برآورد ماکسیمم درستنمایی پارامترها در مدل خطی تابعی تعمیم‌یافته

فاطمه پاپی، پرویز ملک زاده، فاطمه حسینی[۱]

تاریخ دریافت: ۹۷/۱۰/۱۸
تاریخ پذیرش: ۹۹/۳/۱

چکیده:

گاهی در عمل داده‌ها به صورت تابعی از یک متغیر دیگر هستند که به این نوع داده‌ها، داده‌های تابعی گفته می‌شود. اگر متغیر پاسخ اسکالر و به صورت رسته‌ای یا گسسته باشد و متغیرهای کمکی به صورت تابعی، آنگاه برای تحلیل این نوع داده‌ها از مدل خطی تابعی تعمیم‌یافته استفاده می‌شود. در این مقاله یک مدل بریده‌شده خطی تابعی تعمیم‌یافته بررسی و برای به دست آوردن برآورد پارامترهای مدل از یک رهیافت ماکسیمم درستنمایی استفاده می‌شود. درنهایت در یک مطالعه شبیه‌سازی و دو مثال کاربردی مدل و روش‌های ارائه‌شده پیاده‌سازی می‌شوند.

واژه‌های کلیدی: عملگر کوواریانس، توابع ویژه، رگرسیون تابعی، مدل خطی تابعی تعمیم‌یافته، بسط کارهونن لوئه‌و.

۱ مقدمه

گاهی در عمل متغیر کمی به صورت گسسته یا رسته‌ای است که مدل‌های خطی برای تحلیل این نوع مسائل معتبر نیست و از مدل‌های خطی تعمیم‌یافته[۲] استفاده می‌شود. اولین بار نلدر و ودربرن [۱۲] و به‌طور مفصل‌تر مک‌کلاو و نلدر [۹] مدل‌های خطی تعمیم‌یافته را معرفی نمودند و این مدل‌ها را برای مدل بندی متغیرهای پاسخ رسته‌ای و گسسته پیشنهاد دادند. چون اکثر توزیع‌های گسسته و پیوسته عضو خانواده نمایی هستند در مدل‌های خطی تعمیم‌یافته متغیر پاسخ متعلق به خانواده نمایی فرض می‌شود. گاهی در عمل با داده‌هایی سروکار داریم که به‌طور ذاتی به صورت تابعی از یک متغیر دیگر مانند زمان،... هستند که به این نوع داده‌ها، داده‌های تابعی می‌گویند. ممکن است در عمل با مطالعاتی مواجه شویم که متغیر پاسخ گسسته و متغیرهای کمکی به صورت تابعی هستند که در این صورت از مدل‌های خطی تابعی تعمیم‌یافته[۳] که کلاس بزرگ‌تری از مدل خطی تابعی و شامل آن است استفاده می‌شود. برای تعریف این مدل معمولاً فرض می‌کنند متغیر وابسته Y_i، $i=1,...,n$ یک متغیر تصادفی با مقادیر اسکالر است که پیوسته یا گسسته می‌باشد و متغیر کمکی به صورت یک منحنی تصادفی $\{X_i(t), t\in\tau\}$، $i=1,...,n$ متناظر با یک فرایند تصادفی با میانگین صفر و روی فاصله τ است. مشابه مدل خطی تعمیم‌یافته در مدل خطی تابعی تعمیم‌یافته توزیع متغیر پاسخ متعلق به خانواده نمایی فرض می‌شود و مدل به صورت $Y_i=g(\alpha+\int\beta(t)X_i(t)dt)+e_i$ بیان می‌شود که α عرض از مبدأ و $\beta(t)$ توابع پارامتری رگرسیونی نامیده می‌شوند. مدل‌های خطی تابعی توسط افرادی مثل رمسی و سیلورمن [۱۳]، فاراوی [۸]، کاردوت و همکاران [۲] و...، موردمطالعه قرارگرفته‌اند. در این مقاله مدل خطی تابعی تعمیم‌یافته‌ای موردبررسی قرارگرفته است که متغیر پاسخ اسکالر و به صورت رسته‌ای یا گسسته و متغیرهای کمکی به شکل تابعی باشند. برای مدل بندی توابع و برآورد پارامترها از یک مدل بریده‌شده خطی تابعی تعمیم‌یافته استفاده و از رهیافت شبه درستنمایی به بررسی این مدل پرداخته‌شده است. ساختار مقاله به این صورت است که در بخش دوم به معرفی مدل‌های خطی تابعی تعمیم‌یافته و یک مدل خاص خطی تابعی تعمیم‌یافته پرداخته می‌شود. در بخش سوم برآورد پارامترهای مدل با رهیافت شبه درستنمایی بررسی و با استفاده از یک رهیافت مجانبی آزمون فرضیه و بازه اطمینان تقریبی هم‌زمان برای پارامترهای تابعی رگرسیونی به دست آورده می‌شود. در بخش چهارم در خصوص انتخاب تعداد رتبه در مدل بریده‌شده بحث و درنهایت در بخش پنجم بر روی چند مثال مدل و روش برآورد پیاده‌سازی می‌شود.

[۱] دانشگاه سمنان، گروه آمار fatemeh.hoseini@semnan.ac.ir

[2] Generalized Linear Model
[3] Generalized Functional Linear Model

۲ مدل خطی تابعی تعمیم‌یافته

اگر هدف بررسی ارتباط متغیر پاسخ رسته‌ای با برخی متغیرهای کمکی تابعی است، از مدل خطی تابعی تعمیم‌یافته که کلاس بزرگ‌تری از مدل خطی تابعی و شامل مدل خطی تابعی است، می‌توان استفاده نمود. فرض کنید مشاهده برای i امین واحد آزمایش به صورت $(\{X_i(t), t\in\tau\}, Y_i)$، $i=1,\ldots,n$ باشد که مستقل و هم‌توزیع هستند. متغیر کمکی $t\in\tau$ و $X(t)$ به صورت یک منحنی تصادفی است و متناظر با یک فرآیند تصادفی با میانگین صفر و روی فاصله τ است. متغیر وابسته یا پاسخ Y یک متغیر تصادفی با مقادیر اسکالر است که می‌تواند پیوسته یا گسسته باشد. برای مثال حالتی خاص از رگرسیون تابعی دوجمله‌ای، یعنی $Y\in\{0,1\}$ داشته باشیم. فرض کنید $g(\cdot)$ یک تابع پیوند یکنوا مشتق‌پذیر معلوم است، $\beta(\cdot)$ پارامترهای رگرسیونی تابعی و $\sigma^2(\cdot)$ واریانس باشند در این صورت با تعیین یک اندازه مثل $d\omega$ روی τ، پیشگوی خطی مدل به صورت

$$\eta=\alpha+\int\beta(t)X(t)d\omega(\eta)$$

تعریف می‌شود که

$$E(Y\mid X(t),t\in\tau)=\mu,\ \ \mu=g(\eta)$$
$$\mathrm{Var}(Y\mid X(t),t\in\tau)=\sigma^2(\mu)=\tilde{\sigma}^2(\eta)$$

هستند. مشابه مدل خطی تعمیم‌یافته در مدل خطی تعمیم‌یافته تابع توزیعی Y متعلق به خانواده نمایی فرض می‌شود. مدل خطی تعمیم‌یافته تابعی معمولاً به صورت

$$Y_i=g\big(\alpha+\int\beta(t)X_i(t)d\omega(t)\big)+e_i,\quad i=1,\ldots,n \qquad (1)$$

بیان می‌شود که در آن فرض می‌شود

$$E(e\mid X(t),t\in\tau)=0$$
$$\mathrm{Var}(e\mid X(t),t\in\tau)=\sigma^2(\mu)=\tilde{\sigma}^2(\eta)$$

و α یک ثابت عرض از مبدأ است و برای هر t، $E(X(t))=0$ و خطاها (e_i) مستقل هستند و انتگرال با توجه به‌اندازه $d\omega(t)$ برای توازن وزن نامنفی $\nu(\cdot)$ به طوری که اگر $t\in\tau$، $\nu(t)>0$ و اگر $t\notin\tau$، $\nu(t)=0$ و $d\omega(t)=\nu(t)d(t)$ در نظر گرفته می‌شود و معمولاً $\nu(t)=1_{\{t\in\tau\}}$ و $d\omega(t)=dt$ است. فرض کنید $\sigma^2=E\{\tilde{\sigma}^2(\eta)\}$ آنگاه

$$\mathrm{Var}(e)=\mathrm{Var}\{E(e\mid X(t),t\in\tau\}+E(\mathrm{Var}(e\mid X(t),t\in\tau\}$$
$$=E\{\tilde{\sigma}^2(\eta)\}=\sigma^2$$

