# رهیافت بیزی نیمه پارامتری و کاربرد آن در مدل‌های با اثرات آمیخته

کیوان عسلی صاف [۱]

چکیده:

متداول ترین مدل‌ها برای تحلیل داده‌های وابسته در علوم مختلف، مدل‌های خطی تعمیم یافته با اثرهای آمیخته است. یک مسأله اساسی در استفاده از این مدل‌ها بکارگیری فرآیند مناسبی برای تولید توزیع پیشین اثرهای آمیخته است. با فرض این که توزیع این اثرها از فرآیند دیریکله با توزیع پایه و پارامتر دقت مشخص پیروی می‌کنند، مدل‌های بیزی سلسله مراتبی نیمه پارامتری برای تحلیل داده‌های وابسته در رگرسیون خطی تعمیم یافته مورد بررسی قرار می‌گیرند. بدین منظور، ابتدا فرآیند دیریکله معرفی و سپس روش بکارگیری آن به‌عنوان فرآیند تولید توزیع‌های پیشین توضیح داده خواهد شد. در ادامه، توزیع‌های پسین شرطی کامل پارامترهای مدل‌های رگرسیونی با اثرهای آمیخته را یافته و توسط رهیافت بیز، استنباط آماری را با الگوریتم نمونه گیری گیبز انجام خواهیم داد. در نهایت در قالب مثالی، مدل مذکور به داده‌های واقعی برازش داده خواهد شد تا اهمیت کاربرد روش های بیز نیمه پارامتری در مقایسه با روش‌های متداول مشخص می‌شود.

**واژه‌های کلیدی:** مدل‌های رگرسیون خطی تعمیم‌یافته، مدل‌های بیز سلسله‌مراتبی، اثرهای آمیخته، فرآیند دیریکله، نمونه‌گیری گیبز، توزیع‌های پسین شرطی کامل.

## ۱ مقدمه

واقف بودن محقق به پاره‌ای از اطلاعات قبلی در مورد مدل‌های با اثرات آمیخته و پارامترهای آن اهمیت ویژه‌ای در حصول نتایج برآوردیابی دارد. بنابراین یک مسأله‌ی اساسی در استفاده از این مدل‌ها بکارگیری فرآیند مناسبی برای تولید توزیع پیشین اثرهای آمیخته است. در این مقاله، فرض می‌کنیم که توزیع این اثرها از فرآیند دیریکله با توزیع پایه و پارامتر دقت مشخص پیروی می‌کنند، ایده ی اصلی این روش بر طبق فرگوسن [۲] [۴] و آنتونیاک [۳] [۱] می‌باشد. در نهایت برای مقایسه‌ی روش‌های ارائه شده با روش‌های متداول بیز، به برآورد پارامترهای مجهول و اثرات آمیخته در مدل‌های خطی می‌پردازیم که این امر با محاسبه توزیع‌های شرطی کامل پارامترهای

[۱] کارشناس ارشد آمار، مدرس دانشگاه محقق اردبیلی
[۲] Ferguson
[۳] Antoniak

مدل با اثرات آمیخته بر طبق ایشوارن و جیمز[۴] [۵] میسر می‌باشد. سپس با توجه به معیارهای اعتبار مدل‌ها از قبیل معیار اطلاع پراکندگی (DIC[۵])، پراکندگی[۶]، AIC[۷] و غیره، به مقایسه و سنجش روش ارائه شده با روش‌های معمول بیز خواهیم پرداخت که معیارهای اعتبار ذکر شده برای اولین بار توسط اسپیگل هالتر و همکاران [۸] [۹] مطرح گردید.

در بحث مدل سازی برخی از جوامع مورد بررسی را می‌توان به صورت افرازی از چند جامعه کوچکتر در نظر گرفت و می‌توان فرض کرد که داده‌های مورد بررسی با یک سری احتمالات خاص متعلق به این زیر جوامع هستند. لذا با توجه به افراز داده‌هایی که مورد بررسی محقق می‌باشند، برآورد پارامترهای این نوع جوامع نیازمند مدل سازی‌های مناسب می باشند که بر اساس کاتنر و همکاران[۹] [۷] به این امر پرداخته خواهد شد. علاوه بر این چگونگی اتخاذ توزیع های پیشین برای پارامترهای این نوع جوامع و مدل‌های آمیخته در آمار بیز حائز اهمیت فراوان می‌باشد. از این رو با اتخاذ پیشین‌های مناسب، روش مطلوبی را برای برآورد پارامترهای مجهول مدل‌های برازش داده شده به جوامع ذکر شده پیشنهاد می‌دهیم.

## ۲ تعاریف اولیه

قبل از اینکه روش‌های مدلسازی در جوامع آمیخته را شرح دهیم و به چگونگی استفاده از این روش‌ها بپردازیم، برخی از مفاهیمی را که در برازش مدل‌های آمیخته مفید هستند شرح می‌دهیم.

### ۱.۲ توزیع آمیخته

فرض کنید که متغیر تصادفی $Y$ بتواند از یکی از $k$ جامعه (مؤلفه) مفروض به دست آید به‌طوری که هر کدام از این مؤلفه ها دارای یک تابع چگالی احتمال متفاوت با بقیه باشند. اگر $Y$ از $i$امین مؤلفه با احتمال $w_i$ و تابع احتمال $f(Y|\theta_i)$ آمده باشد، آنگاه تابع چگالی احتمال یک مشاهد $Y$ را می‌توان به شکل آمیخته زیر معرفی کرد

$$f(Y|\phi)=\sum_{i=1}^{k} w_i f(Y|\theta_i) \qquad (1)$$

که در آن $\phi = (k, w_1, \cdots, w_k, \theta_1, \cdots, \theta_k)$، $k$ تعداد مؤلفه ها و معلوم و $w_i$ ها وزن‌های آمیختگی هستند که نامنفی بوده و مجموعشان برابر یک است و $\theta_i$ها پارامترهای نامعلوم مربوط به مؤلفه $i$ام را معرفی می‌کند.

### ۲.۲ توزیع دیریکله

توزیع دیریکله ما را به سوی درک فرآیند دیریکله آمیخته که به اختصار با DPM[۱۰] نشان خواهیم

---

[۴] Ishwaran and James
[۵] Deviance Information Criterion
[۶] Deviance
[۷] Akaike Information Criterion
[۸] Spigelhalter and et al.
[۹] Kutner and et al.
[۱۰] Dirichlet Process Mixture

داد، سوق می‌دهد. توزیع دیریکله یک توزیع چند پارامتری تعمیم یافته از توزیع بتا است. فرض کنید $\Theta = \{\theta_1, \theta_2, \cdots, \theta_n\}$ یک توزیع احتمال روی فضای گسسته $X = \{X_1, X_2, \cdots, X_n\}$ باشد، به‌طوری که $P(X = X_i) = \theta_i$، که در آن $X$ یک متغیر تصادفی در فضای $X$ می‌باشد. توزیع دیریکله روی $\Theta$ به صورت زیر است:

$$P(\Theta|\alpha, M) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{n}\Gamma(\alpha m_i)}\prod \theta_i^{\alpha m_i - 1} \quad (2)$$

که در آن $i = 1, 2, \cdots, n$ و $M = \{m_1, m_2, \cdots, m_n\}$ اندازه پایه تعریف شده روی $X$ و بردار میانگین $\Theta$ و $\alpha$ پارامتر دقتی است که چگونگی نمرکز توزیع حول $M$ را نشان می‌دهد.

