرهای $MCMC$ مانند آزمون همگرایی گلمن-روبین (۱۹۹۲) استفاده کرد. این آزمون برای دو مدل بتا و مستطیلی بتا انجام شده و نتایج آن‌ها در جدول ۲.۳ گزارش شده‌اند. مقادیر آماره $PSRF$ نزدیک به یک و برای اکثر پارامترها مقدار یک به‌دست آمده است. بنابراین با توجه به این نتایج می‌توان گفت زنجیر برای هر پارامتر همگرا است. همچنین در این برازش برای داده‌های آلوده‌نشده، $MPSRF = ۱ < ۱/۲$ می‌باشد که بیان‌کننده همگرایی زنجیر به توزیع مانا خواهد بود.

در جدول ۳.۳، میانگین پسین و همچنین فاصله اعتبار ۹۵٪ را برای پارامترهای مدل‌های مستطیلی بتا و بتا گزارش شده‌اند. با توجه به مقادیر اولیه‌ای که برای پارامترها در نظر گرفته شده است و آن‌چه در جدول ملاحظه می‌کنید، با تفاوت اندکی مدل بتا برازش بهتری نسبت به مدل مستطیلی بتا دارد. اما می‌توان گفت که، تقریباً هر دو مدل برازش مشابهی دارند. همچنین، طول فواصل اعتبار، برای هر دو مدل و به ازای هر پارامتر تقریباً برابر هستند. با تفاوت اندکی می‌توان گفت که، نمونه‌های تولید شده از

جدول ۲.۳: نتایج آزمون گلمن-روبین بر داده‌های آلوده‌نشده

مدل	پارامتر	برآورد آماره $PSRF$
بتا	β_0	۱/۰۱
	β_1	۱/۰۰
	ϕ	۱/۰۰
مستطیلی بتا	β_0	۱/۰۰
	β_1	۱/۰۱
	ϕ	۱/۰۰
	α	۱/۰۰

جدول ۳.۳: میانگین پسین و فاصله اعتبار ۹۵٪ برای پارامترهای در مدل بتا و مستطیلی بتا

	بتا			مستطیلی بتا		
پارامترها	میانگین	٪۲/۵	٪۹۷/۵	میانگین	٪۲/۵	٪۹۷/۵
β_0	۰/۵۲۲	۰/۴۶۲	۰/۵۸۹	۰/۵۲۱	۰/۴۵۶	۰/۵۸۵
β_1	۰/۹۸۱	۰/۹۳۵	۱/۰۲۸	۰/۹۷۳	۰/۹۲۵	۱/۰۲۲
ϕ	۲۹/۹۹۳	۲۴/۲۷۲	۳۶/۱۷۴	۳۰/۱۸۱	۲۴/۴۳۰	۳۶/۷۲۳
α				۰/۰۱۹	۰/۰۰۱	۰/۰۶۵

مدل بتا و مستطیلی بتا برای داده‌های آلوده‌نشده تقریباً از کیفیت یکسانی برخوردار هستند.

مقایسه دو مدل با معیارهای انتخاب مدل

جدول ۴.۳: معیارهای مقایسه مدل‌های بتا و مستطیلی بتا

مدل	DIC	$EAIC$	$EBIC$
مستطیلی بتا	−۵۳۷/۱۹۵	−۵۴۲/۲۵۷	−۵۲۹/۰۶۳
بتا	−۵۴۷/۵۲۳	−۵۴۴/۵۴۵	−۵۳۴/۶۵۰

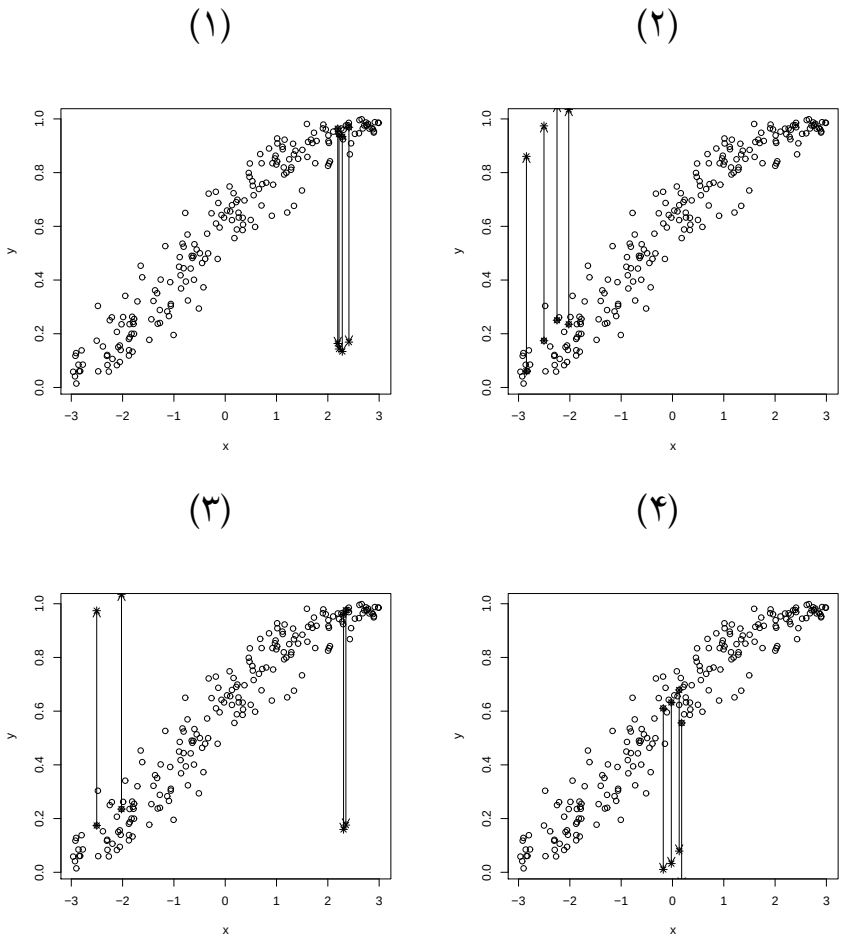
چون برازش مدل‌های بتا و مستطیلی بتا برای پارامترهای رگرسیون طبق داده‌های آلوده‌نشده انجام شده است، معیارهای انتخاب مدل‌ها $(DIC, EAIC, EBIC)$ در جدول ۴.۳ گزارش شدند. همان‌طور که ملاحظه می‌کنید، مقادیر به‌دست آمده از معیارهای DIC، $EAIC$ و $EBIC$ برای مدل بتا کمتر از مستطیلی بتا است. در نتیجه برای داده‌های آلوده‌نشده، مدل بتا عملکرد بهتری نسبت به مستطیلی بتا دارد.

در بخش بعدی، با در نظر گرفتن الگوی نقاط پرت راهکارهایی را برای آلوده کردن داده‌های بتا

معرفی می‌کنیم. سپس، براساس داده‌های آلوده‌شده، مدل‌های مورد نظر را با هم مقایسه می‌کنیم.

۲.۲.۳ تحلیل حساسیت مدل‌ها برای داده‌های آلوده‌شده

داده‌هایی که از دنیای واقعی به‌دست می‌آیند، معمولاً با خطا همراه بوده و در هر مجموعه، داده پرت وجود دارد. برای مقایسه عملکرد مدل رگرسیونی مستطیلی بتا نسبت به بتا در حضور داده‌های پرت و ارزیابی حساسیت آنها، داده‌های شبیه‌سازی‌شده بخش ۱.۲.۳ را با داده‌های پرت آلوده کردیم. روش آلوده کردن را در ادامه توضیح خواهیم داد.



شکل ۵.۳: تولید داده‌های آلوده شده تحت چهار الگوی اختلال: نمودار (۱) کاهش Δ مقادیر پاسخ برای xهای بالا و نمودار (۲) افزایش مقادیر Δ پاسخ برای x های کوچک، نمودار (۳)، کاهش و افزایش واحدهای Δ پاسخ برای xهای بالا و پایین و همچنین نمودار (۴) کاهش واحدهای Δ مقادیر پاسخ برای xهای مرکزی

اگر میزان آلودگی مشاهدات داده‌های شبیه‌سازی‌شده را ۲٪ فرض کنید، در واقع، ۴ داده از ۲۰۰ نمونه تولید شده را آلوده خواهیم کرد. آلودگی داده‌ها، به این معنی است که مشاهدات آلوده‌شده $y_i^*(\Delta) = y_i^* \pm \Delta$ را با مشاهدات شبیه‌سازی‌شده y_i^* جایگزاری کنیم. برای آلوده کردن داده‌های تولید شده از مدل رگرسیون بتا چهار الگوی اختلال متفاوت را در نظر گرفتیم:

۱. مقادیر پاسخ متناظر با مقادیر بزرگ x را Δ واحد کاهش می‌دهیم.

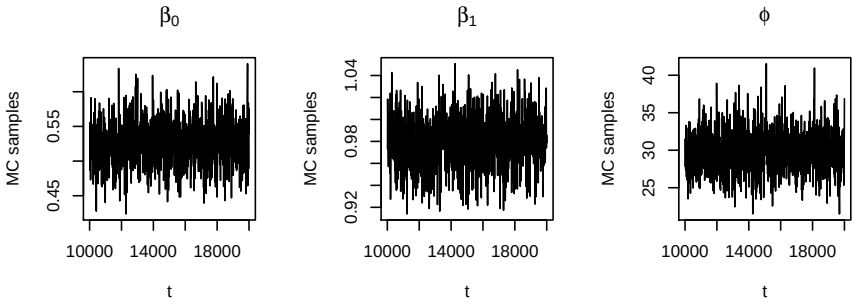
۲. مقادیر پاسخ متناظر با مقادیر کوچک x را Δ واحد افزایش می‌دهیم.

۳. مقادیر پاسخ متناظر با مقادیر بزرگ و کوچک x را به‌ترتیب Δ واحد کاهش و افزایش می‌دهیم.

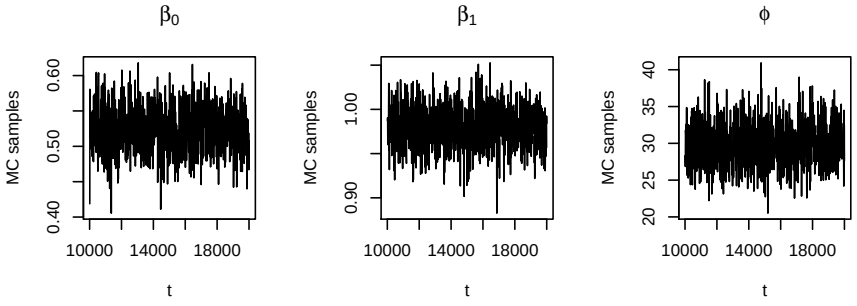
۴. مقادیر پاسخ متناظر با مقادیر میانی x را Δ واحد کاهش می‌دهیم.

الگوهای اختلال ۱، ۲، ۳ و ۴ در شکل ۵.۳ نمایش داده شده‌اند. این الگوها، روش‌های متفاوتی برای آلوده کردن داده‌ها هستند. مقادیر مختلف Δ برای الگوهای ۱، ۲ و ۳ در بازه تغییرات ۰ تا ۰/۸ با فواصل مکانی ۰/۰۵ تغییر می‌کنند (۱۷ مورد). برای الگوی ۴، تغییرات Δ در بازه ۰ تا ۰/۵ با فواصل مکانی ۰/۰۵ می‌باشد (۱۱مورد).

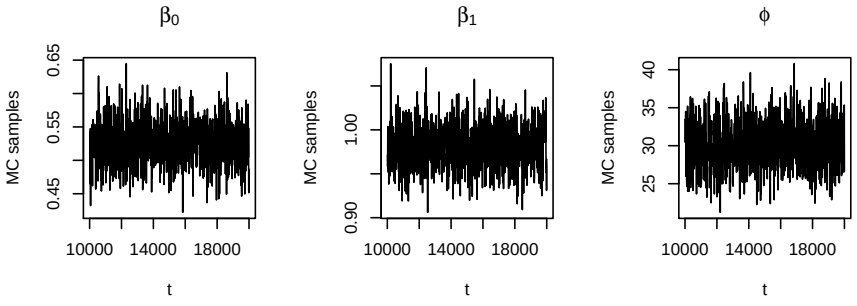
نمودارهای ۶.۳ و ۷.۳ نمودارهای اثر پارامترهای برآوردشده مدل‌های بتا و مستطیلی بتا برای داده‌های آلوده‌شده، تحت چهار الگوی اختلال هستند. نمودارها به ازای بیشترین میزان اختلال برای هر الگو رسم شده‌اند. با رسم این نمودارها می‌توان رفتار آمیختگی زنجیر اول (نمودارها مربوط به زنجیر اول هستند) را در فضای پارامتر مشاهده کرد. با توجه به این نمودارها و نوسانات منظم زنجیرها، در هر یک از آن‌ها می‌توان نتیجه گرفت زنجیر برای هر پارامتر از آمیختگی خوبی برخوردار است. اما در نمودار ۶.۳ که مربوط به پارامترهای مدل رگرسیون بتا است، نوسانات زنجیر به ازای اکثر پارامترها در الگوهای مختلف، حول مقدار واقعی‌شان نیستند و این خود گویای نامناسب بودن برازش مدل بتا در حضور داده‌های پرت است.



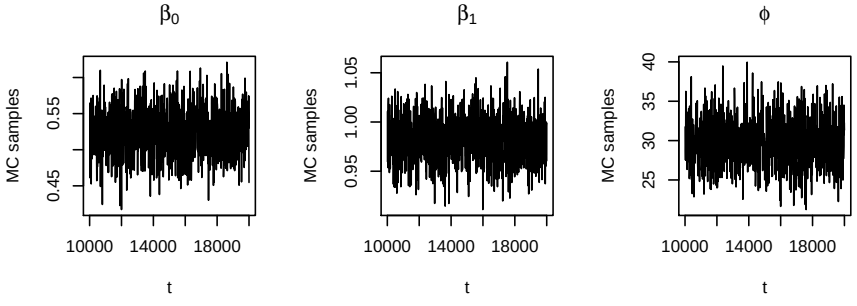
(۱)



(۲)

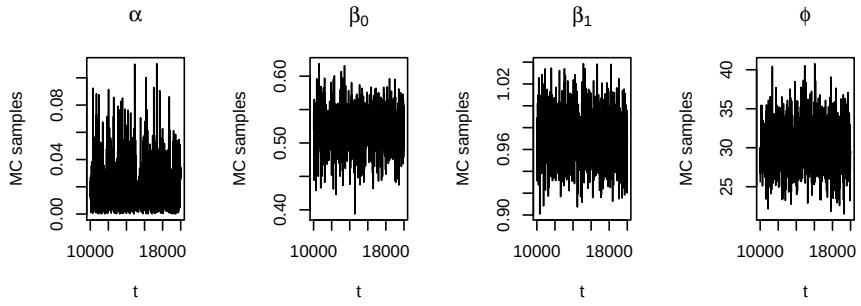


(۳)

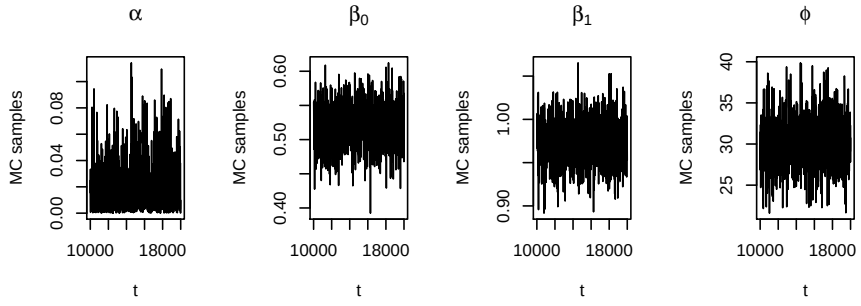


(۴)

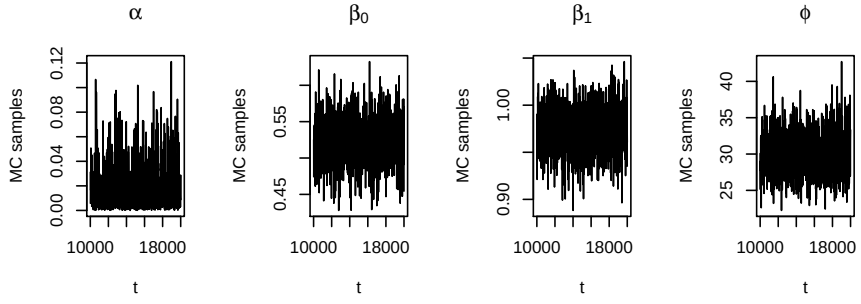
شکل ۶.۳: نمودارهای اثر پارامترهای مدل بتا به ازای داده‌های آلوده‌شده و الگوهای اختلال



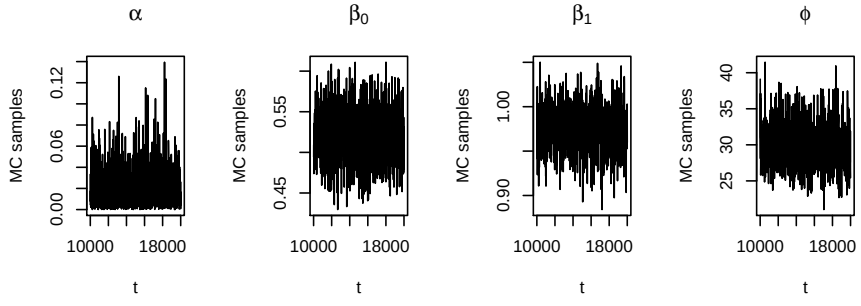
(۱)



(۲)

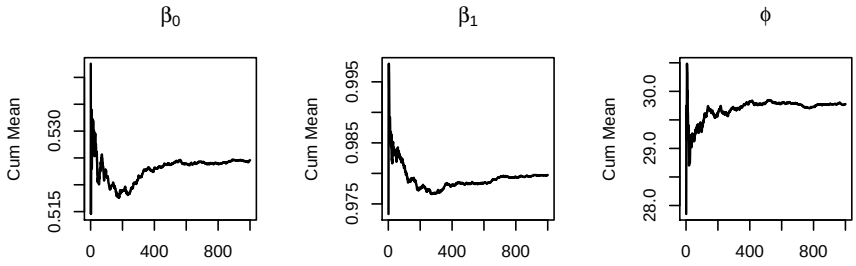


(۳)

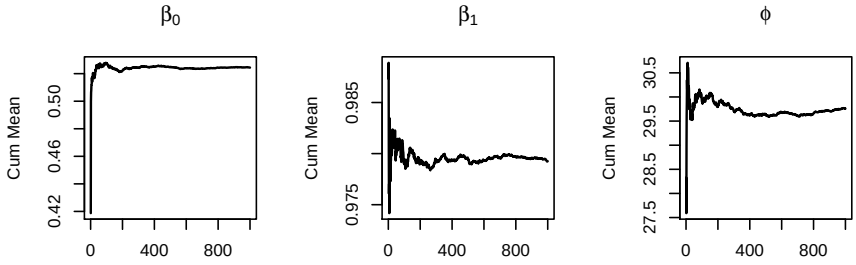


(۴)

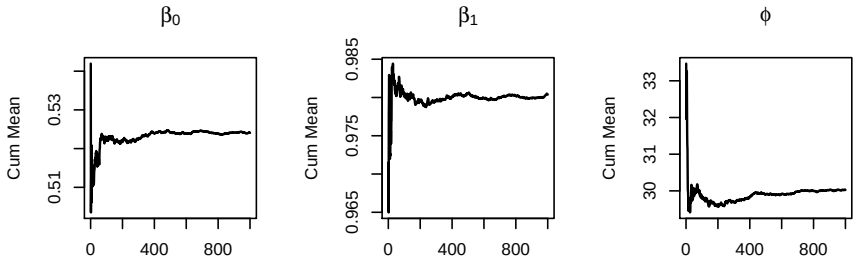
شکل ۷.۳: نمودارهای اثر پارامترهای مدل مستطیلی بتا به ازای داده‌های آلوده‌شده و الگوهای اختلال



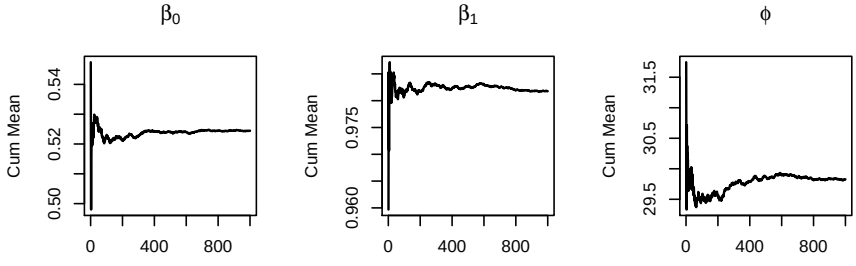
(۱)



(۲)

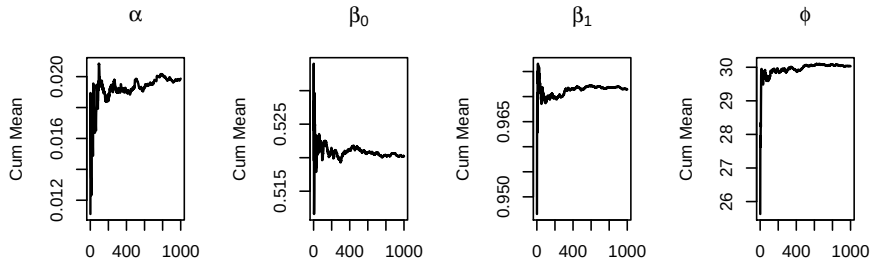


(۳)