همچنین $E(e)=0$. قرار دهید $j=1,2,\ldots$ و ρ_j یک پایه یکامتعامد از فضای تابعی هیلبرت $L^2(d\omega)$ که

$$\int_\tau\rho_j(t)\rho_k(t)d\omega(t)=\delta_{jk},$$

آنگاه فرآیند پیشگویی $X(t)$ و تابع پارامتر $\beta(t)$ با استفاده از قضیه کارهونن-لوئه‌و می‌توانند به صورت

$$\beta(t)=\sum_{j=1}^{\infty}\beta_j\rho_j(t)$$
$$X(t)=\sum_{j=1}^{\infty}\varepsilon_j\rho_j(t),$$

بسط داده شوند، که در آن $E(\varepsilon_j)=0$ و

$$\varepsilon_j=\int X(t)\rho_j(t)d\omega(t),$$
$$\beta_j=\int\beta(t)\rho_j(t)d\omega(t),\ \ \sum\beta_j^2<\infty.$$

با در نظر گرفتن $\sigma_j^2=E(\varepsilon_j^2)$،

$$\sum\sigma_j^2=\int E(X^2(t))d\omega(t)<\infty.$$

اکنون با توجه به توابع یکامتعامد ρ_j می‌توان نتیجه گرفت که

$$\int\beta(t)X(t)d\omega(t)=\sum_{j=1}^{\infty}\beta_j\varepsilon_j.$$

معمولاً از خطاهای استانداردشده که به صورت

$$e'=e\sigma(\mu)=e\tilde{\sigma}(\eta)$$

هستند استفاده می‌شود، به طوری که $E(e'\mid X)=0$، $E(e')=0$ و $E(e'^2)=1$ و فرض می‌کنیم $E(e^4)=\mu_4<\infty$. در مدل (۱) ساختار خطاها (e_i) لازم نیست که حتماً متعلق به خانواده نمایی باشد.

۱.۲ مدل بریده‌شده خطی تابعی تعمیم‌یافته

در این مدل از ایده برش[4] استفاده می‌شود. با تعریف

$$V_p=\sum_{j=p+1}^{\infty}\beta_j\varepsilon_j$$
$$U_p=\alpha+\sum_{j=1}^{p}\beta_j\varepsilon_j$$

آنگاه می‌توان نوشت

$$E(Y\mid X(t),t\in\tau)=g(\alpha+\sum_{j=1}^{\infty}\beta_j\varepsilon_j)=g(U_p+V_p).$$

با شرطی کردن روی p مؤلفه اول و به کار بردن تابع توزیع شرطی $F_{V_p|U_p}$، منجر به تابع پیوند بریده g_p به صورت

$$E(Y\mid U_p)=g_p(U_p)=E[g(U_p+V_p)\mid U_p]$$
$$=\int g(U_p+s)dF_{V_p|U_p}(s)$$

[4] Truncation

می‌شود. برای تقریب زدن مدل کامل با تابع پیوند بریده‌شده، تحت شرط $|g'(\cdot)|^{۲} \leqslant c$ مولر و استدمولر [۱۱] نشان دادند که

$$\begin{aligned}&\left\{\int\Big[g(U_p+V_p)-g(U_p+s)\Big]dF_{V_P|U_p}(s)\right\}^{۲}\\&\leqslant\int g'(\xi)^{۲}(V_p-s)^{۲}dF_{V_P|U_p}(s)\\&\leqslant ۲c\int(V_p^{۲}+s^{۲})dF_{V_P|U_p}(s)\end{aligned}$$

و بنابراین:

$$\begin{aligned}&E\Big(\Big(g(U_p+V_p)-g_p(U_p)\Big)^{۲}\Big)\\&=E\left(\int\Big[g(U_p+Vp)-g(U_p+s)\Big]dF_{V_P|U_p}(s)\right)^{۲}\\&\leqslant ۲cE\Big(V_p^{۲}+E(V_p^{۲}\mid U_p)\Big)=۴cE(V_p^{۲})\\&\leqslant ۴c\sum_{j=p+۱}^{\infty}\beta_j^{۲}\sum_{j=p+۱}^{\infty}\sigma_j^{۲}. \qquad (۲)\end{aligned}$$

خطای تقریبی مدل بریده‌شده به‌طور مستقیم به $\mathrm{Var}(V_p)$ مربوط می‌شود و توسط فرمول $\sigma_j^{۲}=\mathrm{Var}(\varepsilon_j)$، $j=۱,۲,\ldots$ کنترل می‌شود. اکنون با در نظر گرفتن

$$\varepsilon_j^{(i)}=\int X_i(t)p_j(t)d\omega(t),\ i=۱,\ldots,n$$

مدل کامل با خطاهای استانداردشده e_i' به صورت

$$Y_i=g\Big(\alpha+\sum_{j=۱}^{\infty}\beta_j\varepsilon_j^{(i)}\Big)+e_i'\tilde{\sigma}\Big(\alpha+\sum_{j=۱}^{\infty}\beta_j\varepsilon_j^{(i)}\Big),$$

است. با پیشگوهای خطی بریده‌شده η و میانگین‌های μ،

$$\eta_i=\alpha+\sum_{j=۱}^{p}\beta_j\varepsilon_j^{(i)},\ \ \mu_i=g(\eta_i),\ i=۱,\ldots,n$$

و p- مدل بریده‌شده به صورت

$$Y_i^{(p)}=g_p\Big(\alpha+\sum_{j=۱}^{p}\beta_j\varepsilon_j^{(i)}\Big)+e_i'\tilde{\sigma}_p\Big(\alpha+\sum_{j=۱}^{p}\beta_j\varepsilon_j^{(i)}\Big), \qquad (۳)$$

تعریف می‌شود، که در آن $\tilde{\sigma}_p$ و g_p ثابت هستند. در مدل بریده‌شده (۳) برای در نظر گرفتن عرض از مبدأ مجموع از یک قرار داده و فرض می‌شود $\varepsilon_۰^{(i)}=۱$ و $\beta_۰=\alpha$.

۳ برآورد پارامترهای مدل خطی تابعی تعمیم‌یافته

برای برآورد پارامترها $\beta'=(\beta_۰,\ldots\beta_p)$ با ثابت نگه‌داشتن p از روش شبه درستنمایی ودربرن [۱۵] باید معادله امتیاز[5] به صورت

$$U(\beta)=۰ \qquad (۴)$$

حل شود. با فرض معلوم بودن $\sigma^{۲}$ و g و با در نظر گرفتن $\varepsilon^{(i)}=(\varepsilon_۰^{(i)},\ldots,\varepsilon_p^{(i)})'$ و η_i به صورت

$$\eta_i=\sum_{j=۰}^{p}\beta_j\varepsilon_j^{(i)},\mu_i=g(\eta_i),i=۱,\ldots,n$$

تابع امتیاز بردار-مقدار به صورت

$$U(\beta)=\sum_{i=۱}^{n}(Y_i-\mu_i)g'(\eta_i)\varepsilon^{(i)}/\sigma^{۲}(\mu_i) \qquad (۵)$$

تعریف می‌شود، (مولر و استدمولر، [۱۱]). جواب معادله امتیاز (۵) به صورت

$$\hat{\beta}'=(\hat{\beta}_۰,\ldots,\hat{\beta}_p),\ \ \hat{\alpha}=\hat{\beta}_۰ \qquad (۶)$$

فرض می‌شوند. برای حل راحت‌تر معادله امتیاز (۵) ماتریس‌های

$$\begin{aligned}&D=D_{n,p}=\Big(g'(\eta_i)\varepsilon_k^{(i)}/\sigma(\mu_i)\Big),\ \ ۱\leqslant i\leqslant n,\ \ ۰\leqslant k\leqslant p,\\&V=V_{n,p}=diag\Big(\sigma^{۲}(\mu_۱),\ldots,\sigma^{۲}(\mu_n)\Big),\ \ ۱\leqslant i\leqslant n,\end{aligned}$$