## ۳.۲ فرآیند دیریکله

اگر فرض کنیم که $(X, A)$ یک فضای اندازه با یک اندازه احتمال $G_0$ باشد و $\alpha_0$ یک عدد حقیقی مثبت باشد، یک فرآیند دیریکله، توزیع یک اندازه تصادفی $G$ روی فضای $(X, A)$ می‌باشد، به‌طوری که اگر برای هر افراز $(A_1, \cdots, A_k)$ از $X$، بردار تصادفی $(G(A_1), \cdots, G(A_k))$ دارای یک توزیع دیریکله با بعد متناهی به‌صورت زیر باشد:

$$(G(A_1), \cdots, G(A_k)) \sim$$
$$Dir(\alpha_0 G_0(A_1), \cdots, \alpha_0 G_0(A_k))$$

در این صورت $G$ را دارای فرآیند دیریکله با پارامتر تمرکز $\alpha_0$ و اندازه پایهٔ $G_0$ گویند که با نماد زیر نشان داده می‌شود:

$$G \sim DP(\alpha_0 G_0)$$

### ۱.۳.۲ خواص مهم فرآیند دیریکله

فرض کنید که $G$ یک فرآیند دیریکله روی فضای $(X, A)$ با پارامترهای ثابت $G_0$ و $\alpha$ باشد، بنابر فرگوسن [۴] و آنتویاک [۱] نتایج زیر را داریم:

اگر $\pi_1, \cdots, \pi_n$ یک نمونه تصادفی $n$ تایی از $G$ باشد، داریم:

۱— اگر فرض کنیم $\pi \sim G_0$ آنگاه:

$$(\pi_2|\pi_1) \sim \frac{\alpha}{\alpha+1}G_0 + \frac{1}{\alpha+1}\delta_{\pi_1}(\pi_2)$$
$$\vdots$$
$$(\pi_n|\pi_1, \cdots, \pi_{n-1}) \sim \frac{\alpha}{\alpha+n-1}G_0$$
$$+\frac{1}{\alpha+n-1}\sum_{j=1}^{n-1}\delta_{\pi_j}(\pi_n),$$

که در آن $\delta_{\pi_j}(\pi_n)$ مشخص کننده جرم نقطه‌ای واحد در $\pi_n = \pi_j$ می‌باشد، یعنی

$$\delta_{\pi_j}(\pi_n) = \begin{cases} 1 & , \quad \pi_j = \pi_n \\ 0 & , \quad \pi_j \neq \pi_n \end{cases}$$

۲— توزیع شرطی کامل $(\pi_i|\pi_1, \cdots, \pi_{i-1}, \pi_{i+1}, \cdots, \pi_n)$ برای $i = 1, 2, \cdots, n$ به صورت زیر است:

$$f(\pi_i|\pi_1, \cdots, \pi_{i-1}, \pi_{i+1}, \cdots, \pi_n) =$$
$$\frac{\alpha}{\alpha+n-1}G_0 + \frac{1}{\alpha+n-1}\sum_{j\neq i}\delta_{\pi_j}(\pi_i) \quad (3)$$

۳— اگر $\pi = (\pi_1, \cdots, \pi_n)$ یک نمونه تصادفی $n$ تایی از $G$ باشد، آنگاه نمونه متوالی $(n+1)$ ام دارای توزیع زیر است:

$$(\pi_{n+1}|\pi) = \frac{\alpha}{\alpha+n}G_0(\pi_{n+1}) + \frac{1}{\alpha+n}\sum_{i=1}^{n}\delta_{\pi_i}(\pi_{n+1}). \quad (4)$$

بنابراین، نمونه بعدی $\pi_{n+1}$ به شرط $\pi$، مشخص کننده یک مقدار مجزای جدید با احتمال $\frac{\alpha}{\alpha+n}$ می‌باشد و به عبارت دیگر نمونه بعدی به‌طوریکنواخت از بین $n$ مقدار اولیه $\pi_1, \cdots, \pi_n$ استخراج می‌شود که مقدار جدید می‌تواند از توزیع $G_\circ$ نیز استخراج گردد.

## ۳ مدل آمیخته فرآیند دیریکله (DPM)

یک مدل آمیخته به صورت $Y_i \sim \sum_{i=1}^{k} \pi_i f(Y|\theta_i)$ را در نظر بگیرید، که در آن $Y$ به صورت آمیخته ای از توزیع‌های دارای شکل پارامتری یکسان $f$ توزیع شده‌است که این توزیع‌های پارامتری تنها درپارامترهایشان با هم متفاوت می‌باشند. همچنین فرض بر این است که برای $i = 1, \cdots, k$، پارامترهای $\theta_i$ از توزیع معلوم $G_\circ$ استخراج شده‌اند. این مدل آمیخته را می‌توان به صورت سلسله مراتبی به صورت زیر بیان کرد:

$$\begin{aligned} Y_i|c_i, \Theta &\sim f(Y|\theta_{c_i}) \\ c = ((c_1, \cdots, c_k) \mid \pi_{1:k}) &\\ &\sim Discrete(\pi_1, \cdots, \pi_k) \\ \theta_i &\sim G_\circ(\theta), \\ (\pi_1, \cdots, \pi_k) &\sim Dir(\alpha, M) \end{aligned} \qquad (۵)$$

که در آن $c_i$ ها نشانگرها یا برچسب هایی هستند که مقادیر $Y_i$ را به مقدار پارامتر $\theta_{c_i}$ نسبت می دهند و $\pi_i$ ها ضرایب آمیختگی می باشند که از یک توزیع دیریکله استخراج شده‌اند و $G_\circ$ توزیع پیشین معلوم برای پارامترهای مدل یعنی $\theta_i$ها است. برای ضرایب آمیختگی مفروض، متغیرهای نشانگر $c_i$ها دارای توزیع چند جمله‌ای می‌باشند.

توجه کنید که متغیرهای نشانگر پنهان یعنی $c_i$ ها تنها به دلیل سهولت محاسبات بکار برده شده اند. حال اگر تعداد اجزا یا موارد آمیخته از قبل معلوم باشد، پارامترها برای هر مورد می‌توانند از $G_\circ$ استخراج گردند و می‌توان توزیع احتمال دیریکله را به‌عنوان تابع چگالی احتمال ضرایب آمیختگی انتخاب کرد.

در اینجا حالت حدی این مدل را وقتی که $k\tau\infty$ بررسی می‌کنیم. فرگوسن [۴] نشان داد که در حالت حدی، فرآیند دیریکله جایگزین توزیع دیریکله می‌گردد، به طوریکه برای هر نشانگر $c_i$ به شرط نشانگرهای قبلی از توزیع چندجمله ای، $\theta_i$ متناظری وجود دارد که از توزیع $G_\circ$ استخراج می‌شود. در حالت حدی $k\tau\infty$، این برچسب ها دارای توزیعی پیوسته فرض می شوند و می توان از استفاده از این برچسب ها در مدل صرف نظر و در عوض فرض کرد که پارامترها از یک فرآیند دیریکله با اندازه پایه $G_\circ$ استخراج می‌شوند. از این رو مدل آمیخته فرآیند دیریکله $(DPM)$ به صورت زیر معرفی می‌گردد:

$$\begin{aligned} Y_i|\theta_i &\sim f(Y|\theta_i) \\ \theta_i|G &\sim G(\theta) \\ G &\sim DP(\alpha G_\circ(\theta)), \end{aligned} \qquad (۶)$$

که در آن $DP(\alpha G_{\circ}(\theta))$ فرآیند دیریکله با اندازه پایه $G_{\circ}$ و پارامتر دقت $\alpha$ را نشان می‌دهد و $G$ یک توزیع تصادفی استخراج شده از $DP$ می‌باشد. نکته قابل ذکر در اینجا این است که از آنجایی که در بخش اول مدل (٦)، $f(Y|\theta_i)$ یک توزیع معلوم و در بخش دوم $G(\theta)$ یک توزیع عمومی در نظر گرفته می‌شود، در اغلب موارد از مدل (۵) به عنوان یک مدل نیمه پارامتری سلسله مراتبی بیزی[11] نام برده می‌شود.