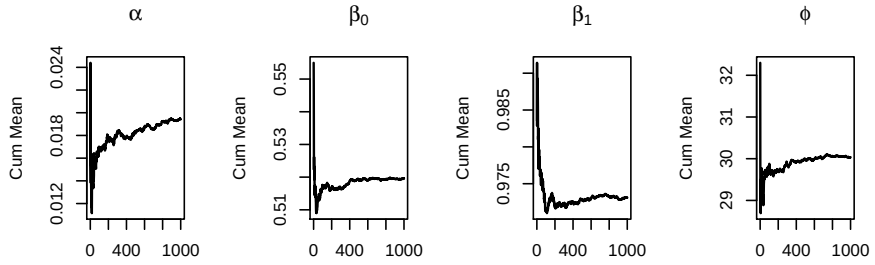


(۴)

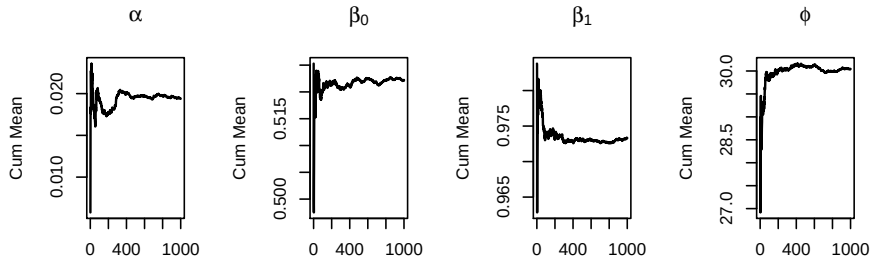
شکل ۸.۳: نمودارهای میانگین تجمعی پارامترهای مدل بتا به ازای داده‌های آلوده‌شده و الگوهای اختلال



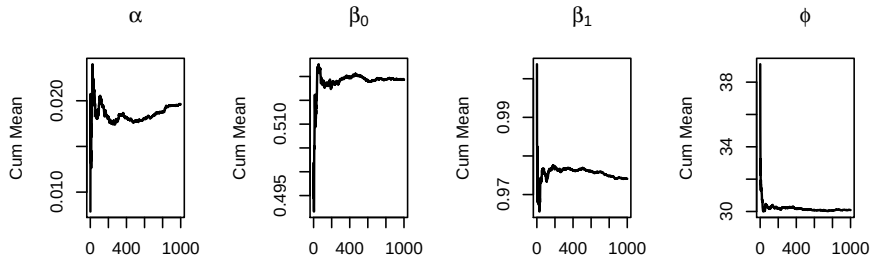
(۱)



(۲)



(۳)



(۴)

شکل ۹.۳: نمودارهای میانگین تجمعی پارامترهای مدل مستطیلی بتا به ازای داده‌های آلوده‌شده و الگوهای اختلال

در نمودارهای ۸.۳ و ۹.۳ میانگین پسین تجمعی در مقابل تکرارها، برای مدل‌های بتا و مستطیلی بتا رسم شده‌اند. این نمودارها نیز، بر اساس بیشترین میزان اختلال هر الگو است. با بررسی این نمودارها، می‌توان همگرایی هر پارامتر به مقدار واقعی آن را مشاهده کرد. همان‌طور که ملاحظه می‌کنید، به ازای بیشترین میزان اختلال، نمودارهای مربوط به مدل مستطیلی بتا به مقادیر واقعی همگرا هستند. در حالی که در نمودار ۸.۳ پارامتر β_0 به ازای الگوهای اختلال اول و چهارم همگرا نیست. پارامتر β_1 برای الگوهای اول و دوم همگرا نمی‌باشد و پارامتر ϕ به ازای تمامی الگوها همگرا نیست. در نتیجه با افزایش اختلال، پارامترهای مدل رگرسیون مستطیلی بتا همگرا می‌باشد، در حالی که پارامترهای مدل بتا در بعضی الگوها یا تمامی آن‌ها همگرا نمی‌باشند و بر این اساس، مدل مستطیلی بتا با افزایش نقاط پرت برازش بهتری دارد.

جدول ۵.۳: نتایج آزمون همگرایی گلمن-روبین برای داده‌های آلوده‌شده مدل بتا تحت چهار الگوی اختلال

الگو اختلال	پارامتر	برآورد آماره $PSRF$
(۱)	β_0	۱/۰۰
	β_1	۰/۹۹
	ϕ	۱/۰۰
(۲)	β_0	۱/۰۰
	β_1	۱/۰۰
	ϕ	۱/۰۰
(۳)	β_0	۱/۰۱
	β_1	۰/۹۹
	ϕ	۱/۰۰
(۴)	β_0	۱/۰۰
	β_1	۱/۰۰
	ϕ	۱/۰۱

همان‌طور که در جداول ۵.۳ و ۶.۳ ملاحظه می‌کنید، مقدار آماره $PSRF$ برای همه پارامترهای مدل بتا تحت چهار الگوی اختلال نزدیک به یک به‌دست آمده و در الگوی دوم برای همه پارامترها یک است و در مدل مستطیلی بتا در الگوی دوم و سوم برای همه پارامترها یک به‌دست می آید در نتیجه زنجیر برای همه پارامترها همگرا است. در برازش مدل‌های بتا و مستطیلی بتا برای داده‌های آلوده‌شده $MPSRF = ۱ < ۱/۲$ است که بیانگر همگرایی زنجیر به توزیع مانا است.

جدول ۶.۳: نتایج آزمون همگرایی گلمن-روبین برای داده‌های آلوده‌شده مدل مستطیلی بتا تحت چهار الگوی اختلال

الگو اختلال	پارامتر	برآورد آماره $PSRF$
(۱)	β_0	۱/۰۰
	β_1	۱/۰۰
	ϕ	۱/۰۰
	α	۱/۰۱
(۲)	β_0	۱/۰۰
	β_1	۱/۰۰
	ϕ	۱/۰۰
	α	۱/۰۰
(۳)	β_0	۱/۰۰
	β_1	۱/۰۰
	ϕ	۱/۰۰
	α	۱/۰۰
(۴)	β_0	۱/۰۰
	β_1	۱/۰۰
	ϕ	۱/۰۱
	α	۱/۰۰

پس از بررسی اولیه برازش در مدل برای الگوهای اختلال معرفی‌شده، به بررسی حساسیت مدل‌ها بر اساس سه مؤلفه مختلف می‌پردازیم:

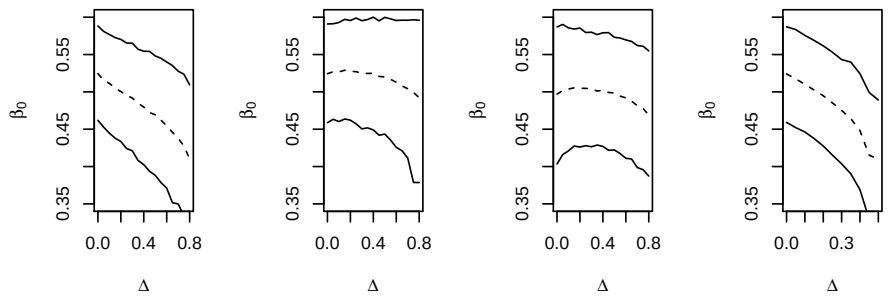
۱. بررسی حساسیت برآورد ضرایب رگرسیون

۲. بررسی حساسیت توزیع پسین همه پارامترها با در نظر گرفتن فاصله کولبک-لیبلر

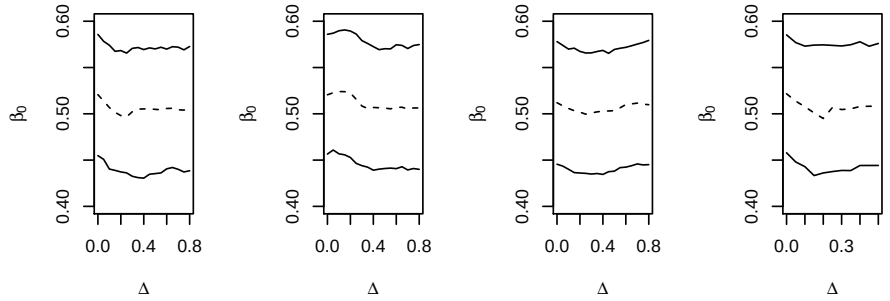
۳. بررسی حساسیت نسبت به معیارهای انتخاب مدل

بررسی حساسیت برآورد ضرایب رگرسیون

در این قسمت، برای هر الگوی اختلال و هر مقدار Δ، مدل‌های بتا و مستطیلی بتا را دوباره برآورد می‌کنیم. اندازه نمونه‌ای که استنباط بر اساس آن انجام می‌شود، ۱۰۰۰ است. میانگین پسین و فاصله اعتبار ۹۵٪ برای پارامترهای رگرسیون β_0 و β_1 در هر الگوی اختلال، برای مدل‌های بتا و مستطیلی بتا در شکل‌های ۱۰.۳ و ۱۱.۳ نشان داده شده‌اند. میانگین پسین توسط خطچین و فاصله اعتبار با خط مشخص شده‌اند.



(الف)



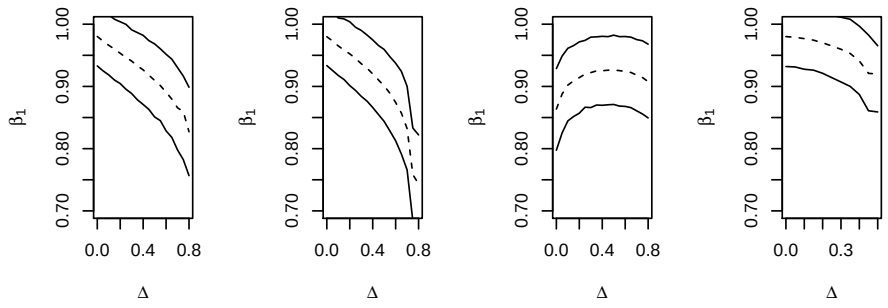
(ب)

شکل ۱۰.۳: میانگین و فاصله اعتبار ۹۵٪ برای ضریب β_0 طبق مدل‌های بتا (الف) و مستطیلی بتا (ب) برای مقادیر مختلف Δ تحت چهار الگوی اختلال نقاط پرت برای داده‌های آلوده

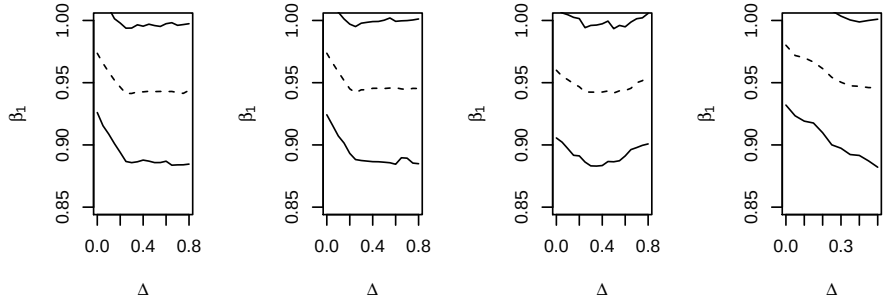
همان‌طور که در شکل‌های ۱۰.۳ و ۱۱.۳ ملاحظه می‌کنید، مدل رگرسیون مستطیلی بتا نسبت به تغییرات Δ، در هر الگوی اختلال از حساسیت قابل ملاحظه کمتری نسبت به مدل رگرسیون بتا برخوردار است. می‌توان گفت، مدل مستطیلی بتا نسبت به هر Δ و هر الگوی اختلال تنومند است. تغییرات در Δ تأثیر قابل توجهی روی میانگین و فاصله اعتبار مدل بتا داشته‌اند. میانگین و فاصله اعتبار برای همه الگوهای اختلال رسم شده‌اند. لازم به ذکر است، طول فواصل اعتبار زمانی که $|\Delta|$ افزایش می‌یابد، به ویژه برای β_0، افزایش می‌یابد.

بررسی حساسیت روی توزیع پسین همه پارامترها

در این بخش تأثیر حضور نقاط پرت بر روی دو مدل رگرسیونی بتا و مستطیلی بتا را بر اساس سنجش فاصله بین دو توزیع پسین با داده‌های آلوده‌نشده، $p(\boldsymbol{\psi}|\mathbf{Y})$، و داده‌های آلوده‌شده، $p(\boldsymbol{\psi}|\mathbf{Y}(\Delta))$، بررسی می‌کنیم. این بخش بر اساس چهار اختلال معرفی‌شده قبلی صورت می‌پذیرد. این بررسی از آن‌جا که کل توزیع پسین را شامل می‌شود، می‌تواند یک سنجش حساسیت جامع تصور شود. معیار فاصله همان اندازه واگرایی کولبک-لیبلر است که در بخش ۵.۸.۱ معرفی شد. این معیار برای دو توزیع پسین مورد



(الف)



(ب)

شکل ۱۱.۳: میانگین و فاصله اعتبار ۹۵٪ برای ضریب β_1 طبق مدل‌های بتا (الف) و مستطیلی بتا (ب) برای مقادیر مختلف Δ تحت چهار الگوی اختلال نقاط پرت برای داده‌های آلوده

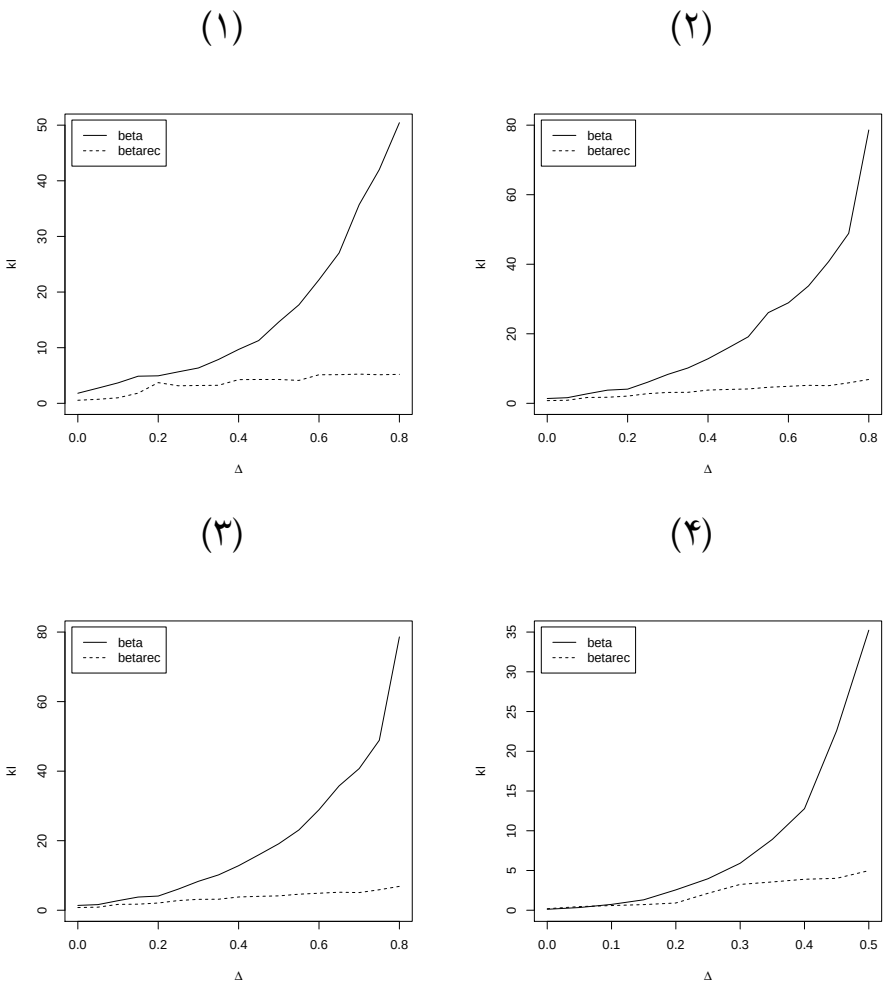
نظر به‌صورت زیر است:

$$KL(\Delta) = \int log(\frac{p(\boldsymbol{\psi}|\mathbf{Y})}{p(\boldsymbol{\psi}|\mathbf{Y}(\Delta))})p(\boldsymbol{\psi}|\mathbf{Y})d\boldsymbol{\psi}$$

برای برآورد این فاصله از روش مونت کارلو و بر اساس نمونه‌های $MCMC$ تولیدشده از استنباط‌های قبلی استفاده می‌کنیم. این برآورد به صورت زیر است:

$$\widehat{KL}(\Delta) = log\frac{\{\Pi_{b=1}^{B}\frac{1}{\vartheta(\boldsymbol{\psi}^b)}\}^{\frac{1}{B}}}{\{B^{-1}\Sigma_{b=1}^{B}\vartheta(\boldsymbol{\psi}^b)\}^{-1}}$$

که در آن $\vartheta(\boldsymbol{\psi}^b) = \frac{p(y_i+\Delta|\boldsymbol{\psi}^s)}{p(y_i|\boldsymbol{\psi}^s)}$، $\{\boldsymbol{\psi}^b\}_{b=1}^{B}$ نمونه‌های $MCMC$ تولیدشده از توزیع پسین $p(\boldsymbol{\psi}|\mathbf{Y})$، و B حجم نمونه مونت کارلو در الگوریتم $MCMC$ هستند. صورت این برآوردگر توسط پنگ و دی (۱۹۹۵) معرفی شد. اندازه $KL(\Delta)$ می‌تواند معیار مناسبی برای سنجش تأثیر نقاط پرت باشد. این اندازه‌ها را برای هر دو مدل بتا و مستطیلی بتا، برآورد می‌کنیم. مدل نیرومند، مدلی است که مقادیر کوچکتری از $KL(\Delta)$ را نتیجه دهد. با این توصیف انتظار داریم مدل رگرسیونی مستطیلی بتا برآوردهای



شکل ۱۲.۳: فاصله کولبک-لیبلر بین توزیع‌های پسین $p(\boldsymbol{\psi}|\mathbf{Y})$ و $p(\boldsymbol{\psi}|\mathbf{Y}(\Delta))$ تحت مدل‌های رگرسیونی بتا و مستطیلی بتا طبق مقادیر مختلف Δ و الگوهای مختلف اختلال برای داده‌های آلوده‌شده

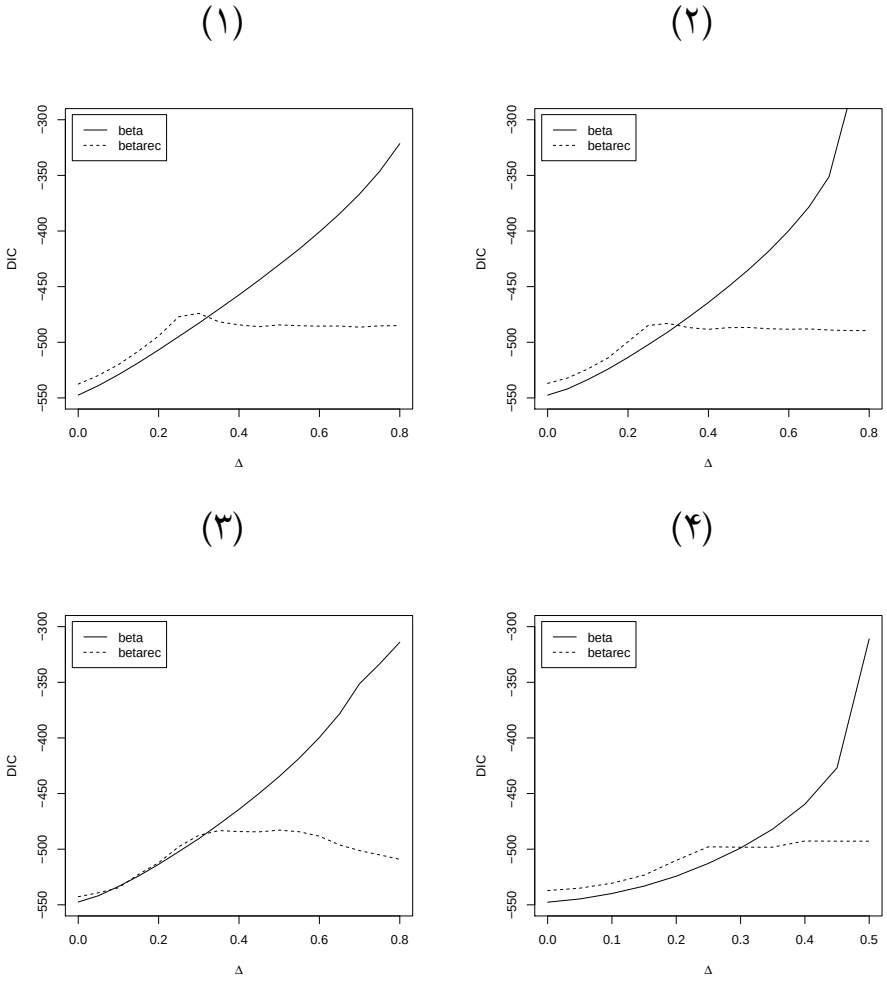
کوچکتری را ارائه دهد.