تعریف می‌شوند. و با به کار بردن μ ،ε ،η و تعریف

$$\begin{aligned}&\Gamma=\Gamma_p=(\gamma_{kl})_{۰\leqslant k,\,l\leqslant p},\\&\gamma_{kl}=E\Big(\frac{g''^{۲}(\eta)}{\sigma^{۲}(\mu)}\varepsilon_l\varepsilon_l\Big),\\&\Gamma^{-۱}=(\xi_{kl})_{۰\leqslant k,l\leqslant p}\end{aligned} \qquad (۷)$$

می‌توان نوشت

$$\Gamma=\frac{۱}{n}E(D'D)$$

که یک ماتریس متقارن معین مثبت است و معکوس آن وجود دارد، که در غیر این صورت

$$E\left(\Big(\sum_{k=۰}^{p}\alpha_k\varepsilon_k g'(\eta)/\sigma(\mu)\Big)^{۲}\right)=۰$$

برای ثابت‌های غیرصفر $\alpha_۰,\ldots,\alpha_p$ برقرار نمی‌شود. اکنون با در نظر گرفتن بردارهای $Y'=(Y_۱,\ldots,Y_n)$، $\mu'=(\mu_۱,\ldots,\mu_n)$ معادله $U(\beta)=۰$ را می‌توان به صورت ماتریس

$$D'V^{\frac{-۱}{۲}}(Y-\mu)=۰$$

نوشت. این معادله معمولاً با روش‌های تکرار حل می‌شود و βها برآورد می‌شوند. تحت فرض $\frac{۱}{n}E(D'D)=\Gamma_P$ یک ماتریس معین مثبت برای هر p است، آنگاه برای هر p ثابت، یک راه‌حل یکتا وجود خواهد داشت. در روش بالا $\sigma^{۲}(\cdot),g(\cdot)$ هر دو معلوم فرض می‌شوند، اما می‌توان حالتی را فرض کرد که هردو نامعلوم هستند که با در نظر اینکه هر دو تابع هموار هستند می‌توان از مدل‌های رگرسیونی شبه درستنمایی نیمه‌پارامتری $SPQR$[6] چیپو و مولر [۴، ۵]

[5] Score Equation

[6] Semi Parametric Quasi-Likelihood Regression Model

استفاده کرد. چون تابع پیوند دلخواه است برای محدودیت‌های همواری و یکنوایی باید پارامترها و برآوردها

$$\|\beta\| = ۱,\ \|\hat{\beta}\| = ۱$$

باشند. قرار دهید $\hat{\eta}_i = \sum_{j=۰}^{p} \hat{\beta}_j \epsilon_j^{(i)}$ و برآورد تابع پیوند به صورت $\hat{g}(\cdot)$ و مشتق اول آن $\hat{g}'(\cdot)$ از هموارسازی نمودار پراکنش $(\hat{\eta}_i, Y_i)$ و برآورد تابع واریانس $\hat{\sigma}^۲(.)$ با هموارسازی نمودار پراکندگی

$$(\hat{\mu}_i, \hat{\epsilon}_i^۲)_{i=۱,\ldots,n}$$

به دست می‌آید، که در آن $\hat{\mu}_i = \hat{g}(\hat{\eta}_i)$ برآورد میانگین پاسخ و مجذور مانده‌ها $\hat{\epsilon}_i^۲ = (Y_i - \hat{\mu}_i)^۲$ است. اکنون با استفاده از معادله امتیاز نیمه‌پارامتری

$$U(\beta) = \sum_{i=۱}^{n} \Big(Y_i - \hat{g}(\eta_i)\Big) \hat{g}'(\eta_i) \epsilon^{(i)} / \hat{\sigma}^۲\Big(\hat{g}(\eta_i)\Big) \qquad (۸)$$

برآوردها به صورت $\hat{\beta} = (\hat{\beta}_۰, \ldots, \hat{\beta}_p)$ نتیجه می‌شود و درنهایت با حل معادله امتیاز (۵) و (۸) برآوردهای پارامترهای رگرسیونی تابعی به صورت

$$\hat{\beta}(t) = \hat{\beta}_۰ + \sum_{j=۱}^{p} \hat{\beta}_j \rho_j(t) \qquad (۹)$$

حاصل می‌شود. ماتریس‌های D و Γ برای حالت $SPQR$ اصلاح و جایگزین برآوردها می‌شوند.

۱.۳ بازه اطمینان و آزمون فرضیه برای پارامترهای رگرسیونی تابعی

یک تابع هسته انتگرال‌پذیر $L^۲$، به صورت $R(s,t) : \tau^۲ \longrightarrow \mathbb{R}$ در نظر بگیرید و تعریف کنید A_R را به صورت یک عملگر انتگرال خطی

$$A_R : L^۲(d\omega) \longrightarrow L^۲(d\omega)$$

روی فضای هیلبرت $L^۲(d\omega)$ برای $f \in L^۲(d\omega)$،

$$\Big(A_R f\Big)(t) = \int f(s) R(s,t) d\omega(s). \qquad (۱۰)$$

عملگرهای A_R عملگرهای هیلبرت–اشمیت خودالحاقی فشرده[۷] هستند. اگر

$$\int |R(s,t)|^۲ d\omega(s) d\omega(t) < \infty,$$

آنگاه مطابق با کانوی [۶] می‌توانند به صورت قطری شده باشند.

حالت‌های خاصی از عملگرهای انتگرال که در این مقاله موردنظر هستند عملگر اتوکوواریانس A_K از X با هسته

$$k(s,t) = cov(X(s), X(t)) = E(X(s)X(t)), \qquad (۱۱)$$

و عملگر اتوکواریانس تعمیم‌یافته A_G با هسته

$$G(s,t) = E\left(\frac{g'(\eta)^۲}{\sigma^۲(\mu)} X(s)X(t)\right), \qquad (۱۲)$$

عملگرهای هیلبرت–اشمیت هستند. A_R یک فضای متریک در $L^۲$ به صورت

$$\begin{aligned} d_R^۲(f,g) &= \int \big(f(t) - g(t)\big)\big(A_R(f-g)\big)(t) d\omega(t) \\ &= \int\int \big(f(s) - g(s)\big)\big(f(t) - g(t)\big) R(s,t) d\omega(s) d\omega(t) \end{aligned}$$

تولید می‌کند، که برای $g \in L^۲(d\omega)$، f و پایه یکامتعامد دلخواه $\{\rho_j, j = ۱,۲,\ldots\}$، هسته هیلبرت–اشمیت R را می‌توان به صورت

$$R(s,t) = \sum r_{kl} \rho_k(s) \rho_l(t),$$

برای ضرایب مناسب $\{r_{kl}, k,l = ۱,۲,\ldots\}$ دانفورد و استوارتز [۷] بیان کرد. با به کار بردن هر تابع مشخص $h \in L^۲$

$$h_{\rho,j} = \int h(s) \rho_j(s) d\omega(s),$$

و با نشان دادن توابع ویژه و مقادیر ویژه A_R با

$$\{\rho_j^R, \lambda_j^R,\ \ j = ۱,۲,\ldots\}$$

فاصله d_R را می‌توان به صورت

$$\begin{aligned} d_R^۲(f,g) &= \sum_{k,l} r_{k,l} \Big(f_{\rho,k} - g_{\rho,k}\Big)\Big(f_{\rho,k} - g_{\rho,l}\Big) \\ &= \sum_{k} \lambda_k^R \Big(f_{\rho^R,k} - g_{\rho^R,k}\Big)^۲, \end{aligned} \qquad (۱۳)$$

بیان کرد. برای مدل p_n – بریده‌شده مربوط به مدل (۱) متر d_G به کار می‌رود که نسبت به متر $L^۲$ تحت شرایط ساده‌تری حدهای مجانبی محاسبه می‌شوند. این متر برای مدل بریده‌شده به صورت (۱)

$$\begin{aligned} d_G^۲(\hat{\beta},\beta) = \int\int &\Big(\hat{\beta}(s) - \beta(s)\Big)\Big(\hat{\beta}(t) - \beta(t)\Big) \\ &E\Big(\frac{g'(\eta)^۲}{\sigma^۲(\mu)} X(s)X(t)\Big) d\omega(s) d\omega(t), \end{aligned}$$

در نظر گرفته می‌شود که برای هر p با

$$d_{G,p}^۲(\hat{\beta},\beta) = (\hat{\beta} - \beta)' \Gamma (\hat{\beta} - \beta),$$