توزیع‌های حاشیه ای پارامترهای مجهول مدل (٦) از آن جهت حائز اهمیت می‌باشد که با در دست داشتن این توزیع‌ها، می‌توانیم با استفاده از روش نمونه‌گیری گیبز به برآورد این پارامترها می پردازیم. بنابراین در این قسمت به یافتن توزیع‌های حاشیه‌ای پارامترهای مجهول مدل می‌پردازیم.

برای یافتن توزیع‌های حاشیه‌ای پارامترهای مجهول مدل (٦)، با توجه به پیشین $\theta_i|G \sim G(\theta)$ و $G \sim DP(\alpha G_{\circ}(\theta))$ در مدل (٦) و با استفاده از خاصیت ۳ از خواص دیریکله ، می‌توانیم برای $i = ۱, \cdots, k$، توزیع حاشیه‌ای $\theta_i|(\theta_۱, \cdots, \theta_{i-۱})$ را به صورت

$$P(\theta_i = \theta|\theta_۱, \cdots, \theta_{i-۱}) = \frac{\alpha}{\alpha + i - ۱} G_{\circ}(\theta) \quad (۷)$$
$$+ \frac{۱}{\alpha + i - ۱} \sum_{j=۱}^{i-۱} \delta_{\theta_j}(\theta)$$

به دست آوریم که این فرمول توسط بلکول و مک کوئین[12] [۲] مطرح شده است.

بنابراین کاربرد $DPM$ برای یک مسأله در حالت کلی به این صورت مطرح می‌گردد که هر مدل پارامتری برای مقادیر $Y_i$ به صورت سلسله مراتبی به حالت

$$Y_i|\theta_i \sim f(Y|\theta_i)$$
$$\theta_i|\Psi \sim h(\theta|\Psi) \quad (۸)$$

می‌تواند با یک مدل $DPM$ از رابطه (٦) جایگزین گردد. البته باید در نظر داشت که در مرحله دوم، می‌توان از پیشین پارامتری $h(\theta|\Psi)$ صرف نظر کرد و در عوض آن را با یک توزیع عمومی $G$ که دارای پیشین فرآیند دیریکله است، جایگزین می‌کنیم.

حال به منظور نشان دادن نحوه بکارگیری فرآیند دیریکله به عنوان توزیع پیشین در آمار بیز و چگونگی استفاده از آن به منظور استخراج نمونه از جوامع آمیخته، در این بخش به تشریح نمونه گیری با استفاده از مدل های آمیخته فرآیند دیریکله می‌پردازیم.

## ۴ نمونه گیری پسین تحت $DPM$

مدل (٦) را در نظر بگیرید که برای $i = ۱, \cdots, n$، توزیع پسین حاشیه ای $\theta_i$ به شرط دیگر پارامترها در آن مطلوب است. بردار $\theta^{(i-)}$ را به مفهوم همهٔ

[11] $Bayesian\ Hierarchical\ Semi-Parametric\ Models$
[12] Blackwell and McQueen

$\theta_i$ها، به جز $\theta_i$، معرفی می‌کنیم. هدف کلی پیدا کردن توزیع پسین $\theta_i$ به شرط داده‌ها می‌باشد که با استفاده از قانون بیز به صورت زیر به دست می آید:

$$P(\theta_i|\theta^{(i-)}, Y_i) \propto P(Y_i|\theta_i)P(\theta_i|\theta^{(i-)}) \quad (۹)$$

که در آن احتمالات به طور ضمنی روی پارامترهای فرآیند دیریکله، شرطی شده‌اند. تابع $P(\theta_i|\theta^{(i-)})$ به‌عنوان توزیع پیشین، طبق (۶) عبارت است از

$$P(\theta_i|\theta^{(i-)}) = \frac{\alpha}{\alpha+n-1}G_0(\theta_i) + \frac{1}{\alpha+n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}\delta_{\theta_j}(\theta_i). \quad (۱۰)$$

توجه شود که $P(Y_i|\theta_i)$ به‌عنوان تابع درستنمایی و معادل $f(Y|\theta_i) = f(Y;\theta)$ در مدل (۵) می‌باشد. بنابراین توزیع پسین $\theta_i$ را می‌توان با استفاده از قانون بیز بدست آورد. در نتیجه داریم:

$$P(\theta_i = \theta|\theta^{(i-)}, Y_i) = \frac{P(Y|\theta_i)P(\theta_i|\theta^{(i-)})}{\int P(Y|\theta_i)P(\theta_i|\theta^{(i-)})d\theta_i} \quad (۱۱)$$

حال با جایگذاری (۱۰) در (۱۱) به دست می آید:

$$P(\theta_i|\theta^{(i-)}, Y_i) = \frac{\alpha P(Y_i|\theta_i)G_0(\theta_i) + P(Y_i|\theta_i)\sum_{\substack{j=1\\ j\neq i}}^{n}\delta_{\theta_j}(\theta_i)}{\int \alpha P(Y_i|\theta_i)G_0(\theta_i)d\theta_i + \int P(Y_i|\theta_i)\sum_{\substack{j=1\\ j\neq i}}^{n}\delta_{\theta_j}(\theta_i)d\theta_i}$$

که با در نظر گرفتن

$$b = \left(\alpha q_0 + \sum_{\substack{j=1\\ j\neq i}}^{n} P(Y_i;\theta_i)\right)^{-1} \quad (۱۲)$$

و

$$q_0 = \int_\theta G_0(\theta)P(Y_i|\theta)d\theta$$

توزیع پسین حاشیه‌ای پارامتر $\theta_i$ به صورت زیر به دست می آید:

$$P(\theta_i|\theta^{(i-)}, Y_i) = b\alpha G_0(\theta_i)P(Y_i;\theta_i) + \sum_{\substack{j=1\\ j\neq i}}^{n} P(Y_i;\theta_i)\delta_{\theta_j}(\theta_i) \quad (۱۳)$$

که در آن به $b$ ثابت نرمال ساز می‌گوییم و $\delta_{\theta_j}(\theta_i)$ تابع جرم نقطه‌ای جرم احتمال در نقطهٔ $\theta_i$ می‌باشد و در واقع $q_0$ نیز توزیع حاشیه‌ای $Y_i$ است. بنابراین با توجه به توزیع پسین $\theta_i$ به صورت

$$h(\theta_i|Y_i) = \frac{G_0(\theta_i)P(Y_i;\theta_i)}{\int_\theta G_0(\theta_i)P(Y_i|\theta_i)},$$

توزیع پسین حاشیه‌ای پارامتر $\theta_i$ را در (۱۳) می‌توان به صورت زیر نوشت:

$$P(\theta_i|\theta^{(i-)}, Y_i) = b\alpha\, q_0\, h(\theta_i|Y_i) + b\sum_{\substack{j=1\\ j\neq i}}^{n} P(Y_i;\theta_j)\delta_{\theta_j}(\theta_i) \quad (۱۴)$$

این عبارت را می‌توان به شکل آمیخته‌ای از توزیع پسین حاشیه‌ای روی $\theta_i$ به صورت زیر نوشت:

$$P(\theta_i|\theta^{(-i)}, Y_i) = \begin{cases} \theta_j & \text{با احتمال } P(Y_i;\theta_j) \\ \sim h(\theta|Y_i) & \text{با احتمال } \alpha q_0 \end{cases} \quad (۱۵)$$

بنابراین با در دست داشتن توزیع پسین حاشیه‌ای پارامترهای مدل، به سادگی می‌توان الگوریتم نمونه گیر گیبز را به‌کار برد.