شکل ۱۲.۳، $\widehat{KL}(\Delta)$ را برای مدل بتا با خط و برای مستطیلی بتا با خط‌چین نشان می‌دهد. برای مدل مستطیلی بتا دو مقدار توزیع پسین تحت داده‌های آلوده‌شده و آلوده‌نشده را به‌دست آورده و فاصله آن‌ها را با معیار کولبک-لیبلر می‌سنجیم، همین کار را برای مدل بتا هم انجام داده و سرانجام به ازای هر اختلال اندازه این معیار برای دو مدل به‌دست می‌آوریم. مقادیر حاصل را به هم وصل نموده و نمودار ۱۲.۳حاصل می‌شود. تحت چهار الگوی اختلال و برای مقادیر متفاوت Δ نمودار مستطیلی بتا پایین نمودار بتا قرار می‌گیرد. به عبارت دیگر، به ازای مقادیر مختلف Δ، مدل مستطیلی بتا برآوردهای کوچکتری از $\widehat{KL}(\Delta)$ را نسبت به مدل بتا اختیار می‌کند. برآورد $KL(\Delta)$ برای مدل بتا تابعی صعودی از Δ است. در حالی که در مدل رگرسیون مستطیلی بتا تابعی تقریباً یکنواخت می‌باشد. به عبارت دیگر، مشاهدات دورافتاده در مدل بتا برآوردهای کران‌داری از $KL(\Delta)$ را نتیجه نمی‌دهند. اما واضح است که برای مدل مستطیلی بتا این برآوردها کران‌دار می‌باشند. همان‌طور که در شکل مشهود است در صورتی که $\Delta < ۰/۲$ باشد، دو مدل عملکرد مشابهی دارند و دو نمودار تفاوت چندانی نداشته و تقریباً

بر هم مماسند. چهار الگوی اختلال، برای هر دو مدل، روند مشابهی دارند. این روند حاکی از آن است که هر چه نقاط پرت را افزایش دهیم، با مقیاس فاصله کولبک-لیبلر مدل بتا از مستطیلی بتا فاصله می‌گیرد. به ازای نقاط پرت بیشتر، مدل بتا کارایی مناسبی ندارد. اما با افزایش این نقاط، مقادیر فاصله کولبک-لیبلر برای مدل مستطیلی بتا دامنه زیادی را دربر نمی‌گیرد. بر اساس معیار کولبک-لیبلر، مدل رگرسیون مستطیلی بتا نسبت به نقاط پرت تنومند است. در نتیجه می‌توان گفت در مقایسه دو مدل، مستطیلی بتا عملکرد بهتری نسبت به مدل بتا دارد.

بررسی حساسیت بر معیارهای مقایسه مدل

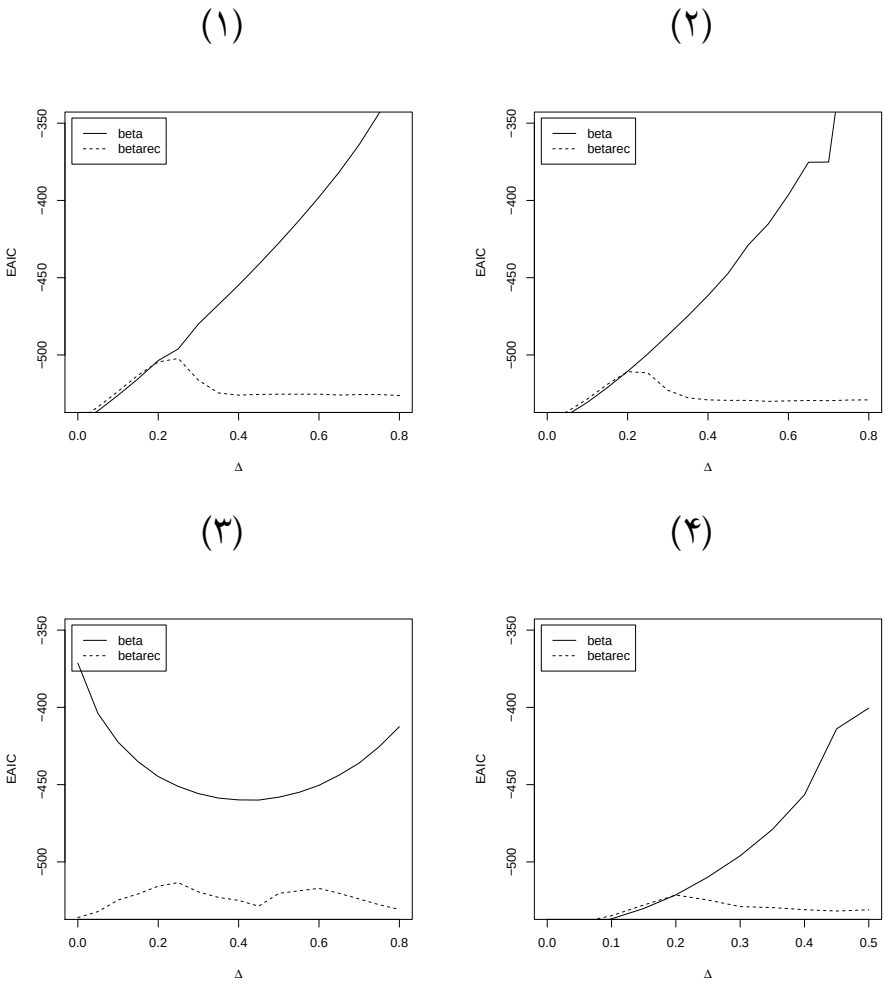
نتیجه بررسی رفتار DIC تحت مشاهدات دورافتاده در شکل ۱۳.۳ نمایش داده شده است. همان‌طور که در شکل می‌بینید، مدل بتا با خط و مستطیلی بتا با خط‌چین به ازای مقادیر DIC نشان داده شده‌اند. بر



شکل ۱۳.۳: معیار اطلاع انحراف طبق مدل‌های رگرسیونی بتا و مستطیلی بتا برای مقادیر گوناگون Δ تحت الگوهای متفاوت نقاط پرت برای داده‌های آلوده‌شده

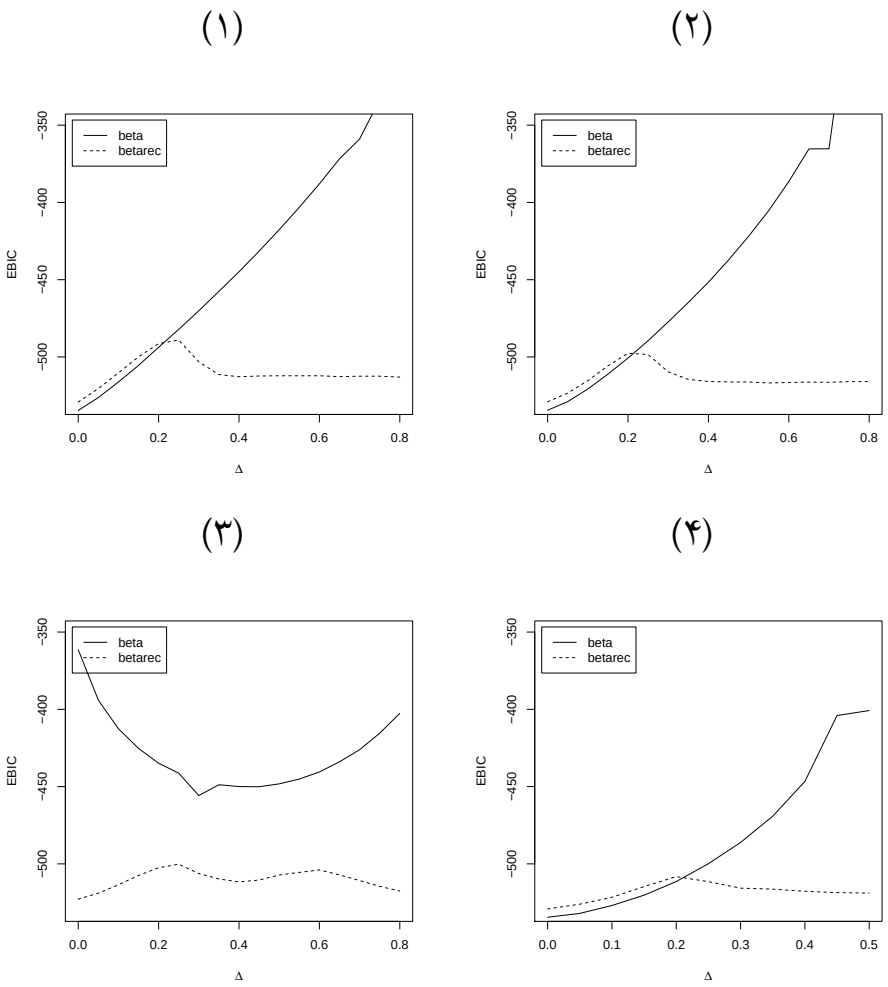
اساس شکل ۱۳.۳، معیار DIC برای آلودگی در بازه $(\Delta < ۰/۳)$ مدل رگرسیون بتا را انتخاب می‌کند. در این بازه نمودارهای دو مدل بتا و مستطیلی بتا تقریباً رفتار یکسانی دارند. دلیل این امر آن است

که مشاهده آلوده‌شده در این بازه، نقاط پرت چندانی را تولید نمی‌کند. همچنین DIC برای چهار نمودار الگوی اختلال کارایی مشابهی را برای موارد آلوده‌شده دارد. مشاهده می‌کنید در تمامی الگوهای اختلال، نمودار مربوط به مدل مستطیلی بتا در پایین نمودار مربوط به مدل بتا قرار می‌گیرد. برای آلودگی‌های بزرگتر ($\Delta > ۰/۳$)، معیار DIC مدل رگرسیون مستطیلی بتا را برمی‌گزینند.



شکل ۱۴.۳: مدل‌های رگرسیونی بتا و مستطیلی بتا برای مقادیر گوناگون $EAIC$ تحت الگوهای متفاوت نقاط پرت برای داده‌های آلوده‌شده

شکل ۱۴.۳، معیار $EAIC$ را برای مقادیر مختلف Δ مورد بررسی قرار می‌دهد. همان‌طور که ملاحظه می‌کنید، به ازای الگوهای ۱، ۲ و ۴ برای مقادیر ($\Delta < ۰/۲$) مدل بتا و مستطیلی بتا عملکرد مشابهی دارند. در حالی که، برای ($\Delta > ۰/۲$) مقادیر $EAIC$ برای مدل بتا با افزایش اختلال، افزایش می‌یابد. امامدل مستطیلی بتا نسبت به اختلالات متفاوت، پایدار است. در الگوی ۳ معیار $EAIC$ نسبت به الگوهای دیگر متفاوت است. در این الگو برای تمامی مقادیر اختلال مدل مستطیلی بتا نسبت به مدل بتا عملکرد بهتری دارد. اما در تمامی الگوهای اختلال، با افزایش نقاط پرت، مقادیر معیار $EAIC$ به ازای مدل رگرسیون مستطیلی بتا پایدار است. در مدل بتا که با خط نشان داده شده است، با افزایش نقاط پرت، مقادیر معیار $EAIC$ آن نیز روندی صعودی را طی می‌کند. به‌عبارتی دیگر، مدل



شکل ۱۵.۳: مدل‌های رگرسیونی بتا و مستطیلی بتا برای مقادیر گوناگون $EBIC$ تحت الگوهای متفاوت نقاط پرت برای داده‌های آلوده‌شده

بتا نسبت به افزایش نقاط پرت به شدت حساس است.

معیار $EBIC$ در شکل ۱۵.۳ چهار الگوی مختلف را نسبت به دو مدل مورد بحث، ارزیابی می‌کند. در این نمودار، در همه الگوهای اختلال نمودار مربوط به مدل مستطیلی بتا (خط‌چین) زیر نمودار مدل بتا (خط) قرار دارد. نمودارهای مربوط به الگوهای ۱، ۲ و ۴ تقریباً مشابه هستند. به ازای ($\Delta < ۰/۲$) نمودار مدل بتا و مستطیلی بتا مشابه هستند و در این حالت، مدل بتا عملکرد بهتری دارد. در صورتی که به ازای ($\Delta > ۰/۲$) مدل مستطیلی بتا بازه مقادیر کوچکی دارد و مدل بتا نموداری صعودی دارد.

بنابراین، هر چه تعداد نقاط پرت برای الگوهای مختلف بیشتر شود، معیارهای مقایسه مدل با احتمال بیشتری مدل مستطیلی بتا را برمی‌گزینند.

فصل ۴

کاربرد مدل رگرسیون مستطیلی بتا

در این فصل، تحلیل بیزی مدل رگرسیون مستطیلی بتا را با دو مثال واقعی تشریح می‌کنیم. این مثال‌ها، شامل نسبت چربی بدن ورزشکاران قایقرانی به جرم بدن آن‌ها در مؤسسه ورزش استرلیا (ais) و تأثیر عواملی چون، آب،[۱] ماسه[۲]، سیمان[۳] و میکروسیلیس[۴] بر مقاومت بتن هستند.

برای برازش مدل‌های مورد نظر در این فصل، ۵۰۰۰۰ از ۱۰۰۰۰۰ نمونه زنجیر $MCMC$ را با گام ۱۰ می‌سوزانیم. بنابراین، استنباط پسین بر اساس نمونه‌ای به حجم ۵۰۰۰ انجام شده است. همچنین، در این فصل در هر دو مدل بتا و مستطیلی بتا برای حالتی که پارامتر دقت ثابت فرض شده، توزیع پیشین ϕ، $gamma(c, c)$ با مقادیر $c = ۰/۱, ۰/۰۱, ۰/۰۰۱$ در نظر گرفته شده است.

۱.۴ مثال ۱: عوامل مؤثر بر جرم بدن قایقران‌ها

از نظر عامه مردم، ورزشکاران افرادی هستند که وزن بدن آنها در محدوده‌ای مشخص، تنظیم و به آسانی حفظ می‌گردد. افزایش وزن یا چربی در بدن برای سلامتی افراد زیانبار است، ولی به طور معمول، میزان چربی مطلوب و مورد قبول برای ورزشکاران کمتر از مقادیر در نظر گرفته‌شده برای افراد عادی است. عوامل متفاوتی برای ذخیره چربی در بدن افراد مؤثر هستند. نمونه‌ای از این عوامل شامل جنسیت، قد، وزن و غیره است.

مؤسسه ورزش استرلیا تعدادی از این عوامل را بر روی مجموعه‌ای از افراد بررسی کرده و تأثیر هر عامل را بر چربی بدن ورزشکاران، اندازه گرفته است. داده‌های حاصل از این پژوهش به عنوان داده‌های ais در بسته sn نرم‌افزار R موجود است. این داده‌ها متعلق به ۱۰۲ مرد و ۱۰۰ زن ورزشکار استرلیایی است که توسط ریچارد تلفورد و راس کانینگام گردآوری شده‌اند. قالب این ۲۰۲ داده، تحت ۱۳ متغیر شکل گرفته است. این متغیرها شامل جنسیت (عاملی با دو سطح مذکر و مؤنث)، ورزش (عاملی

[۱] Water
[۲] Sand
[۳] Cement
[۴] Silicafume

با سطوح شنا، دو میدانی، نت بال، واترپلو و...)، تعداد گلبول قرمز[۵] (RCC)، تعداد گلبول سفید[۶] (WCC)، هماتکریت[۷] (Hc)، هموگلبین[۸] (Hg)، آهن خون[۹] (Fe)، شاخص توده بدن[۱۰] (BMI)، مجموع سلول‌های پوستی[۱۱] (SSF)، نسبت چربی بدن[۱۲] $(Bfat)$، جرم بدن[۱۳] (LBM)، قد[۱۴] (Ht) و وزن[۱۵] (Wt) هستند.

در این مثال، داده‌های مربوط به ۳۷ ورزشکار را از مجموعه داده ais انتخاب کرده‌ایم. در مدل‌بندی این داده‌ها تنها علاقه‌مند به پیش‌بینی درصد چربی بدن $(Bfat)$ هر ورزشکار نسبت به جرم بدن (lbm) آن‌ها هستیم و سایر متغیرهای تبیینی را حذف می‌کنیم. در این بخش، داده‌های چربی بدن ورزشکاران را برای پارامتر دقت ثابت و متغیر در دو حالت بررسی می‌کنیم.

۱.۱.۴ پارامتر دقت ثابت

در این حالت، مدل رگرسیونی مستطیلی بتا با متغیر پاسخ $Bfat_i$ و متغیر توضیحی lbm_i، برای $i = ۱, \cdots, n$ به‌صورت

$$Bfat_i \sim BRr(\gamma_i, \phi, \alpha), \quad logit(\gamma_i) = \beta_۰ + \beta_۱ lbm_i, \quad i = ۱, ۲, \cdots, n$$

و مدل رگرسیون بتا (در صورتی که پارامتر آمیختگی مدل مستطیلی بتا صفر باشد)، با پارامتر دقت ثابت بر داده‌ها برازش می‌شوند. نتایج به‌دست آمده از برازش مدل‌ها، در جدول ۱.۴ ثبت شده‌اند. میانگین پسین پارامترها و همچنین فاصله اعتبار ۹۵٪ آن‌ها برای همه داده‌ها و با حذف نقاط دورافتاده، در این جدول مشهود است. همان‌طور که ملاحظه می‌کنید، بازه اعتبار پارامترهای برآوردشده با همه مشاهدات، عریض‌تر از بازه‌های متناظر با حذف نقاط پرت است.

با محاسبه DIC، $EAIC$ و $EBIC$ در جدول ۲.۴، مدل رگرسیون مستطیلی بتا برای همه مشاهدات، نسبت به مدل بتا برازش بهتری دارد. همچنین، زمان اجرای برنامه برای مدل مستطیلی بتا ۲۹۰٫۳۴ ثانیه بیشتر از مدل بتا است. حال، نقاط پرت را از داده‌ها حذف کرده و مدل‌ها را بر داده‌ها برازش می‌دهیم. زمان اجرای برنامه با حذف نقاط پرت ۲۷۹٫۵۱ ثانیه برای مدل مستطیلی بتا بیشتر از مدل بتا است. طبق جدول ۲.۴، مدل بتا در غیاب نقاط پرت نسبت به مدل دیگر برازش بهتری دارد. اگرچه زمان اجرای برازش مدل مستطیلی بتا بیشتر از مدل بتا است، اما با توجه به این‌که داده‌های پرت

[۵] Red Cell Count
[۶] White Cell Count
[۷] Hematocrit
[۸] Hemoglobin
[۹] Ferritin
[۱۰] Body Mass Index
[۱۱] Sum of Skin Folds
[۱۲] Body Fat Percentage
[۱۳] Lean Body Mass
[۱۴] Height
[۱۵] Weight

جدول ۱.۴: میانگین پسین و فاصله اعتبار ۹۵٪ برای مدل‌های رگرسیون بتا و مستطیلی بتا با پارامتر دقت ثابت برای داده‌های ais در حضور و عدم‌حضور نقاط پرت

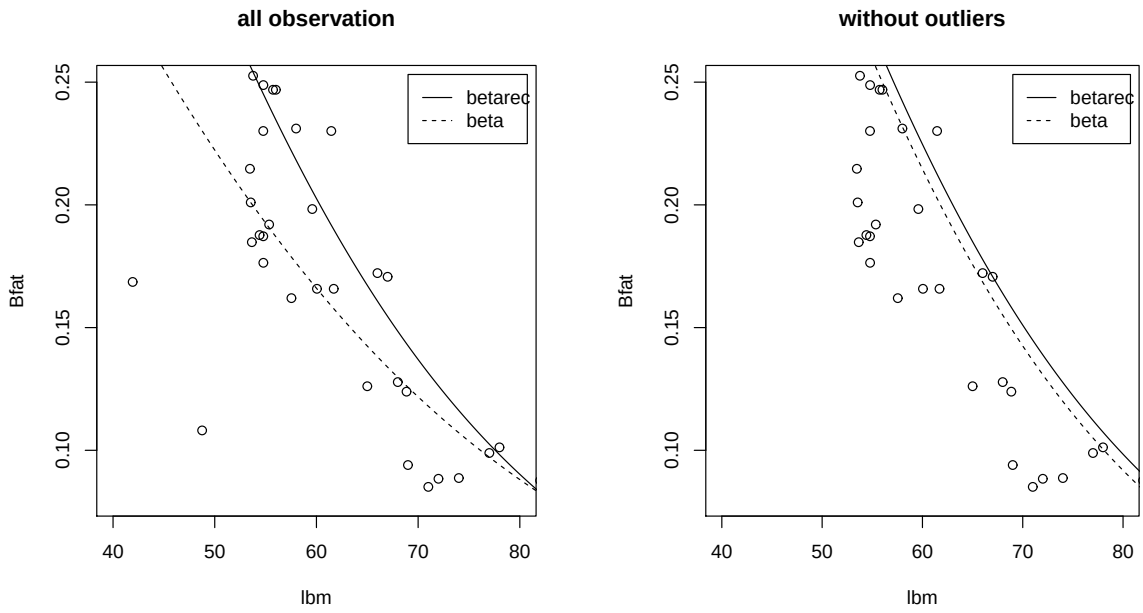
		بتا			مستطیلی بتا		
مشاهدات	پارامتر	میانگین	٪۲/۵	٪۹۷/۵	میانگین	٪۲/۵	٪۹۷/۵
همه	β_0	−۱/۷۵۴	−۱/۸۶۲	−۱/۶۴۳	−۱/۷۳۷	−۱/۷۲۹	−۱/۳۲۲
	β_1	−۰/۰۳۶	−۰/۰۴۷	−۰/۰۲۶	−۰/۰۴۷	−۰/۰۵۷	−۰/۰۳۶
	ϕ	۵۳/۲۳۳	۳۱/۶۸۰	۸۰/۶۰۱	۱۴۳/۶۳۴	۶۸/۹۵۱	۲۲۶/۴۸۴
	α				۰/۱۹۴	۰/۰۳۶	۰/۴۰۵
بدون نقاط پرت	β_0	−۱/۷۸۰	−۱/۸۶۹	−۱/۶۹۸	−۱/۷۱۵	−۱/۸۳۳	−۱/۵۵۶
	β_1	−۰/۰۴۹	−۰/۰۵۹	−۰/۰۴۱	−۰/۰۴۸	−۰/۰۵۸	−۰/۰۴۰
	ϕ	۱۴۵/۹۵۷	۸۶/۳۶۸	۲۲۶/۶۸۶	۱۴۷/۷۲۲	۸۵/۳۳۷	۲۲۶/۹۹۷
	α				۰/۰۷۷	۰/۰۰۲	۰/۲۵۳

جدول ۲.۴: مقایسه مدل‌های بتا و مستطیلی بتا در حضور و عدم‌حضور نقاط پرت با ϕ ثابت و زمان اجرا (بر حسب ثانیه) بر اساس معیارهای DIC، $EAIC$ و $EBIC$

مشاهدات	مدل	زمان اجرا	DIC	$EAIC$	$EBIC$
همه مشاهدات	بتا	۱۶۱/۶۹	−۱۳۳/۶۰۳۷	−۱۲۷/۶۰۳۷	−۱۲۲/۷۷۱
	مستطیلی بتا	۴۵۲/۰۳	−۱۴۸/۹۴۰۱	−۱۴۰/۹۴۰۱	−۱۳۴/۴۹۶۴
حذف نقاط پرت	بتا	۱۳۵/۸۳	−۱۵۴/۳۴۵۶	−۱۴۶/۲۳۹۵	−۱۴۱/۴۰۶۷
	مستطیلی بتا	۴۱۵/۳۴	−۱۵۲/۷۴۱۳	−۱۴۴/۷۲۱۱	−۱۳۹/۲۷۴

می‌توانند از ذات خود مدل آمده باشند و اتفاقاً نقاط مهمی محسوب شوند، برای به‌دست آوردن نتایج کاراتر پرداخت این هزینه می‌تواند توجیه داشته باشد.