تقریب زده می‌شود. دقت کنید که اگر روش نیمه‌پارامتری $SPQR$ به کار می‌رود، چون در این مدل توابع واریانس و پیوند نامعلوم هستند و به‌طور ناپارامتری برآورد می‌شوند، باید شرایطی در نظر گرفته شود که مهم‌ترین آن‌ها عبارت است از این‌که باید تابع پیوند g یکنواخت و معکوس‌پذیر مشتق‌پذیر باشد و برای هر $c \geqslant ۰$، $\|g'(.)\| \leqslant c,\ \|g''(.)\| \leqslant c$. تابع واریانس $\sigma^۲(.)$ به‌طور پیوسته مشتق‌پذیر است و $\delta > ۰$ وجود دارد، به طوری که $\sigma(۰) > \delta$. همچنین در دنباله‌ای از مدل‌های تقریبی بریده‌شده p_n، با تعداد

[۷] Compact Self-adjoint Hilbert-schmidt

جملات پیشگویی p_n، اگر $p_n \longrightarrow \infty$ و $n \longrightarrow \infty$ آنگاه $p_n n^{-\frac{1}{2}} \longrightarrow 0$.
(برای جزئیات شرایط مدل $SPQR$ به مولر و استدمولر، [۱۱] مراجعه شود).
اکنون با در نظر گرفتن فرض‌های اصلی و شرط‌های مدل $SPQR$، می‌توان نشان داد، اگر $n \longrightarrow \infty$ آنگاه

$$\frac{n(\hat{\beta}-\beta)'\Gamma_{p_n}(\hat{\beta}-\beta)-(p_n+1)}{\sqrt{2(p_n+1)}} \xrightarrow{d} N(0,1), \qquad (14)$$

که مولر و استدمولر، [۱۱] با در نظر گرفتن $p_n = p$،
$\tilde{\Gamma} = \frac{1}{n}E(D'D)$ را جایگزین Γ_{p_n} نمودند. در صورتی که فقط پارامترهای شیب $\beta_1, \beta_2, \ldots$ (بدون عرض از مبدأ) موردنظر هستند p_n با $p_n - 1$ و به جای ماتریس $(p+1)\times(p+1)$، Γ از ماتریس $p \times p$، Γ استفاده می‌شود که از حذف اولین سطر و ستون به‌دست‌آمده است. برای همگرایی تابع پارامتر برآورد شده $\hat{\beta}(\cdot)$، فاصله d_G و بسط به صورت

$$\hat{\beta}(t) = \sum \hat{\beta}_{\rho_j^G} \rho_j^G(t)$$

مربوط به تابع پارامتر $\hat{\beta}(\cdot)$ در پایه $\{\rho_j^G,\ j = 1,2,\ldots\}$، باید پایه ویژه عملگر A_G متناسب با مقادیر ویژه λ_j^G به کار گرفته شود. بنابراین

$$\begin{aligned} d_G^2\Big(\hat{\beta}(.),\beta(.)\Big) &= \int\int \Big(\hat{\beta}(s)-\beta(s)\Big)G(s,t) \\ &\quad \Big(\hat{\beta}(t)-\beta(t)\Big)d\omega(s)d\omega(t) \\ &= \sum_{j=1}^{p} \lambda_j^G (\hat{\beta}_{\rho_j^G} - \beta_{\rho_j^G})^2 + \sum_{j=p+1}^{\infty} \lambda_j^G \beta_{\rho_j^G}^2 \\ &= \Big(\hat{\beta}^G - \beta^G\Big)'\Gamma^G\Big(\hat{\beta}^G - \beta^G\Big) + \sum_{j=p+1}^{\infty} \lambda_j^G \beta_{\rho_j^G}^2. \end{aligned}$$

که در آن $\hat{\beta}^G = \Big(\hat{\beta}_{\rho_1^G}, \cdots, \hat{\beta}_{\rho_p^G}\Big)'$، $\beta^G = \Big(\beta_{\rho_1^G}, \cdots, \beta_{\rho_p^G}\Big)'$ و Γ^G ماتریس قطری است. ε_j^G به صورت

$$\varepsilon_j^G = \frac{g'(\eta)}{\sigma(\mu)} \int X(t)\rho_j^G(t)d\omega(t),$$

مشخص می‌شود، به طوری که

$$\begin{aligned} E(\varepsilon_j^G \varepsilon_k^G) &= \int\int G(s,t)\rho_j^G(s)\rho_k^G(t)d\omega(s)d\omega(t) \\ &= \delta_{ij}\lambda_j^G. \end{aligned} \qquad (15)$$

نتیجه ۱.۳. اگر تابع پارامتری $\beta(\cdot)$ ویژگی

$$\sum_{j=p+1}^{\infty} E(\varepsilon_j^{G^{(2)}})\Big[\int \beta(t)\rho_j^G(t)d\omega(t)\Big]^2 = o\Big(\frac{\sqrt{p_n}}{n}\Big), \qquad (16)$$

را داشته باشد، آنگاه وقتی $n \longrightarrow \infty$

$$\frac{n\int\int \Big(\hat{\beta}(s)-\beta(s)\Big)\Big(\hat{\beta}(t)-\beta(t)\Big)G(s,t)d\omega(s)d\omega(t) - (p_n+1)}{\sqrt{2p_n+1}} \xrightarrow{d} N(0,1)$$

(برای اثبات به مولر و استدمولر، [۱۱] مراجعه شود)

در حالتی که توابع پیوند و واریانس نامعلوم هستند در رهیافت $SPQR$، برای به دست آوردن برآورد ناپارامتری از توابع و مشتق‌ها، هموارسازی نمودارهای پراکنش اعمال می‌شود و سپس برآورد پارامتری $\hat{\beta}$ از حل معادله امتیاز نیمه پارامتری (۸) به دست می‌آید. بعد از چند تکرار برآورد ناپارامتری تابع پیوند $\hat{g}$ و مشتق اولش $\hat{g}'$ و تابع واریانس $\hat{\sigma}^2$ به دست می‌آیند. این برآوردگرهای منحنی ناپارامتری توابع با هموارکننده‌های هسته درجه دوم یا خطی موضعی با یک پهنای باند h در هموارسازی به دست می‌آیند. مولر و استدمولر، [۱۱] نشان دادند تحت شرایطی توزیع تقریبی (۱۴) برای رهیافت $SPQR$ نیز برقرار است.

یک مسئله رایج در مدل‌های رگرسیونی به این صورت است که

$$H_0 : \beta = \text{ثابت},$$

که حالت خاصی از آزمون $H_0 : \beta \equiv \beta_0$ برای یک تابع پارامتری مشخص β_0 است. با در نظر گرفتن $\beta_0(t) = \sum \beta_{0j}\rho_j(t)$، فرض $H_0 : \beta_j = \beta_{0j}$، $j = 0,1,2,\ldots$ می‌شود و H_0 رد می‌شود وقتی آماره‌ی آزمون بیان‌شده در (۱۴) بیشتر از مقدار بحرانی $\Phi(1-\alpha)$ باشد.
در عمل بازه‌های اطمینان برای تابع پارامتر رگرسیونی β نیز موردنظر است. در حالتی که $p = p_n$ معلوم است و برآوردهای $\hat{\beta}$ برای بردار β مشخص‌شده‌اند یک ناحیه اطمینان مجانبی $(\alpha-1)$ مطابق با (۱۴) با $(\hat{\beta}-\beta)'\Gamma(\hat{\beta}-\beta) \leqslant C(\alpha)$ مشخص می‌شود که در آن

$$c(\alpha) = \Big[p+1+\sqrt{2(p+1)}\Phi(1-\alpha)\Big]/n,$$

و Γ معمولاً با $\tilde{\Gamma}$ یا $\hat{\Gamma}$ جایگزین می‌شود. اکنون فرض کنید $(e_1,\lambda_1),\ldots,(e_{p+1},\lambda_{p+1})$ بردارهای ویژه و مقادیر ویژه ماتریس Γ باشند، و قرار دهید

$$\omega_k(t) = \sum_{l=1}^{p+1} \rho_l(t)e_{kl},\ k = 1,\cdots,p+1,\ e_k = (e_{k_1},\cdots,e_{k,p+1})',$$

آنگاه برای n بزرگ و p یک بازه اطمینان تقریبی هم‌زمان $(\alpha-1)$ برای تابع $\beta(t)$ به صورت

$$\hat{\beta}(t) = \pm\sqrt{c(\alpha)\sum_{k=1}^{p+1}\frac{\omega_k(t)^2}{\lambda_k}} \qquad (17)$$

است. در عمل بازه اطمینان هم‌زمان با جایگزین کردن برآوردهای ω_k و λ_k که از ماتریس $\tilde{\Gamma}$ یا $\hat{\Gamma}$ به جای Γ به‌دست‌آمده‌اند، محاسبه می‌شود.