## ۵ DPM در مدل رگرسیون خطی چندگانه با اثرهای آمیخته

مدل رگرسیون خطی با اثرهای تصادفی را به صورت زیر در نظر بگیرید:

$$Y_i = X_i\beta + Z_i b_i + e_i,$$

که در آن $i = ۱, \cdots, N$. برای یک $i$ منحصر به فرد با $n_i$ تکرار، $Y_i$ یک بردار $(n_i \times ۱)$ و $X_i$ یک ماتریس با بعد $(n_i \times p)$، به‌عنوان متغیرهای کمکی ثابت و $\beta$ بردار پارامتر $(p \times ۱)$ از ضرایب رگرسیون می باشد که معمولاً در این مدل ها به‌عنوان بردار اثرهای ثابت از آن یاد می‌شود. $Z_i$ هم یک ماتریس با بعد $(n_i \times v)$ از متغیرهای کمکی برای بردار $b_i$ است که $b_i$ نیز یک بردار $(v \times ۱)$ از اثرهای تصادفی می باشد. $e_i$ هم بردار $(n_i \times ۱)$ خطاها در نظر گرفته می‌شود. در بررسی و تحلیل این مدل‌ها، مرسوم است که فرض کنیم $b_i$ و $e_i$ مستقل هستند و هر دو دارای توزیع نرمال می‌باشند، به‌طوری که

$$\sim N_v(۰, V) \quad , \quad e_i \sim N_{n_i}(۰, \sigma^۲ I_{n_i})$$

که در آن $I_{n_i}$ ماتریس همانی با بعد $(n_i \times n_i)$ می‌باشد. تحت این فرض‌ها داریم:

$$Y_i|\beta, b_i \sim N_{n_i}(X_i\beta + Z_i b_i, \sigma^۲ I_{n_i}). \quad (۱۶)$$

مدل (۱۶) را مدل اثرهای تصادفی نرمال می‌نامیم.

از آن‌جایی که در این بخش هدف کاربرد فرآیند دیریکله در این مدل‌هاست، پس فرض نرمال بودن $b_i$ را در نظر نگرفته و در عوض فرض می کنیم $b_i \sim G$، که در آن توزیع $G$ از فرآیند دیریکله با اندازه پایه $G_۰$ پیروی می‌کند. اگر فرض شود که توزیع نمونه ای $Y_i$ها نرمال است، آنگاه یک اندازه پایه نرمال برای اثرهای تصادفی، یک مدل DPM مزدوج را مشخص می‌کنند. حال فرض کنید که توزیع بردار حاصل $Y_i$ برای $i$ امین واحد به صورت زیر باشد:

$$Y_i|\beta, b_i \sim N_{n_i}(X_i\beta + Z_i b_i, \sigma^۲ I_{n_i}).$$

با فرض $\tau = \sigma^{-۲}$، توزیع پیشین برای $\tau$ را به صورت زیر در نظر می‌گیریم:

$$\tau \sim \Gamma(\frac{\alpha_۰}{۲}, \frac{\lambda_۰}{۲}),$$

که در آن $\Gamma$ نشانگر توزیع گاما می‌باشد. بنابراین مدل فرآیند دیریکله برای رگرسیون خطی با اثرهای تصادفی خطی نرمال به صورت

$$\begin{aligned} \beta &\sim N_p(\mu_۰, \Sigma_۰), \\ b_i &\sim G, \qquad (۱۷) \\ G &\sim DP(M.N_v(۰, V)), \end{aligned}$$

معرفی می‌گردد، که در آن $\mu_۰$، $\Sigma_۰$، $M$ و $V$ پارامترهای معلوم فرض می‌شوند.

وقتی که $G$ یک پیشین کاملاً پارامتری باشد، آنگاه بعد از یافتن توزیع پسین کامل پارامترهای مدل، با استفاده از قانون بیز و انجام چند محاسبه جبری و نیز حذف عباراتی که به $\beta$ بستگی ندارند،

توزیع پسین شرطی کامل $\beta$ به صورت زیر به دست می‌آید:

$$\beta|b,\tau,Y \sim N_{n_i}(\hat{\beta},T), \qquad (۱۸)$$

که در آن

$$b=(b_1',\cdots,b_N')$$

$$Y=(Y_1,\cdots,Y_N)$$

$$T=\left(\tau\sum_{i=1}^{N}X_i'X_i+\Sigma_0^{-1}\right)^{-1}$$

$$\hat{\beta}=T\sum_{i=1}^{N}X_i'(Y_i-Z_ib_i)+\Sigma_0^{-1}\mu_0.$$

همچنین توزیع پسین شرطی کامل $\tau$ به صورت زیر محاسبه می‌شود:

$$\tau|b,\beta,Y \sim \Gamma\left(\frac{n+\alpha_0}{2},\frac{\sum_{i=1}^{N}r_i'r_i+\lambda_0}{2}\right), \qquad (۱۹)$$

که در آن

$$r_i=Y_i-X_i\beta-Z_ib_i. \quad \text{و} \quad n=\sum_{i=1}^{N}n_i$$

برای مطالعه جزئیات بیشتر می‌توانید به بوش و مک ایچرن [۱۳] [۳]، کلینمن و ابراهیم [۱۴] [۶] و کاتنر و همکاران [۱۵] [۵] مراجعه کنید.

## ۱.۵ توزیع پسین شرطی کامل اثرات تصادفی در رگرسیون خطی چندگانه با اثرات آمیخته

حال به منظور تحلیل روی هر کدام از $b_i$ ها به دنبال توزیع پسین شرطی کامل $b_i$ به شرط تمام پارامترهای موجود در مدل و از جمله سایر $b_i$ ها، یعنی $b^{(-i)}$ هستیم. وست و همکاران [۱۶] [۱۰] توزیع پسین هر کدام از $b_i$ ها را بر اساس (۷) به صورت زیر به دست آوردند:

$$\begin{aligned}
&\pi(b_i|\beta,\tau,Y,b^{(-i)})\\
&\propto\sum_{i\neq j}\phi_{n_i}(Y_i|X_i\beta+Z_ib_j,\tau^{-1}I_{n_i})\delta_{b_j}\\
&+\{M\int\phi_{n_i}(Y_i|X_i\beta+Z_ib_i^*,\tau^{-1}I_{n_i})\\
&\quad\times\phi_v(b_i^*|0,V)db_i^*\}\\
&\times\phi_{n_i}(Y_i|X_i\beta+Z_ib_i,\tau^{-1}I_{n_i})\phi_v(b_i|0,V),
\end{aligned}$$