شکل ۱.۴ نتیجه برازش مدل‌های مذکور بر داده‌ها است. توجه کنید، در نمودار سمت چپ مدل بتا تحت تأثیر نقاط پرت از نمودار مدل مستطیلی بتا فاصله می‌گیرد. با توجه به شکل و تعریف نقاط پرت ملاحظه می‌کنید که دو نقطه سمت چپ نمودار سمت چپ نسبت به سایر نقاط انحراف زیادی دارد و از روند سایر مشاهدات پیروی نمی‌کند، همین امر باعث می‌شود که این مشاهدات را نقاط پرت در نظر بگیریم. با حذف نقاط پرت از داده‌ها مدل‌ها را بر داده‌های جدید برازش می‌دهیم. همان‌طور که در نمودار سمت راست می‌بینید، دو مدل برازش یکسانی دارند. در نتیجه مدل بتا نسبت به نقاط پرت حساس است. مدل رگرسیون مستطیلی بتا در مواجهه با نقاط پرت، نیرومند بوده و کارایی بهتری دارد. در حضور نقاط پرت مدل مستطیلی بتا و در عدم‌حضور نقاط دورافتاده مدل بتا را برمی‌گزینیم.



شکل ۱.۴: نمودار پراکندگی داده‌های با خطوط رگرسیونی برازش داده‌شده تحت دو مدل رگرسیونی بتا و مستطیلی بتا در حضور و عدم‌حضور نقاط پرت

۲.۱.۴ پارامتر دقت متغیر

اکنون با فرض متغیر بودن ϕ، مدل رگرسیونی مستطیلی بتا به‌صورت

$$Bfat_i \sim BRr(\gamma_i, \phi, \alpha), \quad logit(\gamma_i) = \beta_0 + \beta_1 lbm_i,$$
$$log(\phi_i) = \delta_0 + \delta_1 lbm_i \quad i = 1, 2, \cdots, n$$

خواهد بود. این مدل و مدل رگرسیونی بتا به‌دست آمده از این مدل را نیز برازش می‌دهیم. برازش پارامترهای مدل‌های رگرسیونی بتا و مستطیلی بتا با تمامی مشاهدات و عدم‌حضور نقاط دورافتاده در جدول ۳.۴ گزارش شده‌اند. با مقایسه جداول ۱.۴ و ۳.۴ ملاحظه می‌کنید که برآورد ضرایب رگرسیون با پارامتر دقت ثابت و متغیر تقریباً مشابه هم می‌باشند.

به کمک معیارهای DIC، $EAIC$ و $EBIC$ مدل‌های مورد بحث را در جدول ۴.۴ نسبت به هم قیاس می‌کنیم. با نظر به جدول و تغییر پراکندگی در مدل‌ها و مقادیر حاصل از معیارهای سنجش مدل، درمی‌یابیم که مدل مستطیلی بتا برازش بهتری دارد. اما، با حذف نقاط پرت از داده‌ها، مدل بتا برازش بهتری خواهد داشت. در حالت خاص، مدلی از بتا بهتر است که پارامتر دقت آن ثابت فرض می‌شود. در حالت کلی، مدل‌ها با ثابت فرض کردن پارامتر دقت بهبود می‌یابند.

جدول ۳.۴: میانگین پسین و فاصله اعتبار ۹۵٪ برای مدل‌های رگرسیون بتا و مستطیلی بتا با ϕ متغیر طبق داده‌های ais در حضور و عدم‌حضور نقاط پرت

		بتا			مستطیلی بتا		
مشاهدات	پارامتر	میانگین	٪۲/۵	٪۹۷/۵	میانگین	٪۲/۵	٪۹۷/۵
همه	β_0	−۱/۷۶۵	−۱/۸۶۲	−۱/۶۶۲	−۱/۶۳۹	−۱/۵۶۴	−۱/۳۷۹
	β_1	−۰/۰۴۶	−۰/۰۵۴	−۰/۰۳۶	−۰/۰۴۷	−۰/۰۴۳	−۰/۰۳۶
	δ_0	−۴/۵۳۹	−۴/۹۹۸	−۴/۰۴۱	−۴/۷۲۱	−۴/۴۷۹	−۴/۱۲۸
	δ_1	−۰/۰۶۴	−۰/۱۰۳	−۰/۰۲۳	−۰/۰۴۳	−۰/۰۱۷	۰/۰۳۵
	α				۰/۱۲۹	۰/۰۰۴	۰/۳۶۴
بدون نقاط پرت	β_0	−۱/۷۷۹	−۱/۸۵۷	−۱/۶۹۹	−۱/۷۱۸	−۱/۸۳۲	−۱/۵۴۳
	β_1	−۰/۰۴۹	−۰/۰۵۸	−۰/۰۴۱	−۰/۰۴۹	−۰/۰۵۸	−۰/۰۳۹
	δ_0	−۵/۰۱۳	−۵/۴۷۹	−۴/۴۹۳	−۵/۰۳۴	−۵/۴۹۲	−۴/۵۳۳
	δ_1	−۰/۰۰۵	−۰/۰۵۷	۰/۰۵۱	−۰/۰۰۶	−۰/۰۵۶	۰/۰۴۶
	α				۰/۰۷۴	۰/۰۰۲	۰/۲۴۲

جدول ۴.۴: مقایسه مدل‌های بتا و مستطیلی بتا در حضور و عدم‌حضور نقاط پرت با ϕ متغیر و زمان اجرا (بر حسب ثانیه) با معیارهای DIC، $EAIC$ و $EBIC$ برای داده‌های ais

مشاهدات	مدل	زمان اجرا	DIC	$EAIC$	$EBIC$
همه مشاهدات	بتا	۲۴۶/۹۱	−۱۴۲/۲۵۳	−۱۳۴/۲۵۹	−۱۲۷/۸۰۹
	مستطیلی بتا	۵۸۳/۳۰	−۱۴۴/۳۴۷	−۱۳۵/۶۳۱	−۱۲۹/۵۷۶
حذف نقاط پرت	بتا	۱۸۸/۳۶	−۱۵۱/۳۷۶	−۱۴۳/۵۶۶	−۱۳۷/۱۲۳
	مستطیلی بتا	۴۹۵/۱۴	−۱۵۰/۹۲۱۷	−۱۴۰/۷۲۱۱	−۱۳۲/۲۷۴

۲.۴ مثال ۲: عوامل مؤثر بر مقاومت بتن

یکی از مهم‌ترین و متداول‌ترین مصالح ساختمانی، بتن است که به علت دارا بودن خواصی از جمله مقاومت خوب در برابر آتش‌سوزی، دسترسی آسان به مصالح و مقاومت فشاری خوب، استفاده از آن را با مقبولیت عمومی روبرو کرده است.

بتن به هر ماده یا ترکیبی که از یک ماده چسبنده با خاصیت سیمانی‌شدن تشکیل شده باشد، گفته می‌شود. این ماده چسبنده، عموماً حاصل فعل و انفعال سیمان‌های هیدرولیکی و آب می‌باشد که سخت شدن مخلوط بتن را پس از رسیدن به مقاومت نهایی بتن به همراه دارد. امروزه، چنین تعریفی از بتن شامل طیف وسیعی از محصولات می‌شود. بتن ممکن است، از انواع مختلف سیمان و نیز پوزالان‌ها، سرباره کوره‌ها، مواد مضاف، گوگرد، مواد افزودنی، پلیمرها، الیاف و غیره تهیه شود. همچنین، در نحوه

ساخت آن ممکن است بخارآب، اتوکلاو، خلأ، فشارهای هیدرولیکی و متراکم‌کننده‌های مختلف استفاده شود.

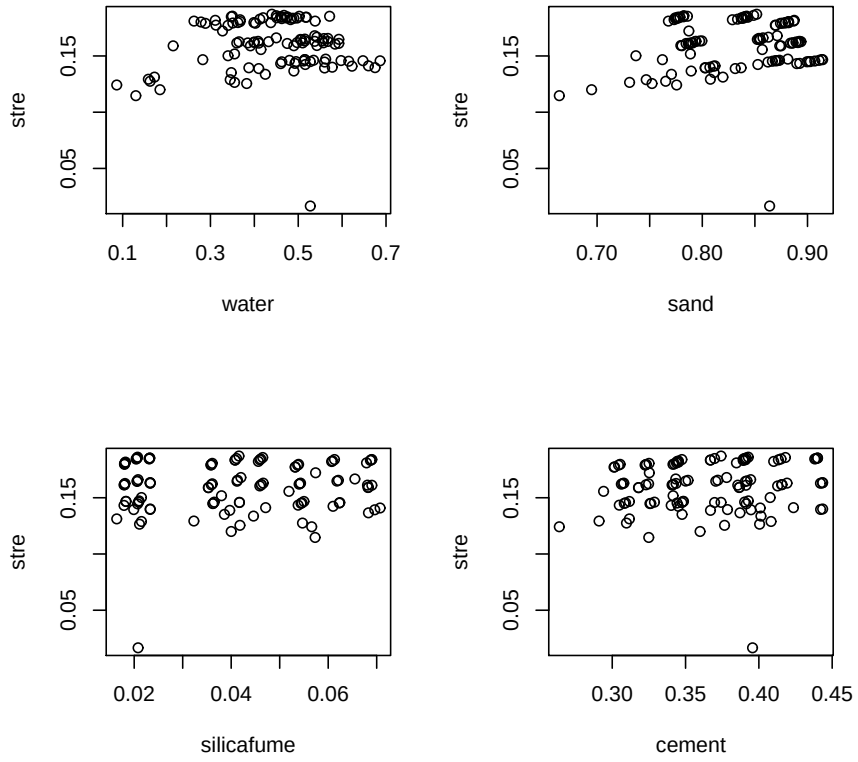
با توجه به گسترش و پیشرفت علم و پیدایش تکنولوژی‌های فراوان در قرن اخیر، شناخت بتن و خواص آن نیز توسعه قابل ملاحظه‌ای داشته است، به نحوی که امروزه شاهد کاربرد انواع مختلف بتن با مصالح مختلف هستیم، که هر یک خواص و کاربردی مخصوص به خود را داراست. در حال حاضر، انواع مختلفی از سیمان‌ها که شامل پوزالان‌ها، سولفورها، پلیمرها، الیاف‌های مختلف و افزودنی‌های متفاوتی هستند، تولید می‌شوند. همچنین، می‌توان خاطرنشان کرد که تولید انواع بتن با استفاده از حرارت، بخار، اتوکلاو، تخلیه هوا، فشار هیدرولیکی ویبره و قالب انجام می‌گیرد. بتن به طور کلی محصولی است که از اختلاط آب با سیمان آبی و سنگدانه‌های مختلف، در اثر واکنش آب با سیمان در شرایط محیطی خاصی حاصل می‌شود و دارای ویژگی‌های خاص است. بتن اینک با گذشت بیش از ۱۷۰ سال از پیدایش سیمان پرتلند به‌صورت کنونی توسط یک بنای لیدزی، دستخوش تحولات و پیشرفت‌های شگرفی شده است. در دسترس بودن مصالح آن، دوام نسبتاً زیاد و نیاز به ساخت و سازهای فراوان سازه‌های بتنی چون ساختمان‌ها، سازه‌ها، سدها، پل‌ها، تونل‌ها و راه‌ها این ماده را بسیار پرمصرف نموده است. اینک، حدود سه تا چهار دهه است که کاربرد این ماده در شرایط خاص مورد استقبال کاربران آن قرار گرفته است.

امروزه، با پیشرفت علم و تکنولوژی مشخص شده است که صرف توجه به مقاومت به عنوان یک معیار برای طرح بتن برای محیط‌های مختلف و کاربردهای مختلف نمی‌تواند جوابگوی مشکلاتی باشد که در دراز مدت در سازه‌های بتنی ایجاد می‌گردد.

چند سالی است که مسأله دوام بتن در محیط‌های مختلف مورد توجه قرار گرفته است. مشاهده خرابی‌هایی با عوامل فیزیکی و شیمیایی در بتن‌ها در اکثر نقاط جهان و با شدتی بیشتر در کشورهای در حال توسعه، افکار و اذهان را به سمت طرح بتن‌هایی با ویژگی خاص و بادوام لازم سوق داده است. در این راستا، در پاره‌ای از کشورها دستورالعمل‌ها و استانداردهایی نیز برای طرح بتن با عملکرد بالا تحصیل شده و طراحان و مجریان در بعضی از کشورهای پیشرفته ملزم به رعایت این دستورالعمل‌ها گشته‌اند.

در این قسمت، تأثیر عوامل متفاوتی چون، آب، سیمان، ماسه و میکروسیلیس را بر مقاومت بتن بررسی می‌کنیم (عجم، ۱۳۹۳؛ ولی‌پور، ۱۳۹۲).

شکل ۲.۴ پراکنش حاشیه‌ای مقاومت بتن را در برابر متغیرهای تبیینی آب، ماسه، سیمان و میکروسیلیس نمایش می‌دهد. مشاهده می‌کنید که در تمامی این پراکنش‌ها حضور یک داده پرت کاملاً واضح است. بنابراین این مجموعه داده برای مقایسه برازش دو مدل رگرسیون بتا و مستطیلی بتا جالب است. همانند مثال اول در این قسمت، داده‌های ساخت بتن را برای دو حالت ثابت و متغیر ϕ انجام می‌دهیم.



شکل ۲.۴: نمودار پراکنش داده‌های آب، ماسه، سیمان و میکروسیلیس در مقابل مقاومت بتن

۱.۲.۴ پارامتر دقت ثابت

در این قسمت، مدل بتا و مستطیلی بتا را بر اساس پارامتر دقت ثابت برازش می‌دهیم. جهت تعیین تابع پیوند و تأثیر عوامل پیشنهادی بر مقاومت بتن، مدل‌های بتا و مستطیلی بتا را با ماسه، آب، سیمان و میکروسیلیس و ترکیب‌هایی از این عوامل به‌عنوان متغیرهای تبیینی برازش دادیم. توجه کنید که در این‌جا متغیرهای تبیینی را مستقل فرض کردیم و اثرهای متقابل را در نظر نگرفتیم. مقادیر DIC این مدل‌ها در جدول ۵.۴ گزارش شده‌اند.

با توجه به جدول ۵.۴، ملاحظه می‌کنید که در تمامی مدل‌ها، مستطیلی بتا بر اساس معیار DIC برازش بهتری نسبت به بتا دارد. برای مدلی با تنها یک متغیر تبیینی، یعنی مدل‌های ۱ تا ۴، بر اساس مقادیر DIC، مدل مستطیلی بتا برای متغیر ماسه کمترین مقدار را دارد. پس، ماسه نسبت به سایر عوامل، برازش بهتری را برای مستطیلی بتا دارد. این در حالی است که برای مدل بتا مدلی با حضور سیمان برازش بهتری را دارد. اگر بخواهیم تأثیر دو عامل را در مقاومت بتن (مدل‌های ۵ تا ۱۰) در نظر بگیریم، مدل نهم با عوامل سیمان و ماسه بهترین برازش را برای مدل مستطیلی بتا دارد و مدل هشتم برای بتا بهترین برازش را نتیجه می‌دهد. با نظر به مدل‌هایی با سه متغیر تبیینی (مدل‌های ۱۱ تا ۱۴)، درمی‌یابیم که پس از عواملی چون ماسه و سیمان، آب نیز در مقاومت بتن در مدل‌های بتا و مستطیلی بتا مؤثر است. در میان عواملی که در دوام بتن مؤثرند، مدلی که در آن میکروسیلیس حضور دارد کمترین

جدول ۵.۴: مقایسه مدل‌های مرکب از متغیرهای تبیینی آب، ماسه، میکروسیلیس و سیمان بر اساس معیار DIC

مدل	DIC		
	$logit(\gamma_i)$	بتا	مستطیلی بتا
۱	$\beta_0 + \beta_1 water$	−۴۴۴/۶۰۶۶	−۵۰۳/۴۷۴۴
۲	$\beta_0 + \beta_2 sand$	−۴۴۵/۱۳۶۶	−۵۰۶/۶۶۷۳
۳	$\beta_0 + \beta_3 silicafume$	−۴۴۵/۵۳۴۵	−۴۹۶/۵۴۰۳
۴	$\beta_0 + \beta_4 cement$	−۴۴۷/۴۵۱۶	−۴۹۹/۵۰۵۱
۵	$\beta_0 + \beta_1 water + \beta_2 sand$	−۴۲۸/۴۸۵۶	−۵۰۵/۷۵۰۵
۶	$\beta_0 + \beta_1 water + \beta_3 silicafume$	−۴۴۳/۶۱۴۶	−۵۰۲/۵۸۴۶
۷	$\beta_0 + \beta_1 water + \beta_4 cement$	−۴۴۲/۴۶۵۵	−۴۹۸/۸۶۷۸
۸	$\beta_0 + \beta_2 sand + \beta_3 silicafume$	−۴۴۴/۹۰۲۴	−۵۰۴/۸۰۴۷
۹	$\beta_0 + \beta_2 sand + \beta_4 cement$	−۴۴۳/۹۰۱۱	−۵۰۸/۶۴۵۰
۱۰	$\beta_0 + \beta_3 silicafume + \beta_4 cement$	−۴۴۳/۹۴۳۳	−۴۹۳/۶۳۷۱
۱۱	$\beta_0 + \beta_1 water + \beta_2 sand + \beta_3 silicafume$	−۴۴۲/۹۴۶۶	−۵۰۱/۵۷۹۸
۱۲	$\beta_0 + \beta_1 water + \beta_2 sand + \beta_4 cement$	−۴۴۱/۹۹۸۶	−۵۱۷/۳۱۰۶
۱۳	$\beta_0 + \beta_1 water + \beta_3 silicafume + \beta_4 cement$	−۴۴۲/۲۳۸۷	−۴۹۹/۳۹۰۶
۱۴	$\beta_0 + \beta_2 sand + \beta_3 silicafume + \beta_4 cement$	−۴۴۴/۸۰۷۳	−۵۱۶/۹۰۶۵
۱۵	$\beta_0 + \beta_1 water + \beta_2 sand + \beta_3 silicafume + \beta_4 cement$	−۴۵۰/۳۵۴۶	−۵۲۹/۲۲۳۸

اقبال را برای هر دو مدل داراست.

با حذف مشاهده پرت از سایر مشاهدات، مدل‌های بتا و مستطیلی بتا را برازش داده و تأثیر عوامل مختلف را در جدول ۶.۴ بر مقاومت بتن بررسی می‌کنیم. در جدول ۶.۴ مشاهده می‌کنید که در عدم حضور نقطه دورافتاده بر اساس مقادیر DIC، مدل بتا در تمامی ترکیب‌های حاصل از عوامل مؤثر بر مقاومت بتن برازش بهتری نسبت به مستطیلی بتا دارد. همچنین، نتایجی که از جدول ۵.۴ نسبت به تأثیر عوامل مؤثر در مقاومت بتن به‌دست آمد، برای این جدول نیز صادق است. اما مدل پانزدهم، یعنی مدلی که معرف تمامی عوامل بر مقاومت بتن است، بهترین برازش را نسبت به سایر مدل‌ها برای تمامی مشاهدات و با حذف نقطه دورافتاده برای بتا و مستطیلی بتا دارد.

