

کاربردی از ارزیابی خطای اندازه‌گیری به‌روش تحلیل رده نهن

حمیدرضا نواب‌پور^۱، اکرم صفرنژاد بروجنی^۲ و طیبه چگینی^۳

تاریخ دریافت: ۱۳۹۴/۷/۳

تاریخ پذیرش: ۱۳۹۶/۱/۳۰

چکیده:

تحلیل رده نهن (LCA) روشی برای ارزیابی خطاهای غیر نمونه‌گیری، به‌خصوص خطای اندازه‌گیری داده‌های رسته‌ای است. [۱]، چهار رهیافت مدل‌بندی رده نهن، یعنی پارامتری‌سازی مدل احتمالاتی، مدل لگ خطی، مدل مسیر تعدیل‌یافته و مدل نموداری را با استفاده از نمودارهای مسیر معرفی کرده است. این مدل‌ها قابل تبدیل به یکدیگرند. مدل‌های احتمالاتی رده نهن، درست‌نمایی جدول رده‌بندی تقاطعی متغیرها را بر حسب احتمال‌های شرطی و حاشیه‌ای مربوط به هر خانه این جدول بیان می‌کند. در این رهیافت پارامترهای مدل با استفاده از الگوریتم EM برآورد می‌شوند. برای آزمون مدل رده نهن، آماره‌خی دو به‌عنوان ملاک نیکویی برازش معرفی شده است. در این مقاله از LCA و داده‌های یک آمارگیری کوچک مقیاس برای محاسبه خطای بدرده‌بندی (که یک نوع خطای اندازه‌گیری است) نسبت دانشجویانی که دست کم در یک درس مردود شده‌اند و نیز خطای بدرده‌بندی نسبت دانشجویانی که دست کم یک‌بار مشروط شده‌اند، استفاده شده است.

واژه‌های کلیدی: خطای کل آمارگیری، خطای اندازه‌گیری، مدل احتمالاتی، تحلیل رده نهن، استاندارد طلایی، خطای بدرده‌بندی.

۱ مقدمه

را ارزیابی کند؛ اما در این مقاله توجه ما بر اندازه‌گیری‌های رسته‌ای و خطاهای رده‌بندی، یعنی بدرده‌بندی کردن یک رسته واقعی به یکی از رسته‌های ممکن آن متغیر (خطای اندازه‌گیری) است. ایده کلی مدل‌های رده نهن به کارهای اولیه لازارسفلد [۹ و ۱۰] برمی‌گردد. [۳ و ۴] استفاده گسترده از این روش تا زمانی که روش ماکسیمم درست‌نمایی را برای برآورد پارامترهای مدل پیشنهاد داد به تأخیر افتاد. [۵ و ۱۱] دانش مدل‌های رده نهن (LC) و تطبیق‌پذیری‌شان را در تحلیل داده‌های رسته‌ای گسترش دادند. کاربردهای جدید LCA نیز کشف و به نوشته‌های جدیدی منتهی شده است که می‌توان به [۲، ۶، ۸ و ۱۲] اشاره کرد. در بخش ۲ در ابتدا ساده‌ترین حالت متغیر رسته‌ای که دارای دو رده است و کم‌ترین تعداد تکرار مورد نیاز برای برآورد خطای اندازه‌گیری که دو اندازه‌گیری روی هر واحد نمونه‌ای است، در نظر گرفته می‌شود و بر اساس LCA، مدل خطای اندازه‌گیری و احتمال خطای پاسخ آن مورد بحث قرار می‌گیرد. سپس این حالت به حالت‌های کلی‌تر مثل دو اندازه‌گیری روی متغیر