که در آن $b^{(-i)}$ به بردار اثرهای تصادفی به استثنای جز $i$ ام اشاره دارد. همچنین چگالی نرمال چندمتغیره با $\phi_{n_i}(Y_i|X_i\beta+Z_ib_i,\tau^{-1}I_{n_i})$ و نیز با $\phi_v(b_i|0,V)$ نشان داده شده اند. پس از انجام چند محاسبه جبری داریم:

$$\begin{aligned}
&\pi(b_i|\beta,\tau,Y,b^{(-i)})\\
&\propto\Big(\sum_{i\neq j}\tau^{\frac{n_i}{2}}\exp\left[\frac{-\tau}{2}(Y_i-X_i\beta+Z_ib_j)'\right.\\
&\quad\left.(Y_i-X_i\beta+Z_ib_j)\right]\delta_{b_j}\Big)\\
&+M|Q_i|^{\frac{1}{2}}|V|^{\frac{1}{2}}\tau^{\frac{n_i}{2}}\exp\Big\{\frac{\tau}{2}[(Y_i-X_i\beta)'\\
&\quad U_i(Y_i-X_i\beta)]\Big\}\\
&\times\phi_{n_i}(Y_i|X_i\beta+Z_ib_i,\tau^{-1}I_{n_i})\phi_v(b_i|0,V),
\end{aligned}$$

که در آن

$$Q_i=(V^{-1}+\tau Z_i'Z_i)^{-1}$$

[۱۳] Bush and McEachern
[۱۴] Kleinman and Ibrahim
[۱۵] Kutner and et al.
[۱۶] West,Muller and Escobar

و (۲۰)

$$U_i = (\tau Z_i Q_i Z_i' - I_n)$$

بنابراین داریم

$$b_i|\beta,\tau,Y \sim N_v\big(\tau Q_i Z_i'(Y_i - X_i\beta), Q_i\big).$$

این نتایج به همانند استخراج نمونه از یک توزیع آمیخته می‌باشد که یک بخش از آن، توزیع نرمال و قسمت‌های دیگرش جرم احتمال هستند.
اغلب ماتریس کواریانس $V$ به عنوان اندازه پایه در مدل احتمال فرآیند دیریکله (۱۶) نامعلوم فرض می‌شود. بنابراین باید یک توزیع پیشین مناسب برای آن تعیین شود تا بتوانیم به برآورد آن بپردازیم. یک فرض آن است که

$$V^{-1} \sim W_v(d_\circ, c_\circ R_\circ),$$

که در آن $W_v(d_\circ, c_\circ R_\circ)$ نشانگر توزیع ویشارت می باشد که $c_\circ > 0$، $d_\circ > v$ و $R_\circ$ ماتریس معین مثبت با بعد $(v \times v)$ می باشد. لذا یک توزیع پیشین را به صورت زیر داریم

$$\pi(V^{-1}|d_\circ, c_\circ R_\circ) \propto$$
$$|V^{-1}|^{(d_\circ - v - 1)/2} \exp\{\frac{-1}{2} tr\big((c_\circ R_\circ)^{-1} V^{-1}\big)\}.$$

توزیع پسین شرطی کامل $V^{-1}$ به صورت زیر بدست می آید:

$$\pi(V^{-1}|b,\beta,\tau,Y) = \pi(V^{-1}|\gamma,\beta,\tau,b)$$
$$\propto |V^{-1}|^{(d_\circ + k - v - 1)/2}$$
$$\times \exp\Big\{ -\frac{1}{2} tr\big((c_\circ R_\circ)^{-1} V^{-1}$$
$$-\frac{1}{2}\sum_{l=1}^{k} \gamma_l' V^{-1} \gamma_l\big)\Big\}.$$

بنابراین

$$(V^{-1}|b,\beta,\tau,Y) \sim$$
$$W_v\big(d_\circ + k, (\frac{1}{c_\circ R_\circ} + \sum_{l=1}^{k} \gamma_l \gamma_l')^{-1}\big). \quad (۲۱)$$

حال در این قسمت به منظور بررسی بیشتر بر روی مدل‌های رگرسیون با اثرات تصادفی و چگونگ بکارگیری فرآیند دیریکله در این مدل‌ها و تحلیل تأثیر این فرآیند در مدل به بیان مثال می‌پردازیم.

**مثال ۱** در این مثال یک مدل اثرهای تصادفی نرمال را مورد بررسی قرار می‌دهیم. تحقیق مورد بحث در یکی از ایالات امریکا به انجام رسیده است. نمرات پیشرفت ریاضی [۱۷] دانش آموزان ۱۶۰ دبیرستان را که متغیر مستقل مدل می باشد، مورد تحقیق قرارمی دهیم و این متغیر را متأثر از چهار متغیر مستقل جنسیت[۱۸]،منطقه سکونت[۱۹]، اقلیت مذهبی [۲۰] و وضعیت اقتصادی [۲۱]دانش آموزان فرض می‌کنیم. با فرض توزیع نرمال برای نمرات ریاضی دانش آموزان به صورت $mathach_{ij} \sim N(0, \sigma_i^2)$ برای $i = 1, \cdots, 160$

[۱۷] $MathAchievement$
[۱۸] $Gender$
[۱۹] $Sector$
[۲۰] $ReligionMinory$
[۲۱] $Socio-economic$

و $j = ۱, \cdots, n_i$ یک مدل رگرسیونی با اثرات آمیخته را به صورت زیر برازش می‌دهیم:

$$\begin{aligned} mathach_{ij} &= \alpha_i + \beta_۱ Gender + \beta_۲ Sector \\ &+ \beta_۳ Religion + \beta_۴ Socio + e_{ij} \end{aligned}$$

که در آن $\alpha_i$ اثر تصادفی برای دبیرستان $i$ام و $n_i$ تعداد دانش آموزان در مدرسه $i$ام می‌باشد و فرض بر این است که برای $i = ۱, \cdots, ۱۶۰$ داریم:

$$\alpha_i \sim N(۰, \sigma_\alpha^۲)$$

لازم به ذکر است که برای $j = ۱, ۲, ۳, ۴$، توزیع پیشین نرمال با پارامترهای معلوم برای $\beta_j$ها فرض شده است و مولفه های واریانس مدل یعنی $\sigma_\alpha^۲$ و $\sigma_e^۲$ جز پارامترهای نامعلوم مدل محسوب می‌شوند و با فرض $\tau_e = \sigma_e^{-۲}$ و $\tau_\alpha = \sigma_\alpha^{-۲}$ توزیع های پیشین یکسان

$$\tau_\alpha \sim \Gamma(۰.۱, ۰.۱) \quad و \quad \tau_e \sim \Gamma(۰.۱, ۰.۱)$$

را برای آنها در نظر می‌گیریم که در آن $\Gamma$ نشان دهنده توزیع گاما می‌باشد.

توسط شبیه‌سازی کامپیوتری توسط نرم افزار $WinBUGS$ به برآورد پارامترهای مجهول مدل می‌پردازیم. با تولید ۶۰۰۰ نمونه و دور ریز ۱۰۰۰ و بعد از اینکه توزیع پارامترهای مجهول مدل به همگرایی مطلوب رسیدند، برآوردهای پارامترهای مجهول مدل را طبق جدول (۱) داریم.