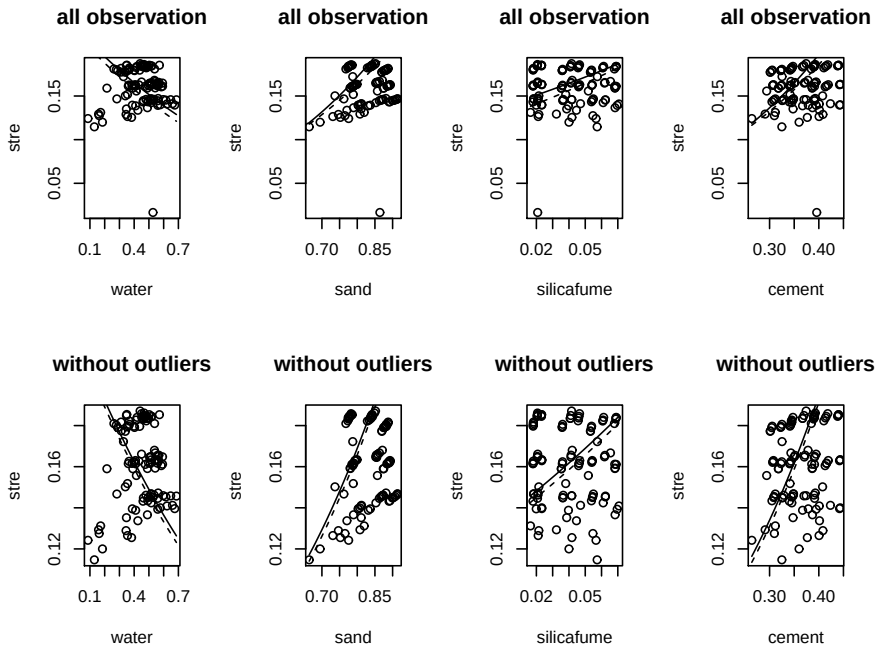
برای برازش این مدل، پارامتر دقت را برابر $\phi = exp(\delta_0)$ تعریف کرده و مدل مستطیلی بتا را به‌صورت

جدول ۶.۴: مقایسه مدل‌های مرکب از متغیرهای تبیینی آب، ماسه، میکروسیلیس و سیمان بر اساس معیار DIC با حذف نقطه پرت

مدل	DIC		
	$logit(\gamma_i)$	بتا	مستطیلی بتا
۱	$\beta_۰ + \beta_۱ water$	$-۵۳۸/۴۶۵۴$	$-۵۱۹/۴۳۴۹$
۲	$\beta_۰ + \beta_۲ sand$	$-۵۴۱/۳۹۳۴$	$-۵۲۲/۷۰۴۰$
۳	$\beta_۰ + \beta_۳ silicafume$	$-۵۳۶/۶۳۴۸$	$-۵۱۸/۱۵۷۲$
۴	$\beta_۰ + \beta_۴ cement$	$-۵۳۹/۳۶۶۹$	$-۵۲۰/۶۷۱۳$
۵	$\beta_۰ + \beta_۱ water + \beta_۲ sand$	$-۵۴۰/۵۱۶۴$	$-۵۲۳/۷۹۸۴$
۶	$\beta_۰ + \beta_۱ water + \beta_۳ silicafume$	$-۵۳۶/۵۱۲۰$	$-۵۱۶/۹۹۹۹$
۷	$\beta_۰ + \beta_۱ water + \beta_۴ cement$	$-۵۳۷/۵۲۹۹$	$-۵۲۰/۳۱۴۷$
۸	$\beta_۰ + \beta_۲ sand + \beta_۳ silicafume$	$-۵۳۹/۳۹۵۸$	$-۵۱۸/۲۲۳۳۳$
۹	$\beta_۰ + \beta_۲ sand + \beta_۴ cement$	$-۵۴۹/۵۴۲۱$	$-۵۲۹/۷۴۵۸$
۱۰	$\beta_۰ + \beta_۳ silicafume + \beta_۴ cement$	$-۵۳۷/۲۰۰۱$	$-۵۱۳/۸۴۷۱$
۱۱	$\beta_۰ + \beta_۱ water + \beta_۲ sand + \beta_۳ silicafume$	$-۵۳۸/۲۷۰۴$	$-۵۲۲/۱۹۶۶$
۱۲	$\beta_۰ + \beta_۱ water + \beta_۲ sand + \beta_۴ cement$	$-۵۵۱/۵۴۴۰$	$-۵۳۱/۹۲۰۷$
۱۳	$\beta_۰ + \beta_۱ water + \beta_۳ silicafume + \beta_۴ cement$	$-۵۳۵/۵۶۳۲$	$-۵۱۷/۰۱۹۲$
۱۴	$\beta_۰ + \beta_۲ sand + \beta_۳ silicafume + \beta_۴ cement$	$-۵۵۰/۵۳۸۵$	$-۵۳۱/۸۵۰۰$
۱۵	$\beta_۰ + \beta_۱ water + \beta_۲ sand + \beta_۳ silicafume + \beta_۴ cement$	$-۵۷۱/۴۵۴۴$	$-۵۵۰/۶۲۸۵$

$$
\begin{aligned}
&stre_i \sim BRr(\gamma_i, \phi, \alpha),\\
&logit(\gamma_i) = \beta_۰ + \beta_۱ water_i + \beta_۲ sand_i + \beta_۳ silicafume_i + \beta_۴ cement_i,\\
&\phi = exp(\delta_۰) \quad i = ۱, ۲, \cdots, n.
\end{aligned}
$$

در نظر می‌گیریم. مدل بتا را نیز برای این داده‌ها (حالتی که ضریب آمیختگی مدل BRr صفر باشد) برازش می‌دهیم. نتیجه برازش مدل‌های بتا و مستطیلی بتا برای فهم تأثیر عواملی چون ماسه، آب، سیمان و میکروسیلیس بر مقاومت بتن در نمودار ۳.۴ و جداول ۷.۴ و ۸.۴ ثبت شده‌اند. ابتدا برازش با تمامی مشاهدات انجام می‌شود. سپس مشاهده ۳۷ام که مشاهده پرت است را از مشاهدات حذف نموده و با سایر مشاهدات مدل‌های مورد نظر را برازش می‌دهیم. همان‌طور که در شکل ۳.۴ (نمودارهای بالای شکل) مشهود است، در صورتی که مدل‌های مورد نظر را بر تمامی مشاهدات برازش دهیم، نمودار مربوط به مدل بتا سمت نقطه پرت منحرف می‌گردد و با حذف نقطه پرت، دو مدل برازش تقریباً یکسانی خواهند داشت.



شکل ۳.۴: نمودار پراکنش داده‌های آب، ماسه، سیمان و میکروسیلیس در قبال مقاومت بتن و برازش مدل‌های بتا (خط‌چین) و مستطیلی بتا (خط) در حضور (نمودارهای سمت راست) و عدم‌حضور (نمودارهای سمت چپ) نقاط پرت برای پارامتر دقت ثابت (نمودارهای بالا) و متغیر (نمودارهای پایین)

جدول ۷.۴: مقایسه مدل‌های بتا و مستطیلی بتا در حضور و عدم حضور نقاط پرت، با پارامتر دقت ثابت و زمان اجرا (بر حسب ثانیه) با معیارهای DIC ، $EAIC$ و $EBIC$

مشاهدات	مدل	زمان اجرا	DIC	$EAIC$	$EBIC$
همه مشاهدات	بتا	۸۵۶/۷۶	−۴۵۰/۳۵۴۶	−۴۴۴/۵۵۶۳	−۴۲۸/۴۶۳۵
	مستطیلی بتا	۲۴۴۶/۲۰	−۵۲۹/۲۳۳۸	−۵۶۳/۹۸۶۲	−۵۴۵/۲۱۱۲
حذف نقاط پرت	بتا	۸۲۴/۶۶	−۵۷۱/۴۵۴۴	−۵۶۵/۴۲۳۸	−۵۴۹/۳۳۶۱
	مستطیلی بتا	۲۳۲۲/۳۱	−۵۵۰/۶۲۸۵	−۵۶۲/۹۹۳۱	−۵۴۴/۲۱۸۲

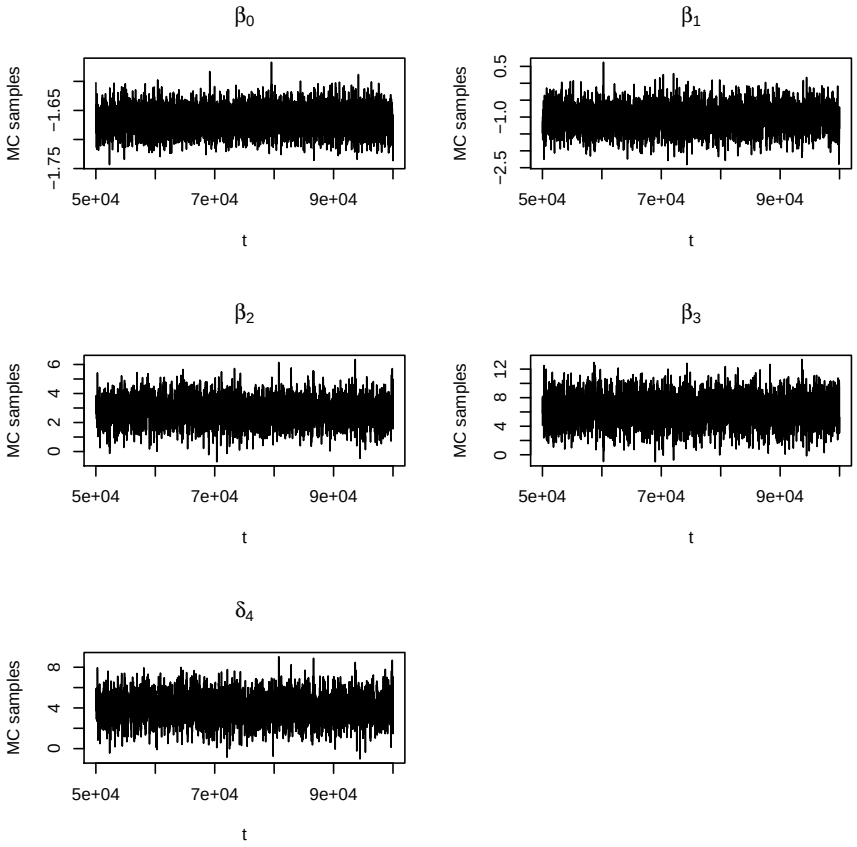
در جدول ۷.۴ مقادیر معیارهای مقایسه مدل برای مدل‌های بتا و مستطیلی بتا با ϕ ثابت، گزارش شده‌اند. طبق این مقادیر، مدل مستطیلی بتا به ازای تمامی مشاهدات برازش بهتری نسبت به مدل بتا دارد. با حذف نقطه دورافتاده مدل بتا برازش بهتری یافته است.

جدول ۸.۴ میانگین پسین و فاصله اعتبار ۹۵٪ را برای پارامترهای مدل‌های رگرسیون بتا و مستطیلی بتا و پارامتر دقت ثابت، طبق داده‌های مقاومت بتن گزارش می‌دهد. این جدول، مدلی با تمامی عوامل مؤثر بر مقاومت بتن را نشان می‌دهد.

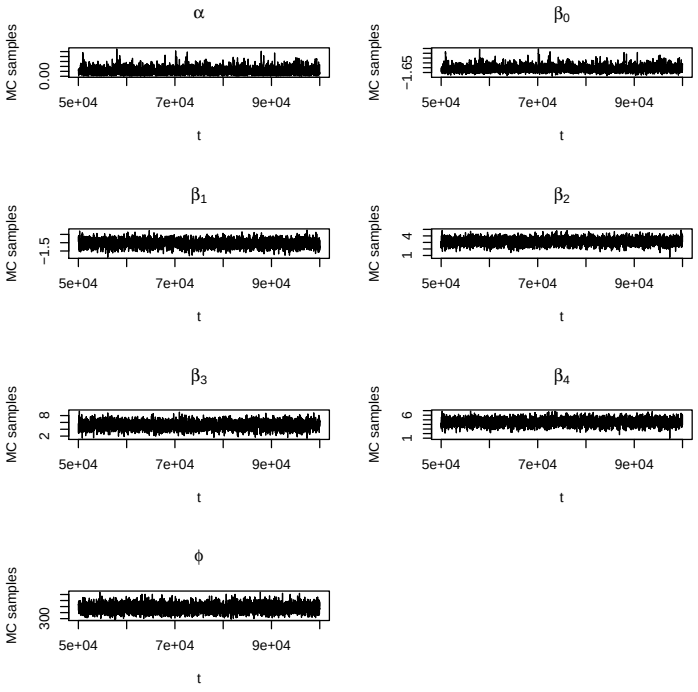
با مشاهده نمودار ۴.۴ و ۵.۴ که براساس زنجیر اول رسم شده‌اند، می‌توان گفت زنجیر اول برای همه پارامترها از آمیختگی خوبی برخوردار است، یعنی نواحی چگال توزیع پسین به خوبی مورد کاوش قرار گرفته‌اند. با استفاده از نمودارهای ۶.۴ و ۷.۴، نمودارهای میانگین تجمعی، می‌توان همگرایی هر پارامتر

جدول ۸.۴: میانگین پسین و فاصله اعتبار ۹۵٪ برای مدل‌های رگرسیون بتا و مستطیلی بتا با پارامتر دقت ثابت طبق داده‌های مقاومت بتن

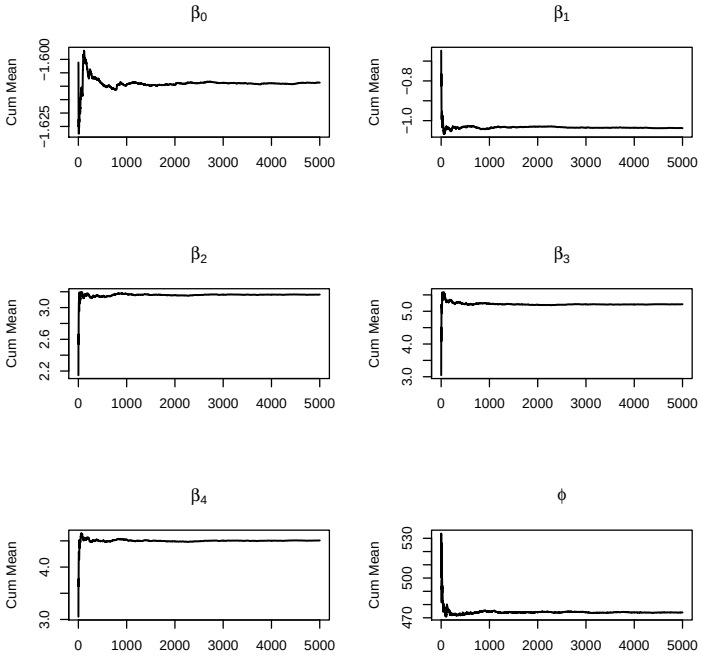
	بتا			مستطیلی بتا		
پارامتر	میانگین	٪۲٫۵	٪۹۷٫۵	میانگین	٪۲٫۵	٪۹۷٫۵
β_0	−۱٫۶۶۶۰	−۱٫۷۱۶۰	−۱٫۶۲۳۶	−۱٫۶۰۹۸	−۱٫۶۵۸۴	−۱٫۵۳۱۳
β_1	−۱٫۰۴۴۰	−۱٫۸۴۱۰	−۰٫۲۶۹۵	−۱٫۰۳۷۳	−۱٫۴۵۹۴	−۰٫۶۰۰۰
β_2	۲٫۷۶۳۰	۱٫۰۲۸۰	۴٫۶۲۴۰	۳٫۱۶۱۲	۲٫۱۹۱۹	۴٫۱۴۰۰
β_3	۶٫۱۲۴۰	۲٫۰۹۳۰	۱۰٫۰۷۹۴	۵٫۱۸۶۷۱	۲٫۹۹۲۵	۷٫۲۶۲۶
β_4	۳٫۹۶۶۰	۱٫۲۷۰۰	۶٫۵۵۷۲	۴٫۵۰۱۱	۳٫۰۱۷۳	۵٫۹۲۷۹
ϕ	۱۵۰٫۶۵۰۰	۱۱۲٫۶۲۱۰	۱۹۵٫۸۳۳۶	۴۷۴٫۰۲۹۳	۳۵۶٫۶۴۲۵	۶۱۶٫۴۷۲۸
α				۰٫۰۵۳۵	۰٫۰۰۷۵	۰٫۱۳۹۷



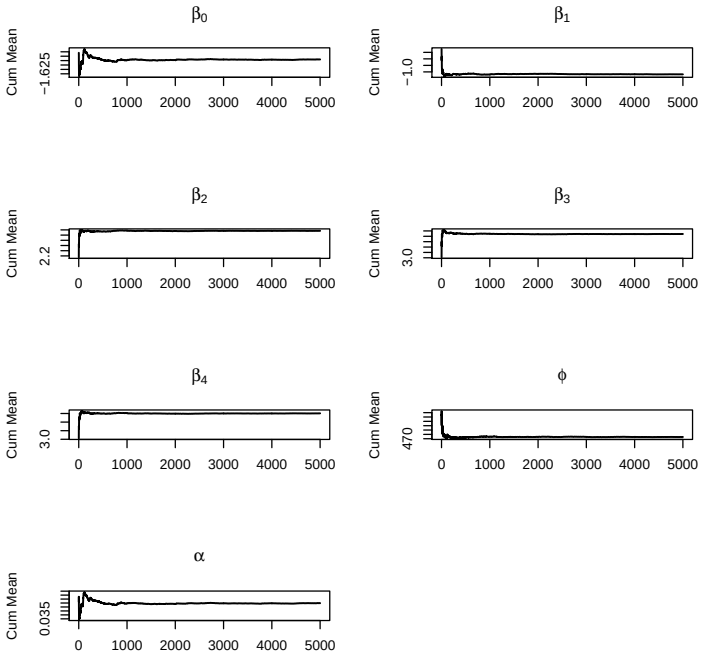
شکل ۴.۴: نمودارهای اثر پارامترهای مدل بتا با تمامی متغیرهای تبیینی برای ϕ ثابت و داده‌های مقاومت بتن



شکل ۵.۴: نمودارهای اثر پارامترهای مدل مستطیلی بتا با تمامی متغیرهای تبیینی برای ϕ ثابت و داده‌های مقاومت بتن



شکل ۶.۴: نمودارهای میانگین تجمعی پارامترهای مدل بتا با تمامی متغیرهای تبیینی برای ϕ ثابت و داده‌های مقاومت بتن



شکل ۷.۴: نمودارهای میانگین تجمعی پارامترهای مدل مستطیلی بتا با تمامی متغیرهای تبیینی برای ϕ ثابت و داده‌های مقاومت بتن

را به ازای مدل‌های بتا و مستطیلی بتا، به مقدار واقعی آن مشاهده کرد. با توجه به جدول ۹.۴ می‌بینید

جدول ۹.۴: نتایج آزمون گلمن-روبین بر داده‌های مقاومت بتن برای پارامتر دقت ثابت

مدل	پارامتر	برآورد آماره $PSRF$
بتا	β_0	۱/۰۰
	β_1	۱/۰۰
	β_2	۱/۰۰
	β_3	۱/۰۱
	β_4	۱/۰۰
	ϕ	۱/۰۱
مستطیلی بتا	β_0	۱/۰۰
	β_1	۱/۰۰
	β_2	۱/۰۰
	β_3	۱/۰۱
	β_4	۱/۰۰
	ϕ	۱/۰۰
	α	۱/۰۱

که مقدار آماره $PSRF$ برای پارامترهای مدل‌های بتا و مستطیلی بتا نزدیک یک و یک به‌دست آمده و در نتیجه زنجیر برای هر پارامتر همگرا است. همچنین در این برازش برای هر دو مدل بتا و مستطیلی بتا، $MPSRF = ۱ < ۱/۲$ می‌باشد که بیان‌کننده همگرایی زنجیر به توزیع مانا است.

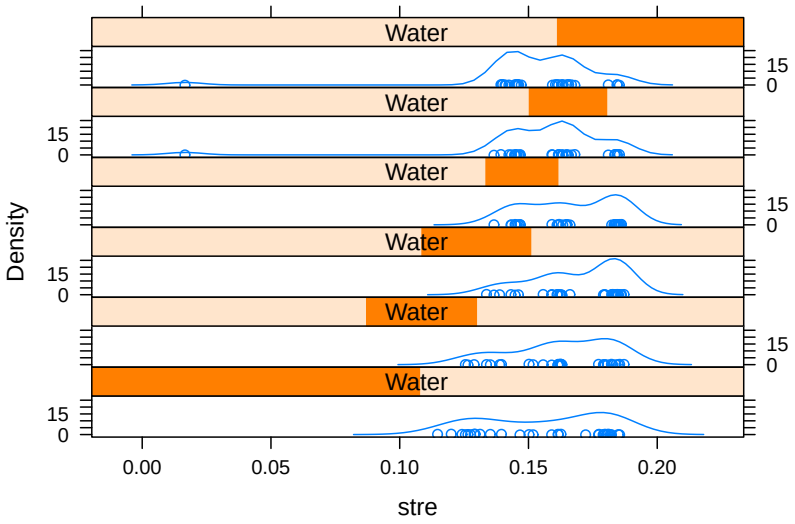
۲.۲.۴ پارامتر دقت متغیر

در این قسمت، مدل‌های مذکور را برای ϕ متغیر برازش می‌دهیم. پیش از برازش این مدل‌ها بهتر است روی داده‌ها، تحلیل اکتشافی انجام شود. برای درک آن‌که کدام متغیرهای تبیینی می‌توانند بر پارامتر دقت تأثیر داشته باشند، از تحلیل اکتشافی استفاده کردیم.

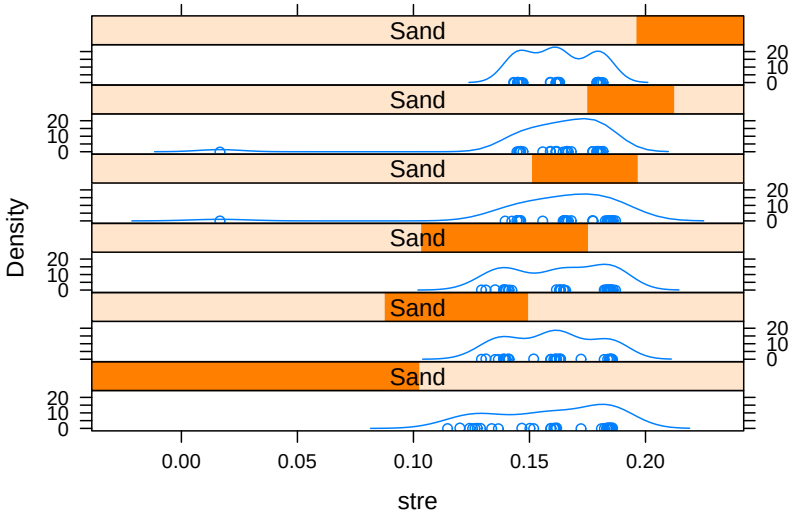
در تحلیل اکتشافی، پژوهشگر به دنبال بررسی داده‌های تجربی به منظور کشف و شناسایی شاخص‌ها و نیز روابط بین آن‌ها است و این کار را بدون تحلیل هرگونه مدل معینی انجام می‌دهد. به بیان دیگر، تحلیل اکتشافی علاوه بر آن‌که ارزش تجسسی یا پیشنهادی داشته باشد، می‌تواند ساختارساز، مدل‌ساز یا فرضیه‌ساز باشد و زمانی به‌کار می‌رود که پژوهشگر شواهد کافی و قبلی و پیش‌تجربی برای تشکیل فرضیه درباره تعداد عامل‌های زیربنایی داده‌ها نداشته و برای تعیین عامل‌هایی که هم‌پراشی بین متغیرها را توجیه می‌کنند، داده‌ها را می‌کاود. بنابراین، تحلیل اکتشافی بیشتر به‌عنوان یک روش ایجاد مدل در نظر گرفته می‌شود. این روش در جهت آزمون مدل به‌کار نمی‌رود. لذا برای تأکید نتیجه تحلیل اکتشافی و فرضیه‌ای که از این طریق به‌دست می‌آید، می‌توان با روش‌های آماری دقیق‌تر، این فرضیه را تأیید یا رد کرد.