۲.۳ برآورد درستنمایی تاوانیده

برای برآورد درستنمایی تاوانیده معمولاً تابع تاوان[8] که توسط مایر و همکاران [۱۰] ارائه‌شده است به صورت

$$P_{\lambda,\varphi}(\beta)=\lambda(\|\beta\|^{۲}+\varphi\|\beta''\|^{۲})^{\frac{۱}{۲}}$$

در نظر گرفته می‌شود، که در آن $\|\beta\|^{۲}=\int_{D_i}\{\beta(t)\}^{۲}dt$ نرم $L^{۲}$ از β و $\beta''(t)=\frac{\partial^{۲}\beta(t)}{\partial t^{۲}}$ است. φ, λ, پارامترهای تاوان غیرمنفی هستند که به ترتیب تنکی و هموواری را کنترل می‌کنند. برای به دست آوردن برآورد ماکسیمم درستنمایی تاوانیده تابع $\beta(\cdot)$ باید عبارت

$$\frac{۱}{n}\sum_{i=۱}^{n}\ell(y_i,\beta(t)X_i(t))-\sum_{j=۱}^{p}P_{\lambda,\varphi}(\beta(t))$$

مینیمم شود به طوری که با در نظر گرفتن

$$\beta(t)=\sum_{j=۱}^{p}\beta_j\rho_j(t)$$

تاوان به صورت

$$P_{\lambda,\varphi}(\beta)=\lambda(\rho'(\Psi+\varphi\Omega)\rho)^{\frac{۱}{۲}}$$

می‌شود، که در آن Ψ ماتریس $p\times p$ که (j,k) امین عنصر آن

$$\Psi_{j,k}=\int\beta_j(t)\beta_k(t)dt,\quad j,k=۱,...,p$$

و Ω یک ماتریس $p\times p$ که (j,k) امین عنصر آن

$$\Omega_{j,k}=\int\beta_j''(t)\beta_k''(t)dt,\quad j,k=۱,...,p$$

هستند. در صورتی که تاوان به صورت

$$P_{\lambda,\varphi}(\beta)=\lambda(\rho'k_\varphi\rho)^{\frac{۱}{۲}}$$

در نظر گرفته شود نوعی تاوان لاسو[9] است که در آن $\Psi+\varphi\Omega=k_\varphi$ یک ماتریس $p\times p$ متقارن معین مثبت است. در صورتی که $k_\varphi=I$ تاوان لاسو معمولی است.

۴ انتخاب رتبه مدل

همان‌طور که در قبل اشاره شد یکی از مطالب مهم در مدل بریده‌شده خطی تابعی تعمیم‌یافته انتخاب مقدار مناسب برای p یا تعداد توابع ویژه‌ای که برای برازش تابع $\beta(t)$ به کار می‌روند، به عبارت دیگر انتخاب رتبه برای مدل است. یک رهیافت برای این انتخاب استفاده از معیار آکائیکه[10] AIC است که تمایل به افزایش رتبه مدل دارد. فرض کنید بردار پیشگوی خطی η_p شامل n مؤلفه $\eta_{p,i}=\sum_{j=۰}^{p}\varepsilon_j^i\beta_j$، $i=۱,...,n$ باشد و بردار $\hat\eta_p$ شامل مؤلفه‌های $\hat\eta_{p,i}=\sum_{j=۰}^{p}\varepsilon_j^i\hat\beta_j$ و بردار η شامل مؤلفه‌های $\sum_{j=۰}^{\infty}\varepsilon_j^i\beta_j$ است. G را مشتق تابع معکوس پیوند g در نظر بگیرید بنابراین Y دارای چگالی

$$f_Y(y)=\exp(y\eta+a(y)-G(\eta))$$

است. در عمل $\tilde\sigma^{۲}(\eta)=g'(\eta)$ می‌باشد. انحراف[11] به صورت

$$\mathscr{D}=-۲\ell_n(Y,\hat\eta_p)+۲\ell_n\Big(Y,g^{-۱}(Y)\Big),$$

تعریف می‌شود، که در آن لگاریتم شبه درستنمایی

$$\ell_n(Y,\hat\eta_p)=\sum_{i=۱}^{n}Y_i\hat\eta_{i,p}-\sum_{i=۱}^{n}G(\hat\eta_{i,p}).$$

است. با استفاده از بسط تیلور

$$\begin{aligned}-۲\ell_n(Y,\hat\eta_p)=&-۲\ell_n(Y,\eta_p)\\&+۲\Big(\nabla_{\beta_p}L_n(Y,\hat\eta_p)\Big)'(\beta_p-\hat\beta_p)\\&+(\beta_p-\hat\beta_p)'\Big(\frac{\partial^{۲}}{\partial\beta_k\partial\beta_l}\ell_n(Y,\tilde\eta_p)\Big)(\beta_p-\hat\beta_p)\end{aligned}$$

که در آن با توجه به معادله امتیاز، عبارت دوم سمت راست صفر و ماتریس به صورت فرم درجه دوم $(D'D)$ است. طبق مولر و استدمولر [۱۱]. مقدار امید ریاضی فرم درجه دوم

$$n(\beta_p-\hat\beta_p)'(\frac{D'D}{n})(\beta_p-\hat\beta_p),$$

به‌طور تقریبی مقدار p است. چون

$$\begin{aligned}-۲\ell_n(Y,\eta_p)=&-۲\ell_n(Y,\eta)-۲\sum_{i=۱}^{n}\Big(Y_i-g(\eta_i)\Big)(\eta_{i,p}-\eta_i)\\&+\sum_{i=۱}^{n}g'(\eta_i)(\eta_{i,p}-\eta_i)^{۲}\end{aligned}\qquad(۱۸)$$

در نتیجه

$$\begin{aligned}E(\mathscr{D})&=n\sum_{k,l=P+۱}E\Big(g'(\eta)\varepsilon_k\varepsilon_l\Big)\beta_k\beta_l-p\big(۱+o(۱)\big)+E_n\\&=n\sum_{k,l=P+۱}E\Big(\frac{g^{۲'}(\eta)}{\tilde\sigma^{۲}(\eta)}\varepsilon_k\varepsilon_l\Big)\beta_k\beta_l-p\big(۱+o(۱)\big)+E_n,\end{aligned}$$

که در آن عبارت E_n به p وابسته نیست. مولر و استدمولر [۱۱] با به‌کارگیری قانون اعداد بزرگ نشان دادند که اگر p در بازه $(p_۰,cn^{\frac{۱}{۲}})$ قرار بگیرد آنگاه

$$\mathscr{D}/E(\mathscr{D})\xrightarrow{P}۱.$$

[8] Penalty Function
[9] Lasso
[10] Akaike Information Criterion
[11] Deviance

همچنین آن‌ها نشان دادند

$$d(\hat{\beta}(\cdot),\beta(\cdot)) = \int\int \Big(\hat{\beta}(s)-\beta(s)\Big)G(s,t)\Big(\hat{\beta}(t)-\beta(t)\Big)d\omega(s)d\omega(t)$$
$$= (\hat{\beta}_p-\beta_p)'\Gamma(\hat{\beta}_p-\beta_p)+\sum_{k,j=p+1}^{\infty}\gamma_{j,k}\beta_j\beta_k$$
$$+2\sum_{j=1}^{p}\sum_{k=p+1}^{\infty}\gamma_{j,k}(\hat{\beta}_j-\beta_j)\beta_k$$

که در آن

$$\gamma_{k,l} = E\Big(\frac{g^{2'}(\eta)}{\tilde{\sigma}^2(\eta)}\varepsilon_k\varepsilon_l\Big).$$

اکنون

$$E\Big(d(\hat{\beta}(.),\beta(.))\Big) = p/n\big(1+o(1)\big)+\sum_{k,j=p+1}^{\infty}\gamma_{j,k}\beta_j\beta_k\big(1+o(1)\big).$$