از جمله روش‌هایی که برای تحلیل خطاهای غیر نمونه‌گیری به‌ویژه خطای اندازه‌گیری وجود دارد، رهیافت تحلیل رده نهن (LCA) است. زمانی که برخی از متغیرها نهن یا مشاهده‌نشده باشند، توزیع متغیر نهن بر اساس چند متغیر مشاهده‌شده نشان‌گر که برای اندازه‌گیری متغیر نهن ساخته شده‌اند، بررسی می‌شود. زمانی LCA به کار گرفته می‌شود که متغیر نهن و متغیرهای نشان‌گر، همگی متغیرهای رسته‌ای باشند. تحلیل رده نهن برای برآورد پارامترهای مدل خطای آمارگیری به کار می‌رود. این برآوردها، در ارزیابی میانگین توان دوم خطاها (MSE) و خطای کل آمارگیری (TSE) مفید هستند. همچنین این روش تحلیل می‌تواند با برآورد احتمال خطای اندازه‌گیری مربوط به یک پرسش، طراح پرسش‌نامه را در انتخاب پرسش‌های بهتر یاری کند. در کل، LCA می‌تواند به‌عنوان ابزاری مفید برای همه هدف‌های ارزیابی خطای تلقی شود و منبع‌های خطای آمارگیری

^۱ عضو هیئت علمی گروه آمار، دانشگاه علامه طباطبایی، ایران

^۲ دانش‌آموخته کارشناسی ارشد آمار، دانشگاه علامه طباطبایی، ایران

^۳ دانش‌آموخته کارشناسی ارشد آمار، دانشگاه علامه طباطبایی، ایران

رسته‌های چندحالتی و همچنین حالت سه اندازه گیری روی متغیر

رسته‌های چندحالتی، تعمیم داده می‌شود. نمادگذاری‌ها بر اساس نمادگذاری [۱] صورت گرفته است. در بخش‌های ۳ و ۴ برای توصیف مدل رده نهان، یکی از رهیافت‌های مدل‌بندی با عنوان پارامتری‌سازی مدل احتمالاتی شرح داده می‌شود. در بخش ۵ شیوه ارزیابی مدل تشریح می‌شود و در بخش ۶ با استفاده از داده‌های یک آمارگیری کوچک مقیاس، خطای بدرده‌بندی با استفاده از روش LCA برآورد می‌شود.

۲ احتمال خطای پاسخ در حالت دو

اندازه گیری روی متغیر رسته‌ای

طبق تعریف $[E(Y_i | i)]$ ، امتیاز واقعی τ_i برای متغیرهای دوحالتی که اگر پاسخ مثبت باشد، مقدار ۱ و اگر پاسخ منفی باشد، مقدار صفر اختیار می‌کند، به صورت «احتمال پاسخ مثبت برای i امین پاسخگو» تعبیر می‌شود. اگر μ_i مقدار واقعی خصیصه مورد نظر که با y_i اندازه گیری می‌شود، ϕ_i احتمال خطای مثبت کاذب^۴ و θ_i احتمال خطای منفی کاذب^۵ در اندازه گیری Y_i باشند، این احتمال‌ها به صورت زیر تعریف می‌شوند:

$$\begin{aligned}\phi_i &= P(Y_i = 1 | \mu_i = 0), \\ \theta_i &= P(Y_i = 0 | \mu_i = 1).\end{aligned}\quad (1)$$

در این صورت τ_i به صورت زیر نوشته می‌شود:

$$\tau_i = P(Y_i = 1 | i) = \begin{cases} 1 - \theta_i; & \mu_i = 1 \text{ اگر} \\ \phi_i; & \mu_i = 0 \text{ اگر} \end{cases}$$

برای نشان دادن ارتباط بین احتمال‌های رده‌بندی و خطای

اندازه گیری، مدل زیر در نظر گرفته می‌شود:

$$Y_i = \mu_i + \epsilon_i, \quad i = 1, \dots, N$$

که در آن ϵ_i مولفه خطای اندازه گیری است و به صورت زیر تعریف می‌شود:

$$\epsilon_i = \begin{cases} 1; & y_i = 1, \mu_i = 0 \\ -1; & y_i = 0, \mu_i = 1 \\ 0; & y_i = \mu_i \end{cases}$$