در تحلیل اکتشافی، متغیرهای معلوم باید با دقت زیادی انتخاب شوند، تا حتی‌الامکان درباره متغیرهای نامعلومی که استخراج می‌شوند، اطلاعات بیشتری فراهم آید. همچنین، تحلیل اکتشافی نیازمند نمونه‌هایی با حجم بسیار زیاد می‌باشد.

حال برای مدل کردن ϕ بر روی متغیرهای تبیینی، ابتدا نمودارهای جعبه‌ای را به همراه نمودار چگالی آن‌ها برای هر متغیر، با متغیر پاسخ رسم می‌کنیم. نمودار جعبه‌ای یکی از بهترین نمودارها برای فهم واضح‌تری از مجموعه داده‌ها می‌باشد. این نمودار به کمک معیارهای مرکزی و پراکندگی، موقعیت آن‌ها را به شکلی بسیار مفید و گویا ارائه می‌دهد. همچنین این نمودار، برای توصیف تغییرات داده‌ها نیز به کار می‌رود. با توجه به آن‌چه ذکر شد، باید متغیرهای تبیینی را بیابیم که نشان‌دهنده بیشترین تغییرات در مدل باشند.



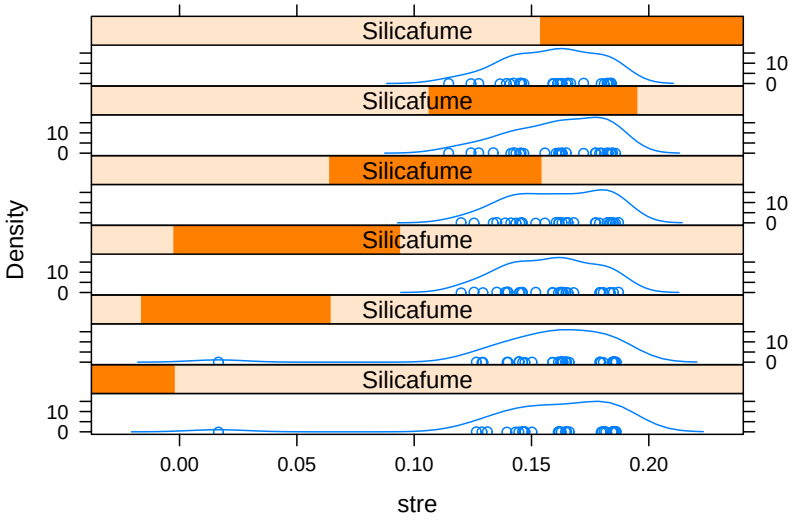
شکل ۸.۴: نمودار جعبه‌ای و چگالی پاسخ مربوط به متغیر آب



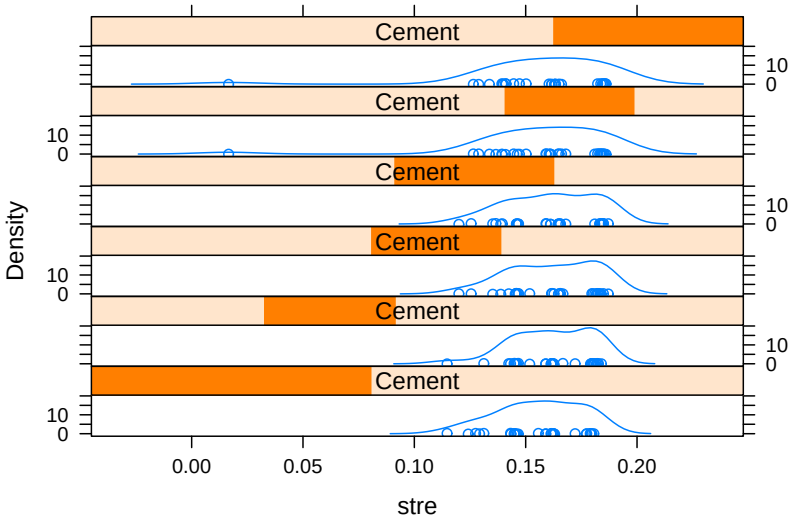
شکل ۹.۴: نمودار جعبه‌ای و چگالی پاسخ مربوط به متغیر ماسه

نمودارهای ۸.۴ تا ۱۱.۴ در همین جهت رسم شده‌اند. با توجه به آن‌ها و آن‌چه گفته شد، طول مستطیل‌های نمودار مربوط به متغیر آب نسبت به سایر متغیرها طول کوچکتری دارد. در نتیجه، این متغیر نسبت به دیگر متغیرها، تغییرات کمتری را نشان داده و مدل نهایی را برای پارامتر دقت متغیر با ماسه، میکروسیلیس و سیمان انجام می‌دهیم. مدل مستطیلی بتا را با پارامتر دقت متغیر به‌صورت

$$
\begin{aligned}
&stre_i \sim BRr(\gamma_i, \phi, \alpha),\\
&logit(\gamma_i) = \beta_۰ + \beta_۱ water_i + \beta_۲ sand_i + \beta_۳ silica fume_i + \beta_۴ cement_i,\\
&log(\phi_i) = \delta_۰ + \delta_۱ sand_i + \delta_۲ silica fume_i + \delta_۳ cement_i \quad i = ۱, ۲, \cdots, n
\end{aligned} \tag{۱.۴}
$$



شکل ۱۰.۴: نمودار جعبه‌ای و چگالی پاسخ مربوط به متغیر میکروسیلیس



شکل ۱۱.۴: نمودار جعبه‌ای و چگالی پاسخ مربوط به متغیر سیمان

در نظر می‌گیریم. مدل رگرسیونی بتا بر اساس آن، بدون در نظر گرفتن پارامتر آمیختگی به‌دست می‌آید. برای آزمون نتیجه حاصل از تحلیل اکتشافی، مدل‌های بتا و مستطیلی بتا را برای ترکیب‌های سه‌تایی و همچنین چهارتایی از عوامل مؤثر بر پارامتر دقت مدل مقاومت بتن برازش دادیم. نتایج برازش این مدل‌ها در جدول ۱۰.۴ گزارش شده‌اند. جدول ۱۰.۴ مدل‌های بتا و مستطیلی بتا را با ترکیب‌های متفاوت برای پارامتر دقت بر اساس معیار DIC مقایسه می‌کند. این جدول نیز بر آن‌چه از تحلیل اکتشافی به‌دست آمد، تأکید می‌کند. یعنی مدلی بدون متغیر آب را برای ϕ توصیه می‌کند. علاوه بر این، برای درک این‌که کدام‌یک از عوامل تأثیر بیشتر و کدام‌یک تأثیر کمتری در پراکندگی مدل نهایی دارند، مدل‌های مورد بحث را با تعداد متغیرهای تبیینی یکی و دوتایی برای پارامتر دقت برازش دادیم. نتیجه

جدول ۱۰.۴: مقایسه مدل‌های مرکب از متغیرهای پراکندگی آب، ماسه، میکروسیلیس و سیمان در حضور تمامی متغیرهای تبیینی بر اساس معیار DIC

DIC	$log(\phi_i)$	بتا	مستطیلی بتا
۱	$\delta_0 + \delta_1 water + \delta_2 sand + \delta_3 silica fume$	−۴۹۳/۸۶۴۲	−۵۱۱/۱۷۰۴
۲	$\delta_0 + \delta_1 water + \delta_2 sand + \delta_4 cement$	−۴۹۲/۹۲۶۲	−۵۱۸/۶۷۷۵
۳	$\delta_0 + \delta_1 water + \delta_3 silica fume + \delta_4 cement$	−۴۹۳/۰۴۵۳	−۵۱۳/۶۶۸۱
۴	$\delta_0 + \delta_2 sand + \delta_3 silica fume + \delta_4 cement$	−۴۹۶/۲۵۹۱	−۵۲۲/۶۹۰۰
۵	$\delta_0 + \delta_1 water + \delta_2 sand + \delta_3 silica fume + \delta_4 cement$	−۴۹۳/۴۴۷۵	−۵۱۳/۷۹۲۶

جدول ۱۱.۴: مقایسه مدل‌های مرکب از متغیرهای پراکندگی آب، ماسه، میکروسیلیس و سیمان در حضور تمامی متغیرهای تبیینی بر اساس معیار DIC

DIC	$log(\phi_i)$	بتا	مستطیلی بتا
۱	$\delta_0 + \delta_1 water$	−۴۷۵/۵۹۸۲	−۵۰۸/۳۸۴۹
۲	$\delta_0 + \delta_2 sand$	−۴۷۹/۴۰۰۹	−۵۲۵/۸۳۷۷
۳	$\delta_0 + \delta_3 silica fume$	−۴۷۸/۳۲۷۸	−۵۲۴/۷۷۳۳
۴	$\delta_0 + \delta_4 cement$	−۴۸۲/۵۰۳۴	−۵۲۹/۹۹۸۰
۵	$\delta_0 + \delta_1 water + \delta_2 sand$	−۴۸۵/۸۸۷۰	−۵۱۰/۴۰۴۷
۶	$\delta_0 + \delta_1 water + \delta_3 silica fume$	−۴۹۶/۴۸۱۵	−۵۱۷/۰۰۴۵
۷	$\delta_0 + \delta_1 water + \delta_4 cement$	−۴۸۵/۶۷۵۰	−۵۱۷/۳۵۷۹
۸	$\delta_0 + \delta_2 sand + \delta_3 silica fume$	−۴۷۵/۳۸۹۹	−۵۲۴/۱۳۰۳
۹	$\delta_0 + \delta_2 sand + \delta_4 cement$	−۴۹۶/۶۴۸۶	−۵۲۵/۲۲۰۲
۱۰	$\delta_0 + \delta_3 silica fume + \delta_4 cement$	−۴۹۴/۰۰۵۴	−۵۲۲/۰۵۶۰

این بررسی در جدول ۱۱.۴ آمده است. همان‌طور که ملاحظه می‌کنید، با مقایسه مدل‌های ۱ تا ۴ عامل سیمان در مدل ۴ بیشترین پراکندگی را برای مدل‌های بتا و مستطیلی بتا مدل می‌کند. پس از آن، ماسه و سپس میکروسیلیس و پس از آن، آب به‌ترتیب پراکندگی از بیشتر به کمتر را تبیین می‌کنند. با نظر به مدل‌های ۵ تا ۱۰ مشاهده می‌کنید، مدل ۹ با ترکیبی شامل سیمان و ماسه کمترین مقدار DIC را به خود اختصاص می‌دهد. مدل‌های حاصل از این ترکیب برازش بهتری خواهند داشت.

حال، پس از بررسی‌های انجام شده مدل (۱.۴) را برازش داده و مدل رگرسیونی بتا را با توجه به آن، با حذف پارامتر آمیختگی برازش می‌دهیم. نتایج این برازش در جداول ۱۲.۴ و ۱۳.۴ گزارش شده‌اند. جدول ۱۲.۴ گویای میانگین پسین و فاصله اعتبار ۹۵٪ را برای پارامترهای مدل‌های بتا و مستطیلی بتا با پارامتر دقت متغیر بر اساس داده‌های مقاومت بتن گزارش می‌دهد. در مقایسه این جدول با جدول ۸.۴ در می‌یابیم که با در نظر گرفتن ϕ متغیر، بازه اعتبار متغیرها افزایش یافته و پراکندگی پارامترها

جدول ۱۲.۴: میانگین پسین و فاصله اعتبار ۹۵٪ برای مدل‌های رگرسیون بتا و مستطیلی بتا با ϕ متغیر طبق داده‌های مقاومت بتن

	بتا			مستطیلی بتا		
پارامتر	میانگین	٪۲/۵	٪۹۷/۵	میانگین	٪۲/۵	٪۹۷/۵
β_0	−۱/۶۵۸۷	−۱/۶۹۴۰	−۱/۶۲۱۶	−۱/۶۰۳۵	−۱/۶۵۳۷	−۱/۵۲۴۵
β_1	−۱/۷۵۱۸	−۱/۳۳۸۰	−۰/۱۶۲۹	−۱/۰۴۵۲	−۱/۴۰۳۱	−۰/۶۹۱۲
β_2	۳/۱۷۵۰	۲/۱۸۸۰	۴/۱۹۶۰	۲/۳۵۷۹	۰/۶۸۱۴	۳/۷۱۵۰
β_3	۵/۶۶۵۱	۲/۹۳۹۵	۸/۳۸۹۳	۴/۳۲۴۹	۱/۷۹۷۳	۶/۷۳۷۸
β_4	۴/۰۸۳۲	۲/۳۱۳۴	۵/۸۳۲۶	۳/۷۸۳۱	۱/۹۱۱۳	۵/۴۵۳۹
δ_0	−۵/۴۹۱۲	−۵/۷۴۹۸	−۵/۲۰۵۱	−۶/۲۹۶۱	−۶/۵۶۹۷	−۶/۰۰۵۳
δ_1	۱۰/۳۱۸۶	۱/۹۰۹۳	۱۷/۸۷۸۳	−۱۰/۵۶۷۲	−۲۱/۷۳۴۲	۰/۳۷۹۱
δ_2	−۱۰/۳۱۸۶	−۲۶/۰۱۲۴	۵/۱۰۵۴	−۴/۱۸۸۷	−۱۸/۸۰۵۴	۱۱/۲۹۳۲
δ_3	۱۸/۶۲۲۳	۱۰/۸۴۵۱	۲۶/۳۳۷۶	−۲/۳۰۷۶	−۱۲/۶۶۹۰	۷/۷۸۹۳
α				۰/۰۵۳۴	۰/۰۰۶۹	۰/۱۴۱۶

جدول ۱۳.۴: مقایسه مدل‌های بتا و مستطیلی بتا با ϕ متغیر و زمان اجرا (بر حسب ثانیه) با معیارهای DIC ، $EAIC$ و $EBIC$ برای داده‌های مقاومت بتن

مدل	زمان اجرا	DIC	$EAIC$	$EBIC$
بتا	۱۳۶۵/۵۸	−۴۹۶/۱۵۹۱	−۴۸۹/۰۵۱۹	−۴۶۴/۹۱۲۷
مستطیلی بتا	۳۵۴۱/۴۷	−۵۲۲/۶۹۰۰	−۵۱۲/۷۶۵۴	−۴۹۹/۶۵۷۸

زیاد شده‌اند.

جدول ۱۳.۴ مدل‌های رگرسیونی بتا و مستطیلی بتا را براساس معیارهای $EAIC$، $EBIC$ و DIC و زمان اجرا (ثانیه) با هم مقایسه می‌کند. در این جدول مقادیری که برای مدل رگرسیونی مستطیلی بتا بر اساس معیارهای مقایسه مدل به‌دست می‌آیند، کمتر از مقادیری هستند که برای مدل رگرسیون بتا به‌دست می‌آیند. پس، مدل رگرسیون مستطیلی بتا نسبت به بتا برازش بهتری دارد. با مشاهده جداول ۷.۴ و ۱۳.۴ در می‌یابیم که بر اساس معیارهای مقایسه مدل با پارامتر دقت ثابت و متغیر مدل رگرسیون مستطیلی بتا نسبت به بتا برازش بهتری دارد. همچنین، توجه کنید که با پارامتر دقت ثابت، مدل مستطیلی بتا برازش بهتری نسبت به مدل مستطیلی بتا با پارامتر دقت متغیر دارد.

۳.۴ نتیجه‌گیری

در این پایان‌نامه، یک مدل رگرسیون تنومند جدید نسبت به نقاط پرت برای پاسخ‌های محدود به فاصله $(۰, ۱)$ ارائه شد. در این مدل، توزیع متغیر پاسخ، آمیخته‌ای از دو توزیع بتا و یکنواخت است. در واقع، رگرسیون مستطیلی بتا از مدل رگرسیون بتا نتیجه می‌شود. در این پایان‌نامه، دیدگاه استنباطی مورد نظر، بیزی انتخاب شد. اما از آن‌جا که توزیع پسین مدل رگرسیون مستطیلی بتا به‌صورت بسته وجود ندارد، از الگوریتم‌های نمونه‌گیری $MCMC$ برای تقریب آن استفاده کردیم. پیچیدگی زیاد توزیع‌های شرطی کامل پارامترها باعث شد تا از نمونه‌گیر ترکیبی متروپلیس–هستینگز و گیبز استفاده کنیم. همه نمونه‌های لازم برای استخراج استنباط‌های بیزی در نرم‌افزار $JAGS$ تولید شدند. ویژگی بارز این نرم‌افزار آن است که توزیع‌های شرطی کامل برای اجرای نمونه‌گیر گیبز را به‌طور خودکار محاسبه و از آن تولید نمونه می‌کند.

عملکرد این مدل در کنار مدل رگرسیون بتا، با دو مطالعه شبیه‌سازی مورد ارزیابی قرار گرفت و کاربرد آن در دو مثال واقعی نشان داده شد. با توجه به نتایج به‌دست آمده از شبیه‌سازی‌ها و پیاده‌سازی مدل روی مثال‌های واقعی، به‌طور خلاصه می‌توان نتایج زیر را عنوان کرد:

- مزیت اساسی مدل مستطیلی بتا، انعطاف بالای آن در برابر داده‌هایی است که مشاهدات پرت دارند.
- مدل رگرسیون مستطیلی بتا، برای مدل‌بندی داده‌های نسبت و درصد تنومند است. در شبیه‌سازی‌ها، این تنومندی در قسمت‌های مختلف مدل‌بندی نشان داده شد. یعنی تنومندی در برآورد ضرایب، کل توزیع پسین و معیارهای انتخاب مدل.
- برخلاف مدل رگرسیون مستطیلی بتا، مدل رگرسیون بتا به برآورد ضرایب رگرسیون حساس است. این حساسیت با استفاده از معیار فاصله کولبک–لیبلر توزیع پسین همه پارامترها مشخص می‌گردد.
- تقریباً در تمام شبیه‌سازی‌ها، مبتنی بر چهار الگوی تولید نقاط پرت متفاوت و معیارهای انتخاب مدل، مدل برتر مدل رگرسیون مستطیلی بتا است.

۴.۴ پیشنهادات برای آینده تحقیق

با توجه به مباحث نظری و مطالعات شبیه‌سازی، می‌توان برای ادامه تحقیق به بندهای زیر اشاره نمود:

- می‌توان به دنبال معرفی برخی ابزار تشخیصی برای رگرسیون مستطیلی بتا بود. برای مثال، به‌کارگیری معیار کولبک-لیبلر، اندازه‌ای تشخیصی برای مشاهدات است که توسط پنگ و دی (۱۹۹۵) و همچنین چو و همکاران (۲۰۰۹) به کار گرفته شده است. اثرپذیری موضعی داده‌ها یا برآوردها و تحلیل باقیمانده‌ها، توسط فراری و همکاران در سال ۲۰۱۱ برای مدل رگرسیون بتا مورد بررسی قرار گرفت که می‌توان آن را برای مدل مستطیلی بتا و همچنین از دیدگاه بیزی فرمول‌بندی نمود.

- در مواردی که حجم داده‌ها زیاد باشد، اجرای استنباط‌های بیزی در نرم‌افزارهای $WinBUGS$ و $JAGS$ کند بوده و کارا نیست. از طرفی، انجام استنباط‌های بیزی، با توجه به پیچیدگی مدل مستطیلی بتا، توسط این نرم‌افزارها ساده است. این دو ویژگی باعث می‌شود که به دنبال الگوریتم‌های نمونه‌گیری جدیدی باشیم که سرعت تولید نمونه در حالتی که حجم داده‌ها زیاد باشند را بالا ببرد (گامرمن و لوپس، ۲۰۰۶). برای مثال، الگوریتم متروپلیس-هستینگز در $C++$ با R به سرعت اجرا می‌شود (ابانک و کاپرسانین، ۲۰۱۲). بنابراین بهبود عملکرد مدل رگرسیونی مستطیلی بتا زمانی که داده‌ها زیاد باشند، می‌تواند موضوع تحقیق مورد توجهی باشد. این بهبودی بر اساس تقریب خوب و سریع توزیع پسین مدل با الگوریتم‌های نمونه‌گیری یا عددی است.

پیوست آ

دستورهای نرم‌افزار $JAGS$

نمونه‌ای از کدهای R برای شبیه‌سازی‌های فصل سوم و همچنین مثال‌های فصل چهارم در این پیوست گزارش شده‌اند.