این استنباط نشان می‌دهد که تابع هدف $d(\hat{\beta}(.),\beta(.))$ با کمینه ساختن به‌طور مجانبی به $E(\mathscr{D}/n)+2p/n$ نزدیک می‌شود. در صورتی که دقت تابع هدف و AIC مطرح باشد، اثبات می‌شود که رتبه p_A با AIC و p_d با مینیم کردن تابع هدف انتخاب می‌شوند. آنگاه اگر $n\longrightarrow\infty$ ،$p_d/p_A\longrightarrow 1$. در عمل AIC و ملاک اطلاع بیزی BIC[12] را با به دست آوردن اولین انحراف یا شبه‌انحراف $\mathscr{D}(p)$، وابسته به مرتبه مدل p بکار می‌بریم. که در شبه درستنمایی یا ماکسیمم درستنمایی اگر با تابع پیوند معلوم باشد به راحتی امکان‌پذیر است و در صورتی که نامعلوم باشد، لگاریتم درستنمایی و رهیافت $SPQR$ با تابع پیوند نامعلوم نیازمند انتگرال‌گیری است. درنهایت عبارت

$$C(p) = \mathscr{D}(p)+\mathscr{P}(p) \qquad (19)$$

کمینه می‌شود که در آن $\mathscr{P}$ جریمه است، به طوری که $2p$ برای AIC و $p\log n$ برای BIC انتخاب می‌شود. روش‌های دیگری برای تعیین p وجود دارد که می‌توان به استفاده از اعتبارسنجی متقابل رایس و سیلورمن، [۱۴] و اختلاف نسبی بین معیار پیرسون و انحراف چیو و مولر، [۴] اشاره نمود.

۵ بررسی چند مثال

در این بخش یک مطالعه شبیه‌سازی و دو مثال کاربردی موردبررسی قرارگرفته است.

۱.۵ شبیه‌سازی

در یک مدل خطی تابعی تعمیم‌یافته علاوه بر انتخاب p، تعیین یک پایه یکامتعامد $\{\rho_j, 1, 2, ...\}$ مهم است. معمولاً دو انتخاب داریم، یا از یک پایه استاندارد ثابت مانند پایه فوریه $\rho_j\equiv\varphi_j\equiv\sqrt{2}\sin(\pi jt)$، $j\geqslant 1$ و $t\in[0,1]$ استفاده می‌شود، یا توابع ویژه عملگر کوواریانس A_K که در بخش ۴ در روابط (۱۱) و (۱۰) دیدیم. در اینجا از الگوریتم دوم استفاده می‌شود و الگوریتمی که برای برآورد توابع ویژه در کاپرا و مولر [۱]؛ رایس و سیلورمن [۱۴] ارائه‌شده، به کار می‌رود. p مؤلفه مدل تعیین می‌شود و بنابراین i امین واحد مشاهده‌شده فرآیند به p پیشگویی‌کننده کاهش می‌یابد

$$\varepsilon_j^{(i)} = \int X_i(t)\rho_j(t)d\omega(t), \quad j=1,...,p.$$

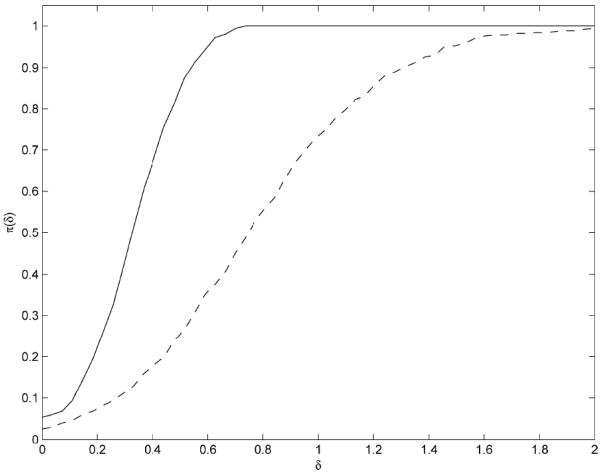
توابع ویژه برآورد شده به جای ρ_j جایگزین و انتگرال‌ها به‌طور عددی محاسبه می‌شوند. پارامترهای α و $\beta_1,...,\beta_p$ در مدل تابعی تعمیم‌یافته با حل معادله امتیاز برآورد می‌شوند. الگوریتم کمترین مربعات بحث شده در مک‌کلا و نلدر [۹] برای حالتی از یک مدل خطی تعمیم‌یافته یا شبه درستنمایی با تابع پیوند معلوم و الگوریتم $QLUE$[13] مطرح‌شده در چیو و مولر [۵] برای مدل $SPQR$ با تابع پیوند نامعلوم تکرار می‌کنیم. هدف از ارائه این مثال شبیه‌سازی مونت‌کارلویی مقایسه AIC و BIC برای تعیین رتبه p است و بررسی توان آزمون‌های آماری اثر رگرسیون در مدل رگرسیون تابعی تعمیم‌یافته و همچنین بررسی رفتار فرآیند نیمه پارامتری $SPQR$ برای رگرسیون تابعی، در مقایسه با پیاده‌سازی شبه درستنمایی و ماکزیمم با تابع پیوند مشخص است. شبیه‌سازی به این صورت است که فرآیندهای شبه‌تصادفی مبنی بر ۲۰ تابع اول پایه فوریه $X(t)=\sum_{j=1}^{20}\varepsilon_j\varphi_j(t)$ با استفاده از متغیرهای شبه‌تصادفی نرمال $\varepsilon_j\sim N(0,1/j^2)$، $j\geqslant 1$ تولیدشده‌اند. فرض کنید $\beta_j=1/j$، $1\leqslant j\leqslant 3$، $\beta_0=1$، $\beta_j=0$ و $j>3$. فرض کنید

$$\beta(t) = \sum_{j=1}^{20}\beta_j\varphi(t)$$
$$P(X(\cdot)) = g(\beta_0+\sum_{j=1}^{20}\beta_j\varepsilon_j)$$

دو پیوند، پیوند لوجیت به طوری که $[g(x)=\exp(x)/(1+\exp(x))]$ و پیوند لگ-لگ مکمل $(c-loglog)$، $[g(x)=\exp(-\exp(-x))]$ در نظر گرفته می‌شود. آنگاه $Y|X\sim Binomial(P(X),1)$ به صورت یک متغیر تصادفی شبه‌برنولی با احتمال $P(X)$ و نمونه $(X_i(t),Y_i)$، $i=1,...,n$ تولید می‌شود. برای نشان دادن نتایج تقریبی که در (۱۴) دیدیم، توابع توان تجربی برای داده‌های تولیدشده و با به کار بردن تابع پیوند لوجیت و آماره آزمون T که در سمت چپ رابطه (۱۵) برای آزمون کردن فرضیه $H_0:\beta_j=0,\ j=1,2,...$ یعنی عدم وجود اثر رگرسیونی دیدیم، محاسبه می‌شوند. این آزمون یک آزمون یک‌طرفه در سطح ۵٪ است، به طوری که اگر $|T|>\Phi^{-1}(0.95)$ فرض H_0 رد می‌شود. با تولید ۵۰۰ نمونه مونت‌کارلو نرخ رد برای حجم نمونه $n=50,200$، به صورت تابع δ، $0\leqslant\delta\leqslant 2$ تعیین شد که بردار پارامتر شرح داده‌شده در پاراگراف قبلی، در δ ضرب شده است و به وسیله‌ی $(\delta,\delta,\delta/2,\delta/3)$ تعیین می‌شود. تابع توان برای دو حجم نمونه ۵۰،۲۰۰ در شکل ۱ رسم شده است. این شکل نشان می‌دهد که اندازه نمونه نقش مهمی توان آزمون دارد.

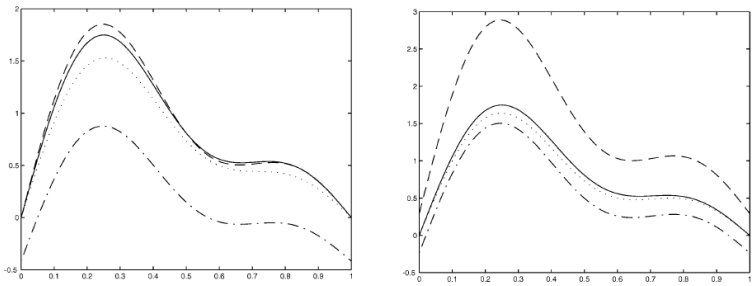
[12] Bayesian Information Criterion

[13] Quasi Likelihood with Unknown linke and variance function Estimation



شکل ۱: تابع توان آزمون معناداری رگرسیون لوژستیک تابعی در سطح ۵٪. بر اساس ۵۰۰ شبیه‌سازی، برای نمونه‌ای به‌اندازه ۵۰ (خط‌چین) و ۲۰۰(خط ممتد)، با $p = ۳$

.