جدول (۱). برآورد پارامترهای مجهول در مثال ۱

| | مقدار برآورد | انحراف استاندارد | کران پایین فاصله طمینان | کران بالا ی فاصله اطمینان |
|---|---|---|---|---|
| $\beta_۱$ | −۲/۸۹۱ | ۰/۲۲۰۱ | −۳/۳۲۹ | −۲/۴۷۹ |
| $\beta_۲$ | −۱/۱۲۸ | ۰/۱۶۵۸ | −۱/۴۳۳ | −۰/۷۹۷۳ |
| $\beta_۳$ | ۱/۹۱۳ | ۰/۱۱۰۶ | ۱/۶۹۳ | ۲/۱۲۴ |
| $\beta_۴$ | −۱۲۲/۰ | ۳/۰۷۷ | −۱۲۶/۷ | −۱۱۵/۶ |
| $\sigma_a^۲$ | ۸۵۳۱/۰ | ۱۰۲۷/۰ | ۶۷۱۹/۰ | ۱۰۷۱۰/۰ |
| $\sigma_e^۲$ | ۳۵/۹۳ | ۰/۶۱۳۷ | ۳۴/۷۴ | ۳۷/۱۵ |

جدول (۲). برآورد اثرهای تصادفی مدارس در مثال ۱

| | مقدار برآورد | انحراف استاندارد | کران پایین فاصله طمینان | کران بالا ی فاصله اطمینان |
|---|---|---|---|---|
| $\alpha_۱$ | ۱۰/۶۵ | ۰/۸۸۰۴ | ۸/۹۴۱ | ۱۲/۳۷ |
| $\alpha_۲$ | ۱۴/۳۶ | ۱/۲۱ | ۱۱/۹۸ | ۱۶/۷۷ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $\alpha_{۳۸}$ | ۱۱۵/۳ | ۲۰/۲۳ | ۷۴/۱۶ | ۱۴۵/۵ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $\alpha_{۱۴۳}$ | ۵/۴۱۹ | ۱/۰۶۱ | ۳/۳۳۳ | ۱۲/۳۷ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $\alpha_{۱۵۹}$ | ۱۲/۴۲ | ۱/۱۲۵ | ۱۰/۲۱ | ۱۴/۶۳ |
| $\alpha_{۱۶۰}$ | ۱۱۱/۱ | ۲۰/۲۲ | ۷۰/۱ | ۱۴۱/۲ |

تعدادی از اثرات تصادفی برآورد شده مربوط به مدارس نیز در جدول (۲) قابل مشاهده است.
با توجه به برآورد ۱۶۰ اثر تصادفی مدارس، بیشترین اثر تصادفی برابر ۱۱۵/۳ و مربوط به مدرسه ۳۸ام و کمترین اثر تصادفی نیز برابر ۵/۴۱۹ و مربوط به مدرسه ۱۴۳ ام می‌باشد. سایر اثرهای تصادفی بین این دو حد ماکزیمم و مینیمم می‌باشند.
حال برای مقایسه این مدل با مدل‌های دیگر از جمله زمانی که فرآیند دیریکله را به عنوان پیشین برای اثرهای تصادفی در مدل مورد استفاده قرار می‌دهیم، معیار $AIC$ را به‌کار می‌بریم. از این رو با توجه به رابطه

$$AIC = Deviance + ۲p_{\psi}$$

بین معیار انحراف و $AIC$ وجود دارد و پراکندگی به صورت جدول (۳) برآورد شده است، بنابراین $AIC$ برابر با ۴۶۳۳۲ به دست می‌آید. لازم به ذکر است که $p_{\psi}$ تعداد پارامترهای مجهول مدل می‌باشد.

جدول (۳) . برآورد $Deviance$ در مثال ۱

| | مقدار برآورد | انحراف استاندارد | کران پایین فاصله طمینان | کران بالا ی فاصله اطمینان |
|---|---|---|---|---|
| $Deviance$ | ۴۶۳۲۰/۰ | ۱۶/۸۴ | ۴۶۳۰۲/۰ | ۴۶۳۴۰/۰ |

**مثال ۲** در ادامه بررسی مثال قبل، در این قسمت نحوه برازش مدل به روش بیز نیمه پارامتری و چگونگی استفاده از این روش را مورد تحلیل قرار می‌دهیم. بدین منظور فرآیند دیریکله را به‌عنوان پیشین برای اثرات آمیخته استفاده می‌کنیم. از آنجایی که جامعه مورد بررسی شامل نمرات ریاضی دانش آموزان ۱۶۰ دبیرستان هستند که این دبیرستان‌ها خوشه‌های مورد بررسی در طرح فوق می‌باشند، بنابراین یک جامعه آمیخته را برای مطالعه پیش رو داریم لذا برای $i = ۱, \cdots, ۱۶۰$، $\alpha_i$ها را به‌عنوان اثرات تصادفی مدل و جزو پارامترهای مجهول در نظر می‌گیریم که بایستی برآورد گردند. یک توزیع پیشین دیریکله را با پارامترهای معلوم برای $\alpha_i$ها در نظر می‌گیریم. بنابراین نشانگرهای خوشه‌ای $Z_i$ را برای خوشه $i$ام به صورت

$$Z_i \sim Categorical(p)$$

تعریف می‌کنیم که در آن $p = (p_۱, \cdots, p_k)$ به‌عنوان بردار ضرایب آمیختگی می‌باشد که برای $j = ۱, \cdots, k$ داریم $۰ \leq p_j \leq ۱$ و $\sum_{j=۱}^{k} p_j = ۱$. با فرض $k \tau \infty$ با توجه به مفهوم فرآیند دیریکله می‌توان فرض کرد که $\alpha_j$ ها از جامعه آمیخته

$$q(.) = \sum_{j=۱}^{\infty} p_j h(.|\alpha_j)$$

استخراج می‌شوند که در آن $h(.|\alpha_j)$ تابع چگالی احتمال خوشه $i$ام می‌باشد که $\alpha_j$ از آن استخراج

می‌گردد. لازم به ذکر است که اندازه پایه مربوط به فرآیند دیریکله را نیز یک توزیع نرمال با میانگین صفر و واریانس نامعلوم و پارامتر دقت مرتبط با آن را یک اسکالر معلوم فرض می‌کنیم. اثرات تصادفی مربوط به ۱۶۰ مدرسه به صورت بالقوه در ۲ خوشه، دسته بندی می‌شوند. بنابراین تعداد خوشه ها را ۲ در نظر می‌گیریم. بعد از اجرای برنامه کامپیوتری پارامترهای مجهول به صورت جدول (۴) برآورد می‌گردند.