[۱] به بررسی اثر خطاها در برآورد شیوع جامعه‌ای $\pi = \frac{1}{N} \sum_{i=1}^N \mu_i$ طبق این مدل پرداخت. او با استفاده از احتمال‌های خطای θ_i و ϕ_i ، مفهوم‌های اعتبار، قابلیت اعتماد، واریانس پاسخ ساده، ارزیابی اندازه گیری و خطای بدرده‌بندی در داده‌های دوحالتی را ارزیابی کرده و نشان داده است که این معیارها بر اساس سه متغیر θ_i و ϕ_i و شیوع جامعه‌ای π ، به طور کامل شناسایی می‌شوند. میانگین احتمال نادرست مثبت (ϕ) و میانگین احتمال نادرست منفی (θ) به صورت زیر تعریف می‌شوند:

$$\phi = \frac{1}{N} \sum_{i=1}^N \phi_i, \quad \theta = \frac{1}{N_1} \sum_{i=1}^{N_1} \theta_i$$

که جمع اول روی عضوهایی از جامعه است که مثبت واقعی هستند (یعنی $\mu_i = 1$)، جمع دوم روی عضوهایی از جامعه عمل می‌کند که منفی واقعی هستند (یعنی $\mu_i = 0$)، $N_1 = \sum_{i=1}^N \mu_i$ تعداد افراد مثبت واقعی در جامعه است و $N_0 = N - N_1$ تعداد افراد منفی واقعی در جامعه می‌باشد.

جدول رده‌بندی تقاطعی \times متغیرهای $y_i = l$ و $\mu_i = l'$ با نسبت‌های خانه‌ای $P_{ll'}$ (که l و l' مقدارهای ۰ و ۱ را اختیار می‌کنند) را در نظر بگیرید (جدول ۱).

جدول ۱. جدول رده‌بندی تقاطعی متغیرهای دوحالتی y_i و μ_i

| | $\mu_i = 1$ | $\mu_i = 0$ | |
|-----------|----------------------------|----------------------------|----------------------------|
| $y_i = 1$ | $P_{11} = \frac{a}{n}$ | $P_{10} = \frac{b}{n}$ | $P_{1+} = P_{11} + P_{10}$ |
| $y_i = 0$ | $P_{01} = \frac{c}{n}$ | $P_{00} = \frac{d}{n}$ | $P_{0+} = P_{01} + P_{00}$ |
| | $P_{+1} = P_{11} + P_{01}$ | $P_{+0} = P_{10} + P_{00}$ | |

^۴ False-Positive

^۵ False-Negative

احتمال‌های خطا: $\pi_{1|2}^{A|X}, \pi_{2|1}^{A|X}, \pi_{1|2}^{B|X}, \pi_{2|1}^{B|X}$

برآوردگرهای ناریب ϕ و θ در حالت معلوم بودن مقادیرهای واقعی به صورت زیر به دست می‌آیند.

$$\hat{\phi} = \frac{P_{1\cdot}}{P_{1\cdot} + P_{2\cdot}} = \frac{b}{b+d},$$

$$\hat{\theta} = \frac{P_{\cdot 1}}{P_{\cdot 1} + P_{\cdot 2}} = \frac{c}{c+a}. \quad (2)$$

احتمال خطای نادرست منفی و احتمال خطای مثبت کاذب تعریف شده در (۱) متناظر با اندازه‌گیری اول، طبق نمادگذاری بالا به صورت زیر هستند:

$$\theta = P(A = 2 | X = 1) = \pi_{2|1}^{A|X}$$

$$\phi = P(A = 1 | X = 2) = \pi_{1|2}^{A|X} \quad (6-)$$

[۷] نیز روشی برای برآورد θ و π و ϕ زمانی که دو اندازه‌گیری برای متغیر Y موجود است، ارائه دادند. در روش آنها نیازی به دسترسی به اندازه‌های استاندارد طلایی و نیز فرض موازی بودن دو اندازه‌گیری y_1 و y_2 نیست، ولی به فرض استقلال موضعی دو اندازه‌گیری نیاز است. همچنین فرض می‌کنند که جامعه می‌تواند به دو گروه تقسیم شود به طوری که θ و ϕ برای دو گروه، یکسان، ولی ممکن است π برای این دو گروه متفاوت باشد. در این قسمت، روش آنها که مقدمه‌ای بر رهیافت LCA است، بیش‌تر شرح داده می‌شود.