شکل ۲: شکل سمت راست داده‌ها از پیوند لگ-لگ مکمل و شکل چپ از تابع پیوند لوجیت تولیدشده است. خط ممتد در هر دو نمودار خط هدف است و برآوردها با فرض پیوند لوجیت با (خط‌چین)، پیوند با (خط‌چین-نقطه) و روش لگ-لگ مکمل *SPQR* با (نقطه‌چین) مشخص‌شده است.

برای این‌که خوب بودن روش نیمه‌پارامتری *SPQR* بررسی شود، میانگین برآورد تابع پارامتر $\hat{\beta}(.)$ با ۵۰ بار اجرای مونت‌کارلو در حالات زیر

۱. استفاده از تابع پیوند لوجیت

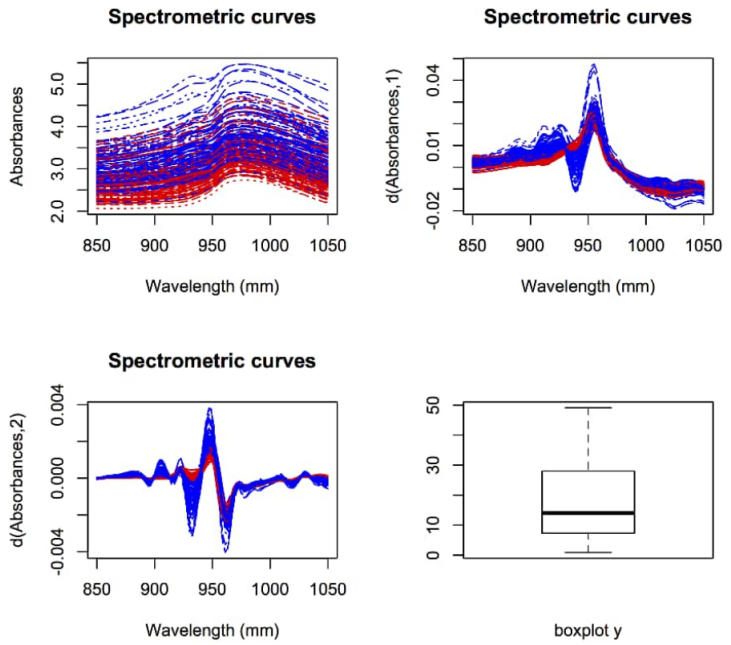
۲. استفاده از تابع پیوند لگ-لگ مکمل

۳. رهیافت نیمه پارامتری *SPQR* با پیوند نامعلوم

محاسبه شد. با توجه به شکل ۲ درمی‌یابیم که روش *SPQR* یک رهیافت مفید است. بنابراین نتیجه می‌شود که انتخاب تابع پیوند یک مسئله مهم در این مدل‌ها می‌باشد و انتخاب نادرست برای آن می‌تواند باعث بروز مشکلاتی در استنباط‌ها شود.

۲.۵ مثال کاربردی ۱

فرض کنید در یک بررسی متغیر پاسخ چربی گوشت است که برای چربی کمتر از ۱۵ عدد صفر و برای چربی بیشتر از ۱۵ عدد یک در نظر گرفته‌شده است و متغیر پاسخ به صورت صفر و یک است. میزان جذب اندازه‌گیری شده در ۱۰۰ طول موج بین ۸۵۰ تا ۱۰۵۰ میلی‌متر به عنوان متغیر کمکی تابعی فرض می‌شود. این مجموعه داده بخشی از مجموعه داده اصلی است که در http://stat.cmu.edu در دسترس است. نمودار متغیر تابعی میزان جذب و مشتق اول و دوم و نمودار جعبه‌ای متغیر پاسخ قبل از رسته‌ای کردن این متغیر در شکل ۳ مشاهده می‌شود.



شکل ۳: رنگ آبی مربوط به چربی کمتر از ۱۵ و رنگ قرمز مربوط به چربی بیشتر از ۱۵ است.

برای تحلیل داده‌ها با استفاده از مدل خطی تابعی تعمیم‌یافته از بسته نرم‌افزاری fda.usc و تابع fregre.glm در نرم‌افزار R استفاده شد. با توجه به اینکه متغیر پاسخ دودویی است می‌توان از توابع پیوند لوجیت[14] و پروبیت[15] و لگ-لگ مکمل (cloglog) استفاده کرد که می‌دانیم برای پیوند لوجیت

$$P_i = P(Y=1 \mid X_i(t), t\in\tau) = \frac{\exp(\alpha+\int\beta(t)X_i(t)dt)}{1+\exp(\alpha+\int\beta(t)X_i(t)dt)}$$

برای پیوند پروبیت

$$P_i = P(Y=1 \mid X_i(t), t\in\tau) = \Phi(\alpha+\int\beta(t)X_i(t)dt)$$

که در آن Φ تابع توزیع تجمعی نرمال استاندارد است استفاده می‌شود. برای مدل لگ-لگ مکمل تابع پیوند به صورت

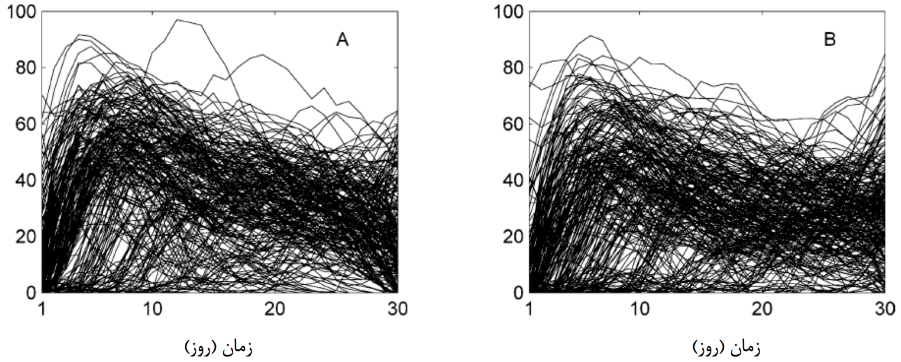
$$\begin{aligned} P_i &= P(Y=1 \mid X_i(t), t\in\tau) \\ &= 1-\exp(-\exp(\alpha+\int\beta(t)X_i(t)dt)) \end{aligned}$$

است. مقدار AIC سه مدل لوجیت، پروبیت و لگ-لگ مکمل به ترتیب ۲۷٫۶۲۷، ۲۷٫۱۹۴ و ۲۷٫۰۳۲ به دست آمد که مدل لگ-لگ مکمل با توجه به AIC کمتر مدل مناسب است. همچنین نرخ طبقه‌بندی نادرست برای مدل لوجیت، پروبیت، لگ-لگ مکمل و $SPQR$ به ترتیب ۳٫۵٪، ۳٫۵٪، ۷.۳٪ و ۳٫۸٪ به دست آمد که نشان می‌دهد که نتایج مدل لگ-لگ مکمل به رهیافت $SPQR$ نزدیک‌تر است و برای این داده‌ها از دقت بیشتری برخوردار است.

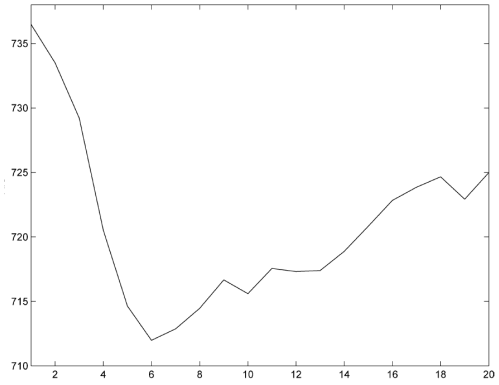
۳.۵ مثال کاربردی ۲

از هزار مگس مدیترانه‌ای مطرح‌شده در کری و همکاران [۳]، مگس‌هایی که ۳۴ روز بعد زنده مانده‌اند را انتخاب می‌کنیم و نمونه‌ای از ۵۳۴ مگس مدیترانه‌ای فراهم و برای پیش‌بینی، خط مسیر تخم‌گذاری از ۰ تا ۳۰ روز برای فرآیندهای پیشگوی $X_i(t)$، $t\in[0,30]$ $i=1,\ldots,534$ استفاده می‌شود. اگر عمر مگس در ۳۰ روز گذشته ۱۴ روز یا بیشتر باشد در طبقه‌بندی طول عمر بلندمدت قرار می‌گیرد، در غیر این صورت طول عمر کوتاه دارد. از $n=534$ مگس، ۲۵۶ مگس طول عمر کوتاه و ۲۷۸ مگس دارای طول عمر بلند هستند. با انتخاب پیوند لوجیت، یک رگرسیون تابعی لوجیت به داده‌ها برازش می‌شود. خط سیر تولیدمثل برای مگس‌های با طول عمر بلند و عمر کوتاه به صورت جداگانه در شکل ۴ ترسیم‌شده است. تفاوت بین دو گروه به دلیل ازدحام منحنی‌های رسم شده به صورت چشمی قابل‌تشخیص نیست. برای انتخاب تعداد مؤلفه‌های مدل ملاک AIC به کار می‌رود. همان‌طور که در شکل ۵ مشاهده می‌شود، ملاک AIC به مرتبه مدل p وابسته است و مقدار مینیمم در ۶ است. بنابراین $p=6$ انتخاب می‌شود. برای انتخاب مقدار مناسب $p=6$ از معیار خطای پیشگویی اعتبارسنجی متقابل $PE=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{p}_i^{(-i)}\right)^2$ برآورد $\hat{p}_i^{(-i)}$ که در آن p_i با خروج iامین مشاهده