می‌توانیم با توجه به حدسی که در مورد تعداد خوشه‌ها زدیم، اثرات تصادفی را، در ۲ خوشه دسته‌بندی کنیم که خوشه اول اثرات تصادفی را بین صفر و ۲۰ و خوشه دوم را نیز بین ۱۰۰ و ۲۰۰ در نظر می‌گیریم. بنابراین $p_1$ و $p_2$ به ترتیب احتمال متعلق بودن اثرت تصادفی به خوشه اول و دوم را نشان می‌دهند که به صورت جدول (۵) برآورد می‌گردند. تعدادی از اثرات تصادفی مربوط به مدارس نیز به صورت جدول (۶)برآورد شده‌اند. با توجه به جدول (۶) همانند مثال یک بیشترین اثر تصادفی مربوط به مدرسه۳۸ام و برابر ۱۴/۶۲ و کمترین اثر تصادفی برآورد شده برابر ۱۱/۳۵ و مربوط به مدرسه ۱۴۳ام می باشد.حال برای مقایسه دو مدل بررسی شده در دو مثال فوق و بیان اهمیت کاربرد فرآیند دیریکله به‌عنوان توزیع پیشین برای اثرات تصادفی، به محاسبه معیار $AIC$ و مقایسه آن با مدل نخست می‌پردازیم. با توجه به مقدار انحراف معیار که به صورت جدول (۷) برآورد شده است و با توجه به رابطه معیار پراکندگی و $AIC$ به صورت

$$AIC = Deviance + ۲p_\psi$$

مقدار $AIC$ برابر ۴۶۱۲۲ به دست می آید که با مقایسه این معیار با مدل مثال یک که برابر ۴۶۳۳۲ به دست آمده بود، به این نتیجه می‌رسیم که مدل دوم با در نظر گرفتن فرآیند دیریکله به عنوان پیشین برای اثرات تصادفی نسبت به مدل اول برتری دارد. برای مطالعه جزئیات بیشتر در مورد معیارهای اعتبار مدل‌ها می‌توانید به اسپیگل هالتر و همکاران [۹] مراجعه کنید.

جدول (۴) . براورد پارامترهای مجهول در مثال ۲

| | مقدار برآورد | انحراف استاندارد | کران پایین فاصله طمینان | کران بالا ی فاصله اطمینان |
|---|---|---|---|---|
| $\beta_1$ | −۳/۳۷۴ | ۰/۲۱۶۱ | −۳/۸۰۶ | −۲/۹۴۸ − ۲/۴۷۹ |
| $\beta_2$ | −۱/۴۵۹ | ۰/۱۵۸۳ | −۱/۴۶ | −۱/۱۴۵ |
| $\beta_3$ | ۱/۸۸۴ | ۰/۱۱۱۳ | ۱/۸۸۳ | ۲/۰۹۹ |
| $\beta_4$ | ۳/۱۱ | ۰/۰۰۱۸ | ۳/۱۱۲ | ۳/۵۸۳ |
| $\sigma_e^2$ | ۳۶/۹۲ | ۰/۰۱۱۴ | ۳۶/۹۱ | ۳۸/۱۷ |

جدول (۵) . برآورد احتمالات متعلق بودن اثرات تصادفی به خوشه ۱ و ۲ در مثال ۲

| | مقدار برآورد | انحراف استاندارد | کران پایین فاصله طمینان | کران بالا ی فاصله اطمینان |
|---|---|---|---|---|
| $p_1$ | ۰/۴۹۳۸ | ۰/۰۵۲۹۱ | ۰/۳۹۰۶ | ۰/۵۹۶۲ |
| $p_2$ | ۰/۵۰۶۲ | ۰/۰۵۳۴۲ | ۰/۴۰۳۰ | ۰/۶۰۹۴ |

جدول (۶). برآورد اثرات تصادفی مدارس در مثال ۲

| | مقدار برآورد | انحراف استاندارد | کران پایین فاصله طمینان | کران بالا ی فاصله اطمینان |
|---|---|---|---|---|
| $\alpha_1$ | ۱۱/۳۵ | ۰/۲۹۸۱ | ۱۰/۷۲ | ۱۱/۹ |
| $\alpha_2$ | ۱۴/۴۵ | ۰/۷۵۰۱ | ۱۱/۴۹ | ۱۵/۱۵ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $\alpha_{38}$ | ۱۴/۶۲ | ۰/۲۷۴ | ۱۴/۱۳ | ۱۵/۱۶ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $\alpha_{143}$ | ۱۱/۳۵ | ۰/۲۹۸۸ | ۱۰/۷۲ | ۱۱/۹ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $\alpha_{159}$ | ۱۲/۸۲ | ۱/۵۸۳ | ۱۰/۹۴ | ۱۴/۹۸ |
| $\alpha_{160}$ | ۱۴/۰۳ | ۱/۲۷۱ | ۱۱/۰۶ | ۱۵/۱۴ |

جدول (۷) . برآورد $Deviance$ در مثال ۲

| | مقدار برآورد | انحراف استاندارد | کران پایین فاصله طمینان | کران بالا ی فاصله اطمینان |
|---|---|---|---|---|
| $Deviance$ | ۴۶۱۱۰/۰ | ۱۸/۴ | ۴۶۰۸۰/۰ | |

برای سنجش و مقایسه دو مدل فوق علاوه بر معیار $AIC$ می‌توانیم تک تک اثرات تصادفی را به همراه انحراف استاندارد برآورد شده برای هر کدام را نیز با هم مقایسه کنیم.

## ۶ $MDP$ در مدل رگرسیون خطی تعمیم یافته با اثرهای آمیخته

فرض کنید برای $i = ۱, \cdots, N$ و $t = ۱, \cdots, n_i$ توزیع متغیرهای تصادفی $Y_{ij}$ متعلق به خانواده نمایی و به صورت زیر باشد:

$$f(Y_{it}|\theta_{it}, \tau) = exp\{\tau(Y_{it}\theta_{it} - q(\theta_{it})) + c(Y_{it}, \tau)\}$$

که در آن

$$\begin{aligned} E(Y_{it}|\theta_{it}, \tau) &= \mu_{it} \\ &= \frac{dq(\theta_{it})}{d\theta_{it}} \end{aligned}$$

و $\tau$ یک اسکالر به‌عنوان پارامتر پراکندگی می باشد. در مدل آمیخته خطی تعمیم یافته فوق، پارامتر کانونی $\theta_{it}$ توسط تابع پیوند $h(.)$ با

متغیرهای توضیحی $X$ و متغیرهای کمکی $Z$ به صورت $\theta_{it} = h(\eta_{it})$ مرتبط می‌شود که در آن پیش بینی کننده خطی به صورت $\eta_{it} = x'_{it} + z'_{it}b_i$ است که در آن $x'_{it}$ و $z'_{it}$ به ترتب سطرهای ماتریس $X_i$ و $Z_i$ می‌باشند و تابع $h(.)$ را یک تابع مشتق پذیر یکنوا فرض می‌کنیم. لازم به ذکر است که $\eta_{it}$ به پیش بینی کننده خطی معروف است. بنابراین داریم:

$$f(Y_{it}|\theta_{it},\tau) \equiv f(Y_{it}|\beta,b_i,\tau)$$

که در آن

$$f(Y_{it}|\beta,b_i,\tau) = \qquad (۲۲)$$
$$exp\{\tau(Y_{it}h(\eta_{it}) - q(h(\eta_{it}))) + c(Y_{it},\tau)\}$$

می‌دانیم که اگر $\theta_{it} = h(\eta_{it}) = \eta_{it}$ آنگاه $h(.)$ را تابع پیوند کانونی گویند.