شبیه‌سازی اول فصل سوم

بسته‌های فراخوانی شده در شبیه‌سازی اول

```
library(rjags)
library(coda)
library(lattice)
```

در این بخش، ابتدا داده‌های آلوده‌شده را تولید می‌کنیم.

```
logit.inv = function(x){
exp(x)/(1+exp(x))
}

data.gen = function(n,phi,mu=0.2,r){
y = rbeta(n,shape1=mu*phi, shape2=(1-mu)*phi)
index = sample(1:n, size = r*n/100)
q = quantile(y, prob = 0.999)
y[index] = runif(length(index),q,1)
return(y)
}
```

سپس ۱۰۰ تکرار زنجیر مارکوف مونت کارلو را شبیه‌سازی نموده و مقادیر MSE، DIC و اریبی را برای آن‌ها به‌دست می‌آوریم.

```
simulate <-function(rep,n,phi,mu,r){
param.values1 = matrix(NA,nrow=rep, ncol=3) # parameter of beta rect reg model
param.values2 = matrix(NA,nrow=rep, ncol=2) # parameter of beta reg model
DICrectangular <- DICclassic <- rep(0,rep)
for (i in 1:rep){
y = data.gen(n,phi,mu,r)
###
cat("
model{
## Likelihood
for (i in 1:n){
y[i]~dbeta(a[i,u[i]],b[i,u[i]])
v[i]~dbern(theta[i])
u[i]<-v[i]+1
a[i,1] <- mu[i]*phi
b[i,1] <- (1-mu[i])*phi
a[i,2]<-1
b[i,2]<-1
logit(gamma[i])<- beta
theta[i]<-alpha.star*(1-abs((2*gamma[i])-1))
mu[i]<-(gamma[i]-0.5*theta[i])/(1-theta[i])
}
#### Priors
beta~dnorm(0,0.01)
phi ~dgamma(0.1,0.1)
alpha.star~dunif(0,1)
}
", fill=TRUE, file="model1.jag")
cat("
model{
## Likelihood
for (i in 1:n){
```

```
y[i]~dbeta(a[i],b[i])
a[i] <- gamma[i]*phi
b[i] <- (1-gamma[i])*phi
logit(gamma[i])<- beta
}
##priors
beta~dnorm(0,0.01)
phi ~dgamma(0.1,0.1)
}", fill=TRUE, file="model2.jag")

jags_d <- list(y=y, n=n)

d.ini1<-list(list(beta=0,alpha.star=0.02, phi=10)
              ,list(beta=1,alpha.star=0.01, phi=9)
              ,list(beta=-1,alpha.star=0.01, phi=10))
d.ini2<-list(list(beta=0, phi=10)
              ,list(beta=1, phi=9)
              ,list(beta=-1, phi=10))

mod1 <- jags.model("model1.jag",
                   data=jags_d, n.chains=3,inits=d.ini1 ,n.adapt=1000)
mod2 <- jags.model("model2.jag",
                   data=jags_d, n.chains=3,inits=d.ini2 ,n.adapt=1000)

out1 <- coda.samples(mod1, variable.names=c("beta", "alpha.star",
    "phi"), n.iter=1000,thin=10)
out2 <- coda.samples(mod2, variable.names=c("beta", "phi"),
                     n.iter=1000,thin=10)

dic1 <- dic.samples(mod1, n.iter=1000, thin = 10, type="pD" )
DICrectangular[i] = sum(dic1$deviance)+sum(dic1$penalty)

dic2 <- dic.samples(mod2, n.iter=1000, thin=10, type="pD")
DICclassic[i] = sum(dic2$deviance)+sum(dic2$penalty)
```

```
par1 <- rbind(out1[[1]],out1[[2]],out1[[3]])
par2 <- rbind(out2[[1]],out2[[2]],out2[[3]])
param.values1[i,] = apply(par1,2,mean)
param.values2[i,] = apply(par2,2,mean)
cat("iter=", "->" , i ,"\n")
}
out=return(list("beta.rectangular"=param.values1,
                "beta.classic"=param.values2,
                "DICrectangular"=DICrectangular,
                "DICclassic"=DICclassic))
 }


n = {50,100,200}
phi ={10,30}
mu = 0.2
r = {\%2,\%5,\%8}

object = simulate(rep=100, n=n, phi=phi, mu=mu, r=r)

### for beta rectangular model
beta.param1 = logit.inv(object$beta.rectangular[,2])
beta.est1 = mean(beta.param1)
estimated.bias1 = beta.est1 - mu

### for beta model
beta.param2 = logit.inv(object$beta.classic[,1])
beta.est2 = mean(beta.param2)
estimated.bias2 = beta.est2 - mu

### for beta rectangular model
MSE1 = mean((beta.param1 - mu)^2)
### for beta model
```

```
MSE2 = mean((beta.param2 - mu)^2)

sum(as.numeric(object$DICrectangular<object$DICclassic)/100))
```

شبیه‌سازی دوم فصل سوم

پس از تولید داده‌های آلوده‌نشده، مدل‌های بتا و مستطیلی بتا را با این داده‌ها شبیه‌سازی می‌کنیم.

```
n<-200
beta0.star <- 0.5
beta1 <- 1
phi <- 30
x <- runif(n,-3,3)
logit.link <- function(x){
exp(x)/(1+exp(x))
}
eta <- beta0.star+(beta1*(x-mean(x)))
gamma <- logit.link(eta)
mu <- gamma
y = rbeta(n ,mu*phi,(1-mu)*phi)

cat("
model{
## Likelihood
for (i in 1:length(y)){
logit(gamma[i])<- beta0.star+(beta1*(x[i]-mean(x[])))
mu[i] <- gamma[i]
y[i]~ dbeta(a[i],b[i])
a[i] <- mu[i]*phi
b[i] <- (1-mu[i])*phi
}
#### Priors
beta1~dnorm(1,0.9)
beta0.star ~ dnorm(0.5,0.9)
beta0 <- beta0.star-beta1*mean(x[])
```

```
phi ~ dgamma(0.01,0.01)
}
", fill=TRUE, file="beta.jag")

cat("
model{
## Likelihood
for (i in 1:length(y)){
y[i]~ dbeta(a[i,u[i]],b[i,u[i]])
logit(gamma[i])<- beta0.star+(beta1*(x[i]-mean(x[])))
theta[i]<- alpha.star*(1-abs((2*gamma[i])-1))
mu[i]<-(gamma[i]-0.5*theta[i])/(1-theta[i])
v[i]~ dbern(theta[i])
u[i]<- v[i]+1
a[i,1]<- (mu[i]*phi)
b[i,1]<- ((1-mu[i])*phi)
a[i,2]<- 1
b[i,2]<- 1
}
#### Priors
beta1 ~ dnorm(1,0.9)
beta0.star ~ dnorm(0.5,0.9)
beta0 <- beta0.star-beta1*mean(x[])
phi ~ dgamma(0.01,0.01)
alpha.star ~ dunif(0,0.5)
}
", fill=TRUE, file="betarec.jag")

jags_d <- list(x=x, y=y)

d.ini1<-list(list(beta1=1,beta0.star=0.45,phi=30)
,list(beta1=0.9,beta0.star=0.5,phi=29)
,list(beta1=1,beta0.star=0.5,phi=30))
```

```
d.ini2<-list(list(beta1=1,beta0.star=0.48,alpha.star=0.02,phi=27)
,list(beta1=0.9,beta0.star=0.5,alpha.star=0.03,phi=25)
,list(beta1=1,beta0.star=0.5,alpha.star=0.02,phi=28))

mod1 <- jags.model("beta.jag",
data=jags_d, n.chains=3,inits=d.ini1 ,n.adapt=10000)

mod2 <- jags.model("betarec.jag",
data=jags_d, n.chains=3,inits=d.ini2 ,n.adapt=10000)

out1 <- coda.samples(mod1, variable.names=c("beta1", "beta0.star","phi"),
n.iter=10000, thin=10
)
out1.beta <- out1[[1]]
out2.beta <- out1[[2]]
out3.beta <- out1[[3]]
outn.beta <- rbind(out1.beta,out2.beta,out3.beta)
coef.beta <- apply(outn.beta,2,mean)

summary(out1)

out2 <- coda.samples(mod2, variable.names=c("beta1", "beta0.star",
"alpha.star","phi"),
n.iter=10000, thin=10
)
out1.betarec <- out2[[1]]
out2.betarec <- out2[[2]]
out3.betarec <- out2[[3]]
outn.betarec <- rbind(out1.betarec,out2.betarec,out3.betarec)
coef.betarec <- apply(outn.betarec,2,mean)

summary(out2)
###DIC beta
dic.beta <- dic.samples(mod1, n.iter=10000, thin = 10, type="pD" )
```

```
 DICbeta = sum(dic.beta$deviance)+ sum(dic.beta$penalty)
###EBIC beta
p <- 3
EBIC.beta <- sum(dic.beta$deviance) + (p*log(n))
###EAIC beta
p <- 3
EAIC.beta <- sum(dic.beta$deviance) + (2*p)
###DIC betarec
dic.betarec <- dic.samples(mod2, n.iter=10000, thin = 10, type="pD" )
DICbetarec = sum(dic.betarec$deviance)+ sum(dic.betarec$penalty)
###EBIC betarec
p <- 4
EBIC.betarec <- sum(dic.betarec$deviance) + (p*log(n))
###EAIC betarec
p <- 4
EAIC.betarec <- sum(dic.betarec$deviance) + (2*p)
```

چهار الگوی اختلال برای ایجاد داده‌های آلوده‌شده

```
###nmodar1 shabihsazi2
library(rjags)
library(coda)
library(lattice)
rm(list=ls())
set.seed(12354)
n<-200
beta0.star<-0.5
beta1<-1
phi<-30
x<-runif(n,-3,3)
logit.link<-function(x){
exp(x)/(1+exp(x))
}
eta<-beta0.star+beta1*x
gamma<-logit.link(eta)
mu<- gamma
```

```
y=rbeta(n,mu*phi,(1-mu)*phi)
###1
par(mfrow=c(2,2))
plot(y~x,main="(1)")
t = which(x>2)
ind = sample(t,size=4)
ind
delta= 0.8
ynew = y[ind]-delta
arrows(c(x[ind[1]],x[ind[2]],x[ind[3]],x[ind[4]]),c(y[ind[1]],y[ind[2]]
,y[ind[3]],y[ind[4]]),c(x[ind[1]],x[ind[2]],x[ind[3]],x[ind[4]]),
c((y[ind[1]]-0.8),(y[ind[2]]-0.8),(y[ind[3]]-0.8),(y[ind[4]]-0.8))
,length=0.1)points(c(x[ind[1]],x[ind[2]],x[ind[3]],x[ind[4]],x[ind[1]]
,x[ind[2]],x[ind[3]],x[ind[4]]),c(y[ind[1]],y[ind[2]],y[ind[3]],y[ind[4]]
,(y[ind[1]]-0.8),
(y[ind[2]]-0.8),(y[ind[3]]-0.8),(y[ind[4]]-0.8)),pch=8)
###2
plot(y~x,main="(2)")
t = which(x< -2)
ind = sample(t,size=4)
ind
delta= 0.8
ynew = y[ind]+delta
arrows(c(x[ind[1]],x[ind[2]],x[ind[3]],x[ind[4]]),c(y[ind[1]],y[ind[2]]
,y[ind[3]],y[ind[4]]),c(x[ind[1]],x[ind[2]],x[ind[3]],x[ind[4]]),
c((y[ind[1]]+0.8),(y[ind[2]]+0.8),(y[ind[3]]+0.8),(y[ind[4]]+0.8))
,length=0.1)points(c(x[ind[1]],x[ind[2]],x[ind[3]],x[ind[4]],x[ind[1]]
,x[ind[2]],x[ind[3]],x[ind[4]]),c(y[ind[1]],y[ind[2]],y[ind[3]],y[ind[4]]
,(y[ind[1]]+0.8),(y[ind[2]]+0.8),(y[ind[3]]+0.8),(y[ind[4]]+0.8)),pch=8)
###3
plot(y~x,main="(3)")
t1 = which(x< -2)
ind1 = sample(t1,size=2)
ind1
```

```
delta= 0.8
ynew = y[ind1]+delta
arrows(c(x[ind1[1]],x[ind1[2]]),c(y[ind1[1]],y[ind1[2]])
,c(x[ind1[1]],x[ind1[2]]),
c((y[ind1[1]]+0.8),(y[ind1[2]]+0.8)),length=0.1)
points(c(x[ind1[1]],x[ind1[2]],x[ind1[1]],x[ind1[2]])
,c(y[ind1[1]],y[ind1[2]],(y[ind1[1]]+0.8),
(y[ind1[2]]+0.8)),pch=8)
t2 = which(x>2)
ind2 = sample(t2,size=2)
ind2
delta= 0.8
ynew = y[ind2]-delta
arrows(c(x[ind2[1]],x[ind2[2]]),c(y[ind2[1]],y[ind2[2]])
,c(x[ind2[1]],x[ind2[2]]),
c((y[ind2[1]]-0.8),(y[ind2[2]]-0.8)),length=0.1)
points(c(x[ind2[1]],x[ind2[2]],x[ind2[1]],x[ind2[2]])
,c(y[ind2[1]],y[ind2[2]],(y[ind2[1]]-0.8),
(y[ind2[2]]-0.8)),pch=8)
###4
plot(y~x,main="(4)")
t = which(abs(x)<0.5)
ind = sample(t,size=4)
ind
delta= 0.5
ynew = y[ind]-delta
arrows(c(x[ind[1]],x[ind[2]],x[ind[3]],x[ind[4]]),c(y[ind[1]]
,y[ind[2]],y[ind[3]],y[ind[4]]),c(x[ind[1]],x[ind[2]],x[ind[3]]
,x[ind[4]]),c((y[ind[1]]-0.6),(y[ind[2]]-0.6),(y[ind[3]]-0.6)
,(y[ind[4]]-0.6)),length=0.1)points(c(x[ind[1]],x[ind[2]]
,x[ind[3]],x[ind[4]],x[ind[1]],x[ind[2]],x[ind[3]],x[ind[4]])
,c(y[ind[1]],y[ind[2]],y[ind[3]],y[ind[4]],(y[ind[1]]-0.6),
(y[ind[2]]-0.6),(y[ind[3]]-0.6),(y[ind[4]]-0.6)),pch=8)
```

کاربرد اول فصل چهارم

بسته‌های فراخوانی شده در کاربرد اول

```
library(rjags)
library(coda)
library(lattice)
library(sn)
```

برازش مدل‌های بتا و مستطیلی بتا بر داده‌های ais بسته sn و ϕ ثابت

```
n<-37
Bfat <- c(16.86,10.81,20.10,21.47,18.48,25.26,18.72,18.77,17.64,23.01,
24.88,19.20,24.69,24.69,16.20,23.11,19.83,16.58,23.01,16.58,12.61,
17.22,17.07,12.78,12.39,9.40,8.51,8.84,8.87,5.63,6.99,9.89,10.12,5.80,
5.63,8.77,6.26)/100

lbm <- c(41.93,48.76,53.54,53.46,53.65,53.77,54.77,54.41,54.78,54.77,
54.78,55.35,55.73,56.01,57.54,58.00,59.59,60.05,61.46,61.70,65.00,
66.00,67.00,68.00,68.86,69.00,71.00,72.00,74.00,75.00,76.00,77.00,
78.00,79.00,80.00,82.00,86.00)

y <- Bfat
x <- lbm

cat(" model{
## Likelihood
for (i in 1:length(y)){
y[i]~ dbeta(a[i],b[i])
a[i] <- gamma[i]*phi
b[i] <- (1-gamma[i])*phi
   logit(gamma[i])<- beta0.star+(beta1*(x[i]-mean(x[])))
}
 ## Priors
beta1 ~ dnorm(0,0.1)
beta0.star ~ dnorm(0,0.05)
beta0 <- beta0.star-beta1*mean(x[])
phi ~ dgamma(0.01,0.01)
```

```
#alpha.star ~ dnorm(0,1)
}
", fill=TRUE, file="model-beta.jag")

cat(" model{
## Likelihood
for (i in 1:length(y)){
    y[i]~dbeta(a[i,u[i]],b[i,u[i]])
    logit(gamma[i])<- beta0.star+(beta1*(x[i]-mean(x[])))
        theta[i]<- alpha.star*(1-abs((2*gamma[i])-1))
        mu[i]<-(gamma[i]-0.5*theta[i])/(1-theta[i])
    v[i]~dbern(theta[i])
        u[i]<-v[i]+1
        a[i,1]<-mu[i]*phi
        b[i,1]<-(1-mu[i])*phi
        a[i,2]<-1
        b[i,2]<-1
        }
 ## Priors
beta1~ dnorm(0,0.1)
beta0.star~dnorm(0,0.05)
beta0<-beta0.star-beta1*mean(x[])
phi~ dgamma(0.01,0.01)
alpha.star~dunif(0,1)
}
", fill=TRUE, file="model-betarec.jag")

jags_d <- list(x=lbm, y=Bfat )

d.ini1<-list(list(beta1=-0.02,beta0.star=0.089,phi=86)
,list(beta1=-0.03,beta0.star=0.09,phi=87),
list(beta1=-0.02,beta0.star=0.09,phi=87)
)
```

```
d.ini2<-list(list(beta1=-0.03,beta0.star=0.8,phi=203,alpha.star=0.09)
,list(beta1=-0.04,beta0.star=0.9,phi=204,alpha.star=0.1),
list(beta1=-0.03,beta0.star=0.9,phi=203,alpha.star=0.09)
)

mod.beta<-jags.model("model-beta.jag",
data=jags_d,
n.chains=3,
inits=d.ini1,
n.adapt=50000
)

mod.betarec<-jags.model("model-betarec.jag",
data=jags_d,
n.chains=3,
inits=d.ini2,
n.adapt=50000
)

out.beta<- coda.samples( mod.beta,
variable.names = c("beta1","beta0.star","phi"),
n.iter =50000,
thin = 10 )

summary(out.beta)

out.betarec<- coda.samples( mod.betarec,
 variable.names = c("beta1","beta0.star","phi","alpha.star"),
n.iter =50000,
thin = 10 )

summary(out.betarec)

out1.beta <- out.beta[[1]]
```

```
out2.beta <- out.beta[[2]]
out3.beta <- out.beta[[3]]
outn.beta <- rbind(out1.beta,out2.beta,out3.beta)
coef.beta <- apply(outn.beta,2,mean)

out1.betarec <- out.betarec[[1]]
out2.betarec <- out.betarec[[2]]
out3.betarec <- out.betarec[[3]]
outn.betarec <- rbind(out1.betarec,out2.betarec,out3.betarec)
coef.betarec <- apply(outn.betarec,2,mean)
###DIC beta
dic.beta <- dic.samples(mod.beta, n.iter=50000, thin = 10, type="pD" )
DIC.beta = sum(dic.beta$deviance)+sum(dic.beta$penalty)
###EBIC beta
p <- 3
EBIC.beta <- sum(dic.beta$deviance)+ (p*log(n))
###EAIC beta
p <- 3
EAIC.beta <- sum(dic.beta$deviance) + (2*p)
###DIC betarec
dic.betarec <- dic.samples(mod.betarec, n.iter=50000, thin = 10, type="pD" )
DICbeta.betarec = sum(dic.betarec$deviance)+sum(dic.betarec$penalty)
###EBIC betarec
p <- 4
EBIC.betarec <- sum(dic.betarec$deviance)+ (p*log(n))
###EAIC betarec
p <- 4
EAIC.betarec <- sum(dic.betarec$deviance) + (2*p)
###nmodar
logit.link<-function(x){
  exp(x)/(1+exp(x))
}
beta0.beta<- coef.beta[1]
beta1.beta<- coef.beta[2]
```

```
phi.beta<- coef.beta[3]

alpha<- coef.betarec[1]
beta0.betarec<- coef.betarec[2]
beta1.betarec<- coef.betarec[3]
phi.betarec<- coef.betarec[4]

f1 = function(x = lbm){
eta<-beta0.beta+beta1.beta*(x-mean(x))
mu = logit.link(eta)
   return(mu)
}
f2 = function(x = lbm){
eta<- beta0.betarec + beta1.betarec*(x-mean(x))
gamma <- logit.link(eta)
   theta <- alpha*(1-abs((2*gamma)-1))
mu <- (gamma-(0.5*theta))/(1-theta)
    mun <- (theta/2)+((1-theta)*mu)
return(mun)
}
plot(Bfat~lbm,xlim=c(40,80),ylim=c(0.08,0.25))
curve(f1,min(lbm),max(lbm),add=T,lty=2)
curve(f2,min(lbm),max(lbm),add=T,lty=1)
```

برازش مدل‌های بتا و مستطیلی بتا بر داده‌های ais بسته sn با ϕ متغیر

```
###write model beta
cat(" model{
## Likelihood
for (i in 1:length(y)){
logit(gamma[i])<- beta0.star+(beta1*(x[i]-mean(x[])))
log(phi[i]) <- -delta0.star-delta1 * (x[i]-mean(x[]))
y[i]~ dbeta(a[i],b[i])
a[i] <- gamma[i]*phi[i]
b[i] <- (1-gamma[i])*phi[i]
}
```

```
 #### Priors
beta1 ~ dnorm(100,0.01)
beta0.star ~ dnorm(100,0.01)
beta0 <- beta0.star-beta1 * mean(x[])
delta1 ~ dnorm(0.01,0.01)
delta0.star ~ dnorm(0.01,0.01)
delta0 <- delta0.star-delta1 * mean(x[])
}
", fill=TRUE, file="model-beta.jag")

cat(" model{
## Likelihood
for (i in 1:length(x)){
  y[i]~dbeta(a[i,u[i]],b[i,u[i]])
  logit(gamma[i])<- beta0.star+(beta1*(x[i]-mean(x[])))
  log(phi[i]) <- -delta0.star-delta1 * (x[i]-mean(x[]))
      theta[i]<- alpha.star*(1-abs((2*gamma[i])-1))
      mu[i]<-(gamma[i]-(0.5*theta[i]))/(1-theta[i])
  v[i]~ dbern(theta[i])
      u[i]<-v[i]+1
      a[i,1]<- mu[i]*phi[i]
      b[i,1]<-(1-mu[i])*phi[i]
      a[i,2]<- 1
      b[i,2]<- 1
      }
### Priors
beta1 ~ dnorm(0.01,0.01)
beta0.star ~ dnorm(0.01,0.01)
beta0 <- beta0.star-beta1 * mean(x[])
delta1 ~ dnorm(0.01,0.01)
delta0.star ~ dnorm(0.01,0.01)
delta0 <- delta0.star-delta1 * mean(x[])
alpha.star ~ dunif(0,1)
}
```

```
", fill=TRUE, file="model-betarec.jag")

jags_d <- list(x=lbm, y=Bfat)

d.ini1<-list(list(beta1=0.0,beta0.star=-1.3,delta1=0.002,delta0.star=-3)
,list(beta1=0.007,beta0.star=-1.3,delta1=0.02,delta0.star=-3),
list(beta1=0.005,beta0.star=-1.4,delta1=0.002,delta0.star=-3)
)

d.ini2<-list(list(beta1=0.0,beta0.star=-1,delta1=0.0,delta0.star=-3
,alpha.star=0.09)
,list(beta1=0.0,beta0.star=0.0,delta1=0.002,delta0.star=-3,alpha.star=0.06),
list(beta1=0.0,beta0.star=-0.1,delta1=0.001,delta0.star=-3,alpha.star=0.05)
)

mod.beta <- jags.model("model-beta.jag",
data=jags_d,
n.chains=3,
inits=d.ini1,
n.adapt=5000
)
mod.betarec <- jags.model("model-betarec.jag",
data=jags_d,
n.chains=3,
inits=d.ini2,
n.adapt=50000
)

out.beta<- coda.samples( mod.beta,
variable.names = c("beta1","beta0.star","delta1","delta0.star"),
n.iter =50000,
thin = 10 )

summary(out.beta)
```

```
out.betarec<- coda.samples( mod.betarec,
 variable.names = c("beta1","beta0.star","delta1","delta0.star","alpha.star"),
n.iter =50000,
thin = 10 )

summary(out.betarec)
out1.beta <- out.beta[[1]]
out2.beta <- out.beta[[2]]
out3.beta <- out.beta[[3]]
outn.beta <- rbind(out1.beta,out2.beta,out3.beta)
coef.beta <- apply(outn.beta,2,mean)

out1.betarec <- out.betarec[[1]]
out2.betarec <- out.betarec[[2]]
out3.betarec <- out.betarec[[3]]
outn.betarec <- rbind(out1.betarec,out2.betarec,out3.betarec)
coef.betarec <- apply(outn.betarec,2,mean)
### DIC beta
dic.beta <- dic.samples(mod.beta, n.iter=50000, thin = 10, type="pD" )
DIC.beta = sum(dic.beta$deviance)+sum(dic.beta$penalty)
### EBIC beta
p <- 4
EBIC.beta <- sum(dic.beta$deviance)+ (p*log(n))
### EAIC beta
p <- 4
EAIC.beta <- sum(dic.beta$deviance) + (2*p)
### DIC beta rec
dic.betarec <- dic.samples(mod.betarec, n.iter=5000, thin = 10, type="pD" )
DICbeta.betarec = sum(dic.betarec$deviance)+sum(dic.beta$penalty)
### EBIC beta
p <- 5
EBIC.betarec <- sum(dic.betarec$deviance)+ (p*log(n))
### EAIC beta
```

```
p <- 5
EAIC.betarec <- sum(dic.betarec$deviance) + (2*p)
```

مراجع

[1] Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle, *Springer*, Selected Papers of Hirotugu Akaike, 199-213.