جدول رده‌بندی GAB برای مشاهده‌ها تشکیل داده می‌شود. بردار شمارهای خانه‌ای $n = [n_{111}, \dots, n_{222}]$ دارای توزیع چندجمله‌ای با بردار احتمال

$$\pi = [\pi_{111}^{GAB}, \pi_{112}^{GAB}, \dots, \pi_{222}^{GAB}]'$$

فرض کنید X_i مقدار واقعی (مثبت واقعی با مقدار $X = 1$ و منفی واقعی با $X = 2$ نشان داده می‌شود)، A اولین اندازه‌گیری، B دومین اندازه‌گیری و G متغیر گروه‌بندی باشد که عضویت واحدها به زیرگروه‌های جامعه‌ای را نشان می‌دهد. متغیرهای A و B در صورت رده‌بندی مثبت، مقدار ۱ و در صورت رده‌بندی منفی، مقدار ۲ را اختیار می‌کنند.

است به طوری که $n_{gab}; g, a, b = 1, 2$ تعداد مشاهده‌های رده‌بندی شده در خانه $G = g, A = a, B = b$ جدول GAB و π_{gab}^{GAB} احتمال توأم این که واحد نمونه‌ای به خانه (g, a, b) رده‌بندی شود، می‌باشند. برای برآورد احتمال‌ها از روش ماکسیمم درست‌نمایی استفاده می‌شود که هسته لگاریتم درست‌نمایی در این حالت به صورت زیر است:

$$\ell(\pi | n) = \sum_{g=1}^2 \sum_{a=1}^2 \sum_{b=1}^2 n_{gab} \log(\pi_{gab}^{GAB}) \quad (5-)$$

[۷] هدف برآورد احتمال‌های خطا و یافتن برآوردی برای احتمال شیوع جامعه‌ای است. به این منظور نمادهای زیر را برای احتمال‌های مربوط معرفی کردند:

که پارامترهای آن با فرض استقلال موضعی میان دو اندازه‌گیری A و B در هر گروه عبارت‌اند از:

$$\pi_1^X = p(X = 1) \quad \text{احتمال شیوع جامعه‌ای}$$

$$\frac{\pi_1^G \pi_{1|1}^{X|G} \pi_{1|2}^{X|G} \pi_{1|21}^{A|XG} \pi_{1|22}^{A|XG} \pi_{2|11}^{A|XG} \pi_{2|12}^{A|XG}}{\pi_{1|21}^{B|XG} \pi_{1|22}^{B|XG} \pi_{2|21}^{B|XG} \pi_{2|22}^{B|XG}}$$

نسبتی از جامعه که متعلق به زیر گروه ۱ است:

نسبتی از جامعه که متعلق به زیر گروه ۲ است:

احتمال‌های حاشیه‌ای:

احتمال خانه‌های مشاهده شده:

احتمال شیوع مثبت

برای گروه $G = g$:

احتمال شیوع مقدار منفی برای

گروه $G = g$:

احتمال‌های پاسخ مصاحبه اول:

احتمال‌های پاسخ بازمصاحبه:

با توجه به عبارت زیر می‌توان نشان داد که رابطه (۴) دارای ۱۱ پارامتر بالا است [۱].

$$P(G = g, A = a, B = b) = P(G = g)P(A = a, B = b | G = g)$$

$$= \pi_g^G \pi_{ab|g}^{AB|G}$$

بنا بر این نمی‌توان برآوردهای یکتایی برای این پارامترها به دست آورد. برای شناسایی پذیر کردن مدل، باید قیدهایی به این مدل تحمیل شود.