[14] Logit
[15] Probit



شکل ۴: پیش‌بینی خط سیر، مطابق با منحنی‌های تخم‌گذاری روزانه، برای $n = ۵۳۴$ مگس. (چپ) تولیدمثل ۲۵۶ مگس با طول عمر کوتاه و (راست) ۲۷۸ مگس با طول عمر بلند را نمایش می‌دهد.



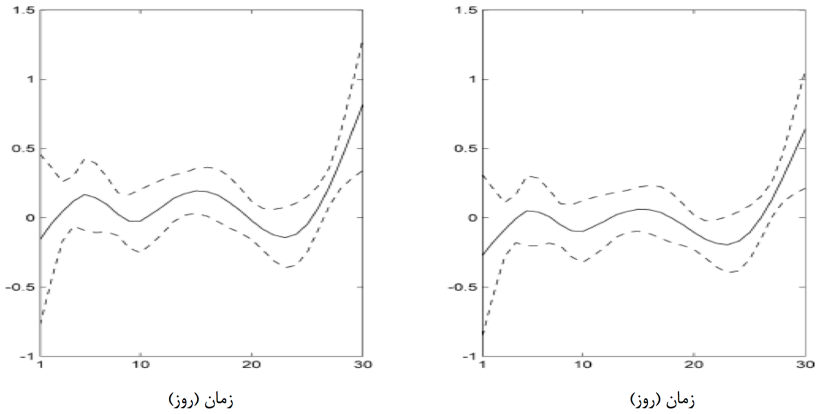
شکل ۵: معیار اطلاعات آکائیکه (AIC) به عنوان تابعی از مقادیر مؤلفه‌های مدل p برای داده مگس مدیترانه‌ای

.

است، استفاده شد. از این معیار هم $p = ۶$ به دست آمد. برآورد نرخ طبقه‌بندی نادرست برای گروه مگس‌های طول عمر بلند، ٪۳۷ با پیوند لوجیت و ٪۳۵ برای پیوند $SPQR$ است. درحالی‌که برای گروه مگس‌های کوتاه عمر ٪۴۷ برای لوجیت و ٪۴۸ برای $SPQR$، است. بنابراین طبقه‌بندی صحیح مگس‌های کوتاه عمر را به سختی نشان می‌دهد. توابع پارامتر $\hat{\beta}(\cdot)$ برای رگرسیون تابعی لوژستیک و $SPQR$ با بازه‌های اطمینان هم‌زمان در شکل ۶ نشان داده‌شده است؛ مشاهده می‌شود که برآورد با پیوند ناپارامتری کاملاً نزدیک به برآورد با پیوند لوژستیک است، بنابراین در این حالت پیوند لوژستیک را انتخاب می‌کنیم. از بازه‌های اطمینان مجانبی نتیجه می‌گیریم تابع پیوند شیب صعودی را در سمت راست برای عمر ۳۰ روزه دارد، و شکل تابع پارامتر $\hat{\beta}(\cdot)$ مشخص می‌کند که دوره‌های تناوب تخم‌گذاری با افزایش طول عمر مرتبط است. تحت تابع پیوند لوجیت، می‌دانیم احتمال طبقه‌بندی مگس با طول عمر بلند $g(\eta) = \exp(\eta)/(۱ + \exp(\eta))$ است. برای برآورد تابع پیوند ناپارامتری از رهیافت $SPQR$ استفاده شد، که هموارسازی خطی موضعی و پهنای باند ۰٫۵۵ برای مراحل هموارسازی انتخاب شد. برای دو تابع پیوند، بزرگ‌ترین پیشگوی خطی η، سپس بزرگ‌ترین مقادیر تابع پارامتر $\beta(\cdot)$، با احتمال افزایش طول عمر مرتبط هستند. چون برآورد تابع پارامتر بین ۱۷ − ۱۲ روز و پس از ۲۶ روز نسبتاً بزرگ است، نتیجه می‌گیریم که در طول این دوره‌ها فعالیت باروری سنگین با افزایش طول عمر در ارتباط است. در مقابل، افزایش تولیدمثل بین ۱۲ − ۸ روز و ۲۶ − ۲۰ روز با کاهش طول عمر در ارتباط است.

بحث و نتیجه‌گیری

یک مدل خطی تابعی تعمیم‌یافته بریده‌شده برای بررسی ارتباط بین یک متغیر پاسخ اسکالر گسسته و متغیرهای کمکی تابعی با رهیافت شبه درستنمایی موردبررسی قرار گرفت. برای این مدل آزمون فرضیه و بازه اطمینان مجانبی برای پارامترهای رگرسیونی تابعی به دست آورده شد. مدل نیمه‌پارامتری $SPQR$ برای حالتی که تابع پیوند نامعلوم است بررسی و درنهایت مدل و روش‌ها بر روی چند مثال پیاده‌سازی شد. همان‌طور که در مثال‌ها مشاهده شد



شکل ۶: برآورد پارامتر تابعی $\hat{\beta}(\cdot)$ (توپر) و بازه‌های اطمینان هم‌زمان، (چپ) پیوند لوجیت، (راست) رهیافت ناپارامتری *SPQR*.

معمولاً رهیافت *SPQR* برای تحلیل مدل خطی تابعی تعمیم‌یافته مناسب است. رهیافت ناپارامتری رهیافت *SPQR* به کار گرفته شود و درنهایت بهترین مدل با توجه به داده‌ها می‌توان توابع پیوند مختلف در نظر گرفته شود و برای اطمینان انتخاب شود. .

مراجع

[1] Capra, W. B. and Muller, H. G. (1997). An accelerated time model for response curves. *Journal of the American Statistical,* **92**, 72–83.

[2] Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters,* **45**, 11–22.

[3] Carey, J. R., Liedo, P., Muller, H. G., Wang, J. L. and Chiou, J. M. (1998). Relationship of age patterns of fecundity to mortality, longevity and lifetime reproduction in a large cohort of Mediterranean fruit fly females. *Journal of the Gerontology: Biological Sciences,* **53A**, B245–B251.

[4] Chiou, J. M. and Muller, H. G. (1998). Quasi-likelihood regression with unknown link and variance functions. *Journal of the American Statistical Association,* **93**, 1376–1387.

[5] Chiou, J. M. and Muller, H. G. (1999). Nonparametric quasi-likelihood. *Journal of the American Statistical,* **27**, 36–64.

[6] Conway, J. B. (1990). *A Course in Functional Analysis, 2nd ed*. Springer, New York.

[7] Dunford, N. and Schwartz, J. T. (1963). Linear operators. *II. Spectral Theory*. Wiley, New York.

[8] Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics,* **39**, 254–261.

[9] McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models, 2nd ed*. Chapman and Hall, London.

[10] Meier, L, Van de Geer, S & B¨uhlmann, P (2009). High-dimensional additive modeling, *The Annals of Statistics*, **37**, 3779–3821.

[11] Muller, H. G. and Stadmuller, U. (2005). Generalized functional linear models, *The Annals of Statistics*, **33**, 774–805.

[12] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models, *Journal of Royal Statistical Society,* **135**, 370–384.

[13] Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Springer, New York.

[14] Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society,* **53**, 233–243.

[15] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss–Newton method. *Biometrika,* **61**, 439–447.