[2] Anderson, D.R., Burnham, K.P. & White, G.C. (2001). Kullback Leibler information in resolving natural resource conflicts when definitive data exist, *Wildlife Society Bulletin*,1260-1270.

[3] Finley, A.O. (2013). Using JAGS in R with the rjags pachage, https://cran.r-project.org/web/packages/rjags.

[4] Bayes, C.L., Bazan, J.L. & Garsia, C. (2012). A new robust Regression model for proportions, *Bayesian Analysis*, **7**, 841-866.

[5] Blalock, HM. (1961). Evaluating the relative importance of variables, *American Sociological Review*, **26**, 866-874.

[6] Bowman, K.O. & Shenton, L.R. (2007). *The Beta distribution, Moment Method, Korl Pearson and Fisher R.A.*, Department of Statistics, U.S.A., 133-164.

[7] Bury, K. (1999). *Statistical Distributions in Engineering*, Cambridge University Press, New York.

[8] Branscum, A.J., Johnson, W.O. & Thurmond, M.C. (2007). Bayesian beta regression: application to household data and genetic distanse between foot-and-mouth disease viruses, *Australian & New Zealand Journal of Statistics*, **49**, 287-301.

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

[10] Breslow, N.E. & Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika*, **8**, 81-91.

[11] Bring J. (1994). How to standardize regression coefficients, *The American Statistican*, **48**, 209-213.

[12] Burnham, Kenneth P. & David R., Anderson (1998). *Model Selection and Inference: A Practical Information Theoretical Approach*, Springer, New York.

[13] Carlin, Bradley P. & Luis, Thomas A. (2008). *Bayesian Methods for Data Analysis*, CRC Press, U.S.A.

[14] Casella, G. (1996). Statistical theory and Monte Carlo algorithms, **5**, 249-344.

[15] Casella, G. & George, E. (1992). *An Introduction to Gibbs Sampling*, The American Statistician, **46**, 167-174.

[16] Celeux, G., Forbes, F., Robert, C. & Titterington, D. (2006). Deviance information criteria for missing data models, *Bayesian Analysis*, **1**, 651-673.

[17] Chen, M., Shao, Q. & Ebrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*, Springer, New York.

[18] Cho, H., Ibrahim, J.G., Sinha, D. & Zhu, H. (2009). Bayesian case influence diagnostics for survial models, *Biometrics*, **65**, 116-124.

[19] Clarke, B.R. & Heathcote, C.R. (1994). Robust estimation of K-component univariate normal mixture, *Annals of Institute of Statistical Mathematics*, **46**, 83-93.

[20] Congdon, P. (2014). *Applied Bayesian Modelling*, U.S., **595**.

[21] Douglas, C. Montgomery, Elizabeth, A peck & G. Geoffrey Vining (1982). *Introduction to Linear Regression Analysis*, translator S.A. Razavi Psrisei, 13-96.

[22] Efron, B. (1986). How biased is the apparent error rate of a prediction rule?, *Journal of the American Statistical Association*, **81**, 461-470.

[23] Eubank, R.L. & Kupresanin, A. (2012). *Statistical Computing in C++ and R*, Journal of Statistical Software.

[24] Ferguson, T.S. (1961). *On the Rejection of Outliers*, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, **1**, 253-287.

[25] Ferrari, S., Espaniheira, P.L. & Cribari, F. (2011). Diagonstic tools in Beta regression with varying dispersion, *Statistical Neerlandica*, **65**, 337-351.

[26] Ferrari, S.L.P., Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions, *Journal of Applied Statistics*, **31**, 799-815.

[27] Galton, F (1885). Photographic composites, *The Photographic News*, **29**, 243-245.

[28] Garcia, C., Garcia, J. & van Dorp, J.R. (2011). Modeling heavy-tailed, skewed and peaked uncertainty phenomena with bounded support, *Statistical Methods and Applications*, **20**, 463–486.

[29] Gelfand, A.E. & Dey, D.K. (1994). Bayesian model choice: asymptotics and exact calculations, *Journal of the Royal Statistical Society*, 501-514.

[30] Gelfand, A., Smith, A. (1990). Sampling based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398-409.

[31] Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (2003). *Bayesian Data Analysis*, Statistical Science.

[32] Gelman, A. & Hill, j. (2006).*Data Analysis Regression and Multilevel/ Hierarchical Models*, Cambridge University Press.

[33] Gelman, A & Rubin, DB (1992). Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**, 457-511.

[34] Geman, S., Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the bayesian restoration of images, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 721-741.

[35] Ghosh, P., Bayes, C.L. & Lachos, V.H. (2009). A robust Bayesian approach to null intercept measurement error model with application to dental data, *Computional Statistics & Data Analysis*, **53**, 1066-1079.

[36] Gilks, W.R., Richardson, S. & Speigelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice.*

[37] Greenland S., Schlesselman J.J.& Criqui M.H. (1986). The fallacy of employing standardized regression cofficients and correlations as measures of effect. *American Journal of Epidemiology*, **125**, 203-208.

[38] Gut A. (2005). *Probability: A Graduate Course*, Springer Texts in Statistics, New York.

[39] Hahn, E.D. (2008). Mixture densities for project management activity times: a robust approach to PERT, *European Journal of Operational Research*, **188**, 450-459.

[40] Hastie, Trevor, J. & Daryl, Pregibon (1992). *Generalized Linear Models In Statistical Models* in S. ed. John M., Chambers: Waolworth and Brooks/Cole Chapter6.

[41] Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their application, *Biometrika*, **57**, 97-109.

[42] Heather & Turner (2008). *Introduction to Generalized Linear Models*, ESRC National Centre fo, UK and Department of Statistics University of Warwick, uk.

[43] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.

[44] Jackman, S. (1989). *Generalized Linear Models*, Stanford University.

[45] John, C.B., Cooper (2007). The first class honours grade: an applocation of the beta distribution, *Mathematical Spectrum*, **39**, 73-74.

[46] Kass, R. & Reftery, A. (1995). Bayes factors and model uncertainty, *Journal of the American Statistical Association*, **22**, 79-86.

[47] Kendall, M.G. & Buckland, W.R. (1957). *A Dictionary of Statistical Terms*, Longman, London.

[48] Kenneth, P., Burnham & David, K., Anderson (2001). Kullback-Leibler information as a basis for strong infrence in ecological studies, *Wildlife research*, **28**, 111-119.

[49] King G. (1986). How not to lie with statistics: avoiding common mistakes in quantitative political science. *American Journal of Political Science*, **30**, 666-687.

[50] Kotz, S., Kozubowski, T.J., Podgorski K. (2001). The Laplace distribution and generalizations, *Birkhuser*, Boston.

[51] Kullback, S & Leibler, R.A. (1951). On information and sufficiency, *The Annals of Mathematical Statistics*, **22**, 79-86.

[52] Levy, H. & Duchin, R. (2004). Asset return distribution and the Investment Horizon, explaining contradictions, *The Journal of Portfolio management*, **30**, 47-62.

[53] Lindsey, James, K. (1997). *Applying Generalized Linear Models*, Springer, New York.

[54] Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions, *Statistics in Medicine*, **28**, 3049–3082.

[55] Lunn, D., Thomas, A., Best, N. & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility, *Statistics and Computing*, **10**, 325–337.

[56] Markatou, M. (1998). Mixture models robustness and the weighted likelihood methodology, *Biometrics*, **56**, 483-486.

[57] Markatou, M. (1996). Robust statistical infrence: Weighted liklihoods or usual M-estimation?, *Communications in Statistics, Theory and Methods*, **25**, 2597-2613.

[58] Miller, R.G.Jr. (1981). *Simultaneous Statistical Inference*, Springer Verlag, New York.

[59] McCullagh P. & Nelder, J.A. (1989). *Generalized Linear Models*, London.

[60] McFall Lamm, Jr.R. (2003). Asymmetric returns and optimal hedge fund portfolios, *The Journal of Alternative Investments*, **6**, 9-21.

[61] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, **21**, 1087-1092.

[62] Nelder, J.A., Wedderburn, R.W.M. (1972). Generalized linear models, *Journal of the Royal Statistical Society*, **135**, 370-384.

[63] Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables, *Political Analysis*, **9**, 325–346.

[64] Peng, F. & Dey, D.K. (1995). Bayesian analysis of outlier problems using divergence measures, *The Canadian Journal of Statistics*, **23**, 199-213.

[65] Pierre Moulin & Patrick, R., Johnstone, Kullback-Leibler Divergence and the Central Limit Theorem.

[66] Pinheiro, J.C., Bates, D.M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects models, *Jornal of Computational and Graphical Statistics*, **4**, 12-35.

[67] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press.

[68] Robert C. & Casella, G. (2010). *Introducing Monte Carlo Methods with R*, Springer, New York.

[69] Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statisticia, *Annals of Statistics*, **12**, 1151–1172.

[70] Schwarz, G. (1987). Further analysis of the data by Akaike's information criterion and the finite corrections, *Communications in Statistics, Theory and Methods*, **1**, 13-26.

[71] Simas, Alexandre B. & Barreto-Souza, Wagner & Rocha, Andrea V. (2010). Improved estimators for a general class of Beta regression, *Computational Statistics and Data Analysis*, **2**, 348-366.

[72] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Lind, A. (2002). Bayesian measures of model complexity and fit, *Jornal of the Royal Statistical Society*, **64**, 583-639.

[73] Stephane H., Michel B., John W., Polak (2006). Discrete mixture models, *Swiss Transport Research Conference.*

[74] Thomas, P., Minka (2001). *Bayesian Linear Regression.*

[75] Vehtari, A. & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison, *Statistics Surveys*, **6**, 142-228.

[76] Verkuilen, J. & Smithson, M. (2006). A better lemon squeezer? maximum likelihood regression with beta-distributed dependent variables, *Psychological Methods*, **11**, 54-71.

[77] Walter, G., Augustin, T. (2009). Bayesian linear regression-different conjugate models and their Sensitivity to prior-data conflict, *Department of Statistics*, **069**.

[78] Watanabe, S. (2013). A widely applicable Bayesian information criterion, *Journal of Machine Learning Research*, **14**, 867-897.

واژه‌نامه فارسی به انگلیسی

آب ..... Water
آماره انحراف ..... Deviance Statistics
انجمن سلطنتی آمار ..... Royal Statistical Society
استنباط بیزی ..... Bayesian Inference
الگوریتم متروپلیس-هستینگز ..... Metropolis-Hastings Algorithm
الگوریتم متروپلیس-هستینگز گام زدن تصادفی ..... Random Walk Metropolis-Hastings Algorithm
الگوریتم متروپلیس-هستینگز مستقل ..... Independent Metropolis-Hastings Algorithm
معیار اطلاع آکائیک مورد انتظار ..... Expected Akaike Information Criterion
معیار اطلاع بیزی مورد انتظار ..... Expected Bayesian Information Criterion
بیش‌پراکندگی ..... Overdispersion
پارامتر آمیختگی ..... Mixture Parameter
پارامتر دقت ..... Precision Parameter
پارامتر شکل ..... Shape Parameter
پارامتر طبیعی ..... Natural Parameter
پاسخ ..... Response
پسین ..... Posterior
پواسن ..... Poisson
پیچیدگی ..... Complexity
پیشگو ..... Predictor
پیشگوی خطی ..... Linear Predictor
جستجوی محلی ..... Local Exploration
توزیع پیشین ..... Prior Distribution
پیشین آگاهی‌بخش ..... Informative Prior
پیشین ناآگاهی‌بخش ..... Non Informative Prior
تابع پیوند لجیت ..... Logit Link Function

تابع پیوند لگاریتمی ..... Log Link Function
تابع پیوند همانی ..... Identity Link Function
تقریب لاپلاس ..... Laplace Approximation
تنومند ..... Robust
توزیع ابزاری ..... Instrumental Distribution
توزیع بتا ..... Beta Distribution
توزیع پیشنهادی ..... Proposal Distribution
توزیع پسین کناری ..... Marginal Posterior Distribution
میانگین توان‌دوم خطا ..... Mean Square Error
توزیع‌های آمیخته ..... Mixture Distributions
ثابت نرمال‌ساز ..... Normalization Constant
چگالی شرطی کامل ..... Full Conditioal Density
چوله ..... Skew
خانواده نمایی ..... Exponential Family
داده پرت ..... Outlier Data
سوزاندن ..... Burn-in
درصد ..... Percentage
درستنمایی ..... Likelihood
دوجمله‌ای ..... Binomial
رگرسیون ..... Regression
رگرسیون بتا ..... Beta Regression
رگرسیون پرابیت ..... Probit Regression
رگرسیون خطی ..... Linear Regression
رگرسیون مستطیلی بتا ..... Beta Rectangular Regression
روش‌های تربیع‌بندی ..... Quadrature Methods
روش‌های رد و پذیرش ..... Accept-Reject Methods
زنجیر مارکوف مونت کارلو ..... Markov Chain Monte Carlo
سیمان ..... Cement
شبیه‌سازی ..... Simulation
فضای پارامتر ..... Parametric Space
قانون قوی اعداد بزرگ ..... Strong Law Number
قضیه بیز ..... Bayes Theory

کولبک-لیبلر Kullback-Leibler
گاما Gamma
لجستیک Logistic
ماسه Sand
مدل‌های تعمیم‌یافته خطی Generalized Linear Models
مقاومت بتن Strength
معیار اطلاع آکائیک Akaike Information Criterion
معیار اطلاع آکائیک مورد انتظار Expected Akaike Information Criterion
معیار اطلاع انحراف Deviance Information Criterion
معیار اطلاع بیزی Bayesian Information Criterion
معیار اطلاع بیزی مورد انتظار Expected Bayesian Information Criterion
منعطف Flexible
مؤسسه ورزش استرلیا Australian Institute of Sport
مؤلفه منظم Systematic Component
میکروسیلیس Silicafume
نرمال Normal
نسبت Proportion
نمودارهای اثر Trace Plots
نمونه‌گیر گیبز Gibbs Sampler
نمونه‌گیری نقاط مهم Importance Sampling
هسته مارکوف Markov Kernel
هماتوکریت Hematocrit
هموگلبین Hemoglobin

واژه‌نامه انگلیسی به فارسی

روش‌های رد و پذیرش ..... Accept-Reject Methods
معیار اطلاع آکائیک ..... Akaike Information Criterion
مؤسسه ورزش استرلیا ..... Australian Institute of Sport
استنباط بیزی ..... Bayesian Inference
معیار اطلاع بیزی ..... Bayesian Information Criterion
قضیه بیز ..... Bayes Theory
توزیع بتا ..... Beta Distribution
رگرسیون مستطیلی بتا ..... Beta Rectangular Regression
رگرسیون بتا ..... Beta Regression
دوجمله‌ای ..... Binomial
سوزاندن ..... Burn-in
سیمان ..... Cement
پیچیدگی ..... Complexity
آماره انحراف ..... Deviance Statistics
معیار اطلاع انحراف ..... Deviance Information Criterion
معیار اطلاع آکائیک مورد انتظار ..... Expected Akaike Information Criterion
معیار اطلاع بیزی مورد انتظار ..... Expected Bayesian Information Criterion
خانواده نمایی ..... Exponential Family
منعطف ..... Flexible
چگالی شرطی کامل ..... Full Conditional Density
گاما ..... Gamma
مدل‌های تعمیم‌یافته خطی ..... Generalized Linear Models
نمونه‌گیر گیبز ..... Gibbs Sampler
هماتوکریت ..... Hematocrit
هموگلبین ..... Hemoglobin

تابع پیوند همانی ..... Identity Link Function
نمونه‌گیری نقاط مهم ..... Importance Sampling
الگوریتم متروپلیس-هستینگز مستقل ..... Independent Metropolis-Hastings Algorithm
پیشین آگاهی‌بخش ..... Informative Prior
توزیع ابزاری ..... Instrumental Distribution
کولبک-لیبلر ..... Kullback-Leibler
تقریب لاپلاس ..... Laplace Approximation
درستنمایی ..... Likelihood
پیشگوی خطی ..... Linear Predictor
رگرسیون خطی ..... Linear Regression
لجستیک ..... Logistic
تابع پیوند لجیت ..... Logit Link Function
تابع پیوند لگاریتمی ..... Log Link Function
جستجوی محلی ..... Local Exploration
توزیع پسین کناری ..... Marginal Posterior Distribution
زنجیر مارکوف مونت کارلو ..... Markov Chain Monte Carlo
هسته مارکوف ..... Markov Kernel
میانگین توان‌دوم خطا ..... Mean Square Error
الگوریتم متروپلیس-هستینگز ..... Metropolis-Hastings Algorithm
توزیع‌های آمیخته ..... Mixture Distributions
پارامتر آمیختگی ..... Mixture Parameter
پارامتر طبیعی ..... Natural Parameter
پارامتر ناآگاهی‌بخش ..... Non Informative Prior
نرمال ..... Normal
ثابت نرمال‌ساز ..... Normalization Constant
داده پرت ..... Outlier Data
بیش‌پراکندگی ..... Overdispersion
فضای پارامتر ..... Parametric Space
توزیع پیشین ..... Prior Distribution
درصد ..... Percentage
پواسن ..... Poisson
پسین ..... Posterior

پارامتر دقت ..... Precision Parameter
پیشگو ..... Predictor
رگرسیون پرابیت ..... Probit Regression
نسبت ..... Proportion
توزیع پیشنهادی ..... Proposal Distribution
روش‌های تربیع‌بندی ..... Quadrature Methods
الگوریتم متروپلیس-هستینگز گام زدن تصادفی ..... Random Walk Metropolis-Hastings Algorithm
رگرسیون ..... Regression
پاسخ ..... Response
تنومند ..... Robust
انجمن سلطنتی آمار ..... Royal Statistical Society
ماسه ..... Sand
پارامتر شکل ..... Shape Parameter
میکروسیلیس ..... Silicafume
شبیه‌سازی ..... Simulation
چوله ..... Skew
مقاومت بتن ..... Strength
قانون قوی اعداد بزرگ ..... Strong Law Number
مؤلفه منظم ..... Systematic Component
نمودارهای اثر ..... Trace Plots
آب ..... Water

Aabstract

In many regression models the response variable is continuous, restricted to the interval $(0, 1)$, such as percentages, proportions and fractions or rates. For analysing such responses, the logistic or probit regression models are not adequate and the popular model is beta regression. However, the beta regression model is not robut when there are outlying observations in the response variable. Therefore, we consider a new regression model constructed based on the beta rectangular distribution. It is shown that beta rectangular regression is more robust than the beta regression model. In this thesis, we first review the beta rectangular distribution and a new parametrization is introduced. Then the beta rectangular regression model is proposed and a Bayesian inference approach is adopted using a Markov chain Monte Carlo (MCMC) algorithm. Two simulation studies are carried out to show the better performance of the new model in the presence of outliers. Finally, application of the proposed model is shown by two applications.

Keywords: Bayesian analysis, Beta rectangular regression, Robust, MCMC, Outliers.

Shahrood University

Faculty of Mathematical Sciences

Beta rectangular regression for modeling proportion responses

Hanie Keihani

Supervisor

Dr. Hossein Baghishani

2015 September