توجه کنید که در مدل‌های خطی تعمیم یافته ، فرض بر آن است که اثرهای تصادفی معمولاً از توزیع نرمال پیروی می‌کنند و مشاهدات در واحد $i$ام از هم مستقل می‌باشند. بنابراین تابع درستنمایی به صورت زیر است:

$$f(Y|\beta,b,\tau) \propto \prod_{۱\le i\le N}\Big(\prod_{۱\le t\le n_i} f(Y_{it}|\beta,b_i,\tau)\Big) \qquad (۲۳)$$

که در آن

$$b = (b_۱,\cdots,b_N)', Y = (Y_{۱۱},\cdots,Y_{Nn_N})$$

حال به منظور برآورد اثرات تصادفی به روش نیمه پارامتری بیزی، با استفاده از فرآیند دیریکله به عنوان توزیع پیشین برای اثرات تصادفی، به برازش مدل می‌پردازیم. لذا برای $i = ۱,\cdots,N$ فرض می‌کنیم که $b_i$ها دارای یک توزیع آزاد مثل $G$ باشند.به عبارت دیگر $G$ یک توزیع عمومی بوده و فرض می‌شود که توزیع $b_i$ها نرمال است. بنابراین یک مدل خطی تعمیم یافته با اثرات تصادفی خواهیم داشت که توزیع آن اثرات از فرآیند دیریکله پیروی می کند.برای روشن شدن بیشتر موضوع فرض کنید که توزیع $Y_{it}$ برای جز $i$ام در زمان $t$، بر طبق (۲۲) برابر $f(Y_{it}|\beta,b_i,\tau)$ باشد. در ادامه ، توزیع‌های پیشین برای پارامترهای مدل آمیخته فرآیند دیریکله در رگرسیون خطی تعمیم یافته را به صورت زیر در نظر می‌گیریم:

$$\tau \sim \Gamma(\alpha_\circ,\lambda_\circ)$$

و

$$\beta \sim N_p(\mu_\circ,\Sigma_\circ)$$

وبا فرض توزیع $G$ برای $b_i$ها، داریم:

$$G \sim DP(M.N_v(\circ,V)) \qquad (۲۴)$$

که در آن برای ماتریس $V^{-۱}$ توزیع پیشین ویشارت با پارامترهای معلوم را فرض می‌کنیم. بنابراین با توجه به پیشین های فوق به منظور برآورد پارامترهای مجهول مدل توزیع پسین توأم پارامترها را به صورت زیر به دست می‌آوریم:

$$\pi(\beta,b,\tau) \propto expLnf(Y|\beta,b,\tau)\pi(\beta,b,\tau)$$
$$\propto exp\{\sum_{i=۱}^{N}\sum_{t=۱}^{n_i} Lnf(Y_{it}|\beta,b_i,\tau)$$
$$-\frac{۱}{۲}(\beta-\mu_\circ)'\Sigma_\circ^{-۱}(\beta-\mu_\circ)\} \qquad (۲۵)$$

$$\times\{-\tau\lambda_{\circ}-\frac{1}{2}\sum_{i=1}^{N}b_i'V^{-1}b_i\}\times\tau^{\alpha_{\circ}-1}$$

که در آن $\pi(\beta,b,\tau)$ نشان‌دهنده چگالی پیشین توام $\beta$ و $b$ و$\tau$ می‌باشد. حال به منظور استنباط برای این پارامترها، ابتدا چگالی پسین شرطی کامل آنها را به‌دست می‌آوریم. از این رو در عبارت (۲۵) با حذف عباراتی که به $\beta$ بستگی ندارند، چگالی پسین حاشیه‌ای شرطی کامل $\beta$ را به صورت

$$\pi(\beta|Y,b,\tau)\propto exp\{\sum_{i=1}^{N}\sum_{t=1}^{n_i}Lnf(Y_{it}|\beta,b_i,\tau)$$
$$-\frac{1}{2}(\beta-\mu_{\circ})'\Sigma_{\circ}^{-1}(\beta-\mu_{\circ})\}$$

به دست می‌آوریم و از آنجایی که این توزیع شکل صریحی ندارد، با کمک روش‌های $MCMC$ و به ویژه نمونه‌گیر گیبز می توان به استخراج نمونه از این توزیع پرداخت. توزیع پسین حاشیه‌ای شرطی کامل نیز به صورت

$$\pi(\tau|\beta,b,Y)$$
$$\propto exp\{\sum_{i=1}^{N}\sum_{t=1}^{n_i}c(Y_{it},\tau)\}$$
$$\times exp\{\sum_{i=1}^{N}\sum_{t=1}^{n_i}(Y_{it}\theta_{it}-q(\theta it))-\tau\lambda_{\circ}\}$$

می‌باشد که با نمونه‌گیری گیبز از این توزیع می توان به برآورد $\tau$ پرداخت. برای مطالعه مثال های بیشتر در زمینه کاربرد $MDP$ در رگرسیون خطی تعمیم یافته با اثرهای آمیخته می‌توانید به اولسن و همکاران[۲۲] [8] مراجعه کنید.

## ۷ نتیجه‌گیری

در این مقاله بعد از آشنایی با اهمیت کاربرد فرآیند دیریکله به عنوان پیشین برای اثرات تصادفی در مدل‌های رگرسیون بیزی با اثرات آمیخته به تعمیم این موضوع پرداختیم و کاربرد مدل‌های دیریکله آمیخته $(DPM)$ را برای مدل‌های رگرسیون خطی تعمیم یافته با اثرات آمیخته در حالت بیزی مورد بررسی قرار دادیم و در ادامه با ارائه مثالی کاربردی و شبیه‌سازی کامپیوتری با نرم افزار $WinBUGS$ نحوه برازش $(DPM)$ برای این مدل‌ها را تشریح کردیم.

در نهایت به مقایسه مدل های برازش داده شده پرداختیم. نتیجه ای که می‌توان بی شک به آن اشاره کرد، بهبود برآوردگرهایی می‌باشد که برای پارامترها و اثرات آمیخته یافتیم و دلیل این مدعا معیارهای سنجش اعتبار مدل از قبیل $AIC$ و $Deviance$ می‌باشند.

## مراجع

[1] Antoniak, C.E. (1974), Mixture of Dirichlet processes with applications to non-parametric problems, *Annals of Statistics*, 2, 1152-1174.

[۲۲] Ohlsen and et al.

[2] Blackwell, D. and McQueen, J.B. (1973), Ferguson distribution via Polya Urn schemes, *Annals of Statistics*, 1, 353-355.

[3] Bush, C. A. and McEachern, S. N. (1996), A semi-parametric Bayesian model for randomized block designs, *Biometrika,* 83, 275-286 .

[4] Ferguson, T.S. (1973), A Bayesian analysis of some nonparametric problems, *Annals of Statistics* 1, 209-230.

[5] Ishwaran, H. and James, L.F. (2001), Gibbs sampling methods for stick-breaking priors, *Journal of the American Statistical Association,* 96, 161-173.

[6] Kleinman, K.P. and ibrahim, J.G. (1998), A semiparametric Bayesian approach to the random effects model, *Biometrics,* 54, 921-938.

[7] Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005), *Applied Linear Statistical Models*, 5th Ed, New York: McGraw-Hill/Irwin.

[8] Ohlssen, D.I., Sharples, L.D. and Spiegelhalter, D.J. (2006), Flexible random-effecs models using Bayesian semiparametric models, *Statistics in Medicine* 26(9), 2088-2112.

[9] Spigelhalter, D.J., Best, N.G., Carlin, B.P. and Linde, A.V. (2002), Bayesian measures of model comolexity and fit, *Journal of the Royal Statistical Society B.* 64(4), 583-639

[10] West, M., Muller, P. and Escobar, M.D. (1994), *Hierarical Priors and Mixture Models with Application in Regression and Density Estination*, Aspects of Uncertainly, eds. P.R. Freeman and A. FM. Smith, New York: Willey, 363-386.