[۷] علاوه بر فرض استقلال موضعی، یعنی

$$\pi_{ab|xg}^{AB|XG} = \pi_{a|xg}^{A|XG} \pi_{b|xg}^{B|XG}$$

کردند:

۱. احتمال های خطای اندازه گیری گروه های مختلف با هم

برابرند؛ یعنی

$$\pi_{a|x}^{A|XG} = \pi_{a|x^2}^{A|XG} = \pi_{a|x}^{A|X}, \quad a \neq x$$

$$\pi_{b|x}^{B|XG} = \pi_{b|x^2}^{B|XG} = \pi_{b|x}^{B|X}, \quad b \neq x$$

و

۲. احتمال شیوع در دو گروه متفاوت است؛ یعنی:

$$\pi_{1|1}^{X|G} \neq \pi_{1|2}^{X|G} \phi.$$

با توجه به این فرض ها، پارامترهای مدل مقید عبارت اند از:

$$\pi_1^G, \pi_{1|1}^{X|G}, \pi_{1|2}^{X|G}, \pi_{1|2}^{A|X}, \pi_{2|1}^{A|X}, \pi_{1|2}^{B|X}, \pi_{2|1}^{B|X}$$

همچنین با تحمیل این قیدها، احتمال رده بندی یک واحد نمونه ای به خانه (g, a, b) به شکل زیر درمی آید:

$$\pi_{gab}^{GAB} = \pi_1^G \left(\pi_{1|g}^{X|G} \pi_{a|1g}^{A|XG} \pi_{b|1g}^{B|XG} + \pi_{2|1}^{X|G} \pi_{a|2g}^{A|XG} \pi_{b|2g}^{B|XG} \right)$$

با جایگذاری این احتمال ها در فرمول هسته لگاریتم درست نمایی (۴) خواهیم داشت:

$$\begin{aligned} \ell(\pi | n) = & \sum_{g=1}^2 \sum_{a=1}^2 \sum_{b=1}^2 n_{gab} \left[\log \pi_g^G \right. \\ & + \log \left(\pi_{1|g}^{X|G} \pi_{a|1g}^{A|XG} \pi_{b|1g}^{B|XG} \right. \\ & \left. \left. + \pi_{2|1}^{X|G} \pi_{a|2g}^{A|XG} \pi_{b|2g}^{B|XG} \right) \right] \end{aligned}$$

برآوردهای ماکسیمم درست نمایی پارامترها (احتمال های خانه ای و به خصوص احتمال های خطا و احتمال شیوع) با استفاده از روش های عددی مثل الگوریتم EM به دست می آیند.

این گونه تحلیل خطای رده بندی به راحتی به حالتی که متغیر مورد نظر، متغیری چندحالتی است (با فرض دو اندازه گیری) تعمیم داده می شود. فرض کنید X و در نتیجه A و B هر یک K رسته دارند و متغیر گروه بندی، جامعه را به L گروه تقسیم کند. همچنین فرض کنید $x = 1, \dots, K$ ، $\pi_{x|g}^{X|G}$ ، شیوع واقعی هر رسته در هر یک از زیر گروه های جامعه ($\sum_{x=1}^K \pi_{x|g}^{X|G} = 1$) و همچنین $\pi_{a|xg}^{A|XG}$ ، $\pi_{b|xg}^{B|XG}$ ، $a \neq x, b \neq x$ و باشند. در این صورت، هسته لگاریتم درست نمایی برای جدول GAB به صورت زیر است:

$$\ell(\pi | GAB) = \sum_g \sum_a \sum_b n_{gab} \log \left(\pi_{gab}^{GAB} \right)$$

که تحت شرط استقلال موضعی خطاهای اندازه گیری داریم:

$$\pi_{gab}^{GAB} = \sum_x \pi_g^G \pi_{x|g}^{X|G} \pi_{a|xg}^{A|XG} \pi_{b|xg}^{B|XG}$$

در این حالت، تعداد خانه های جدول رده بندی تقاطعی GAB برابر با LK^2 و در نتیجه درجه آزادی آن برابر با $LK^2 - 1$ است زیرا مجموع شمارهای همه خانه ها برابر با تعداد کل مشاهده ها (n) است. برای برآورد $1 - (2K - 1)LK$ پارامتر مدل از فرض برابری احتمال های خطا در زیر گروه ها استفاده می شود، بنا بر این تعداد پارامترها $1 - (2k(k - 1)LK)$ است.

برای حالت $L = 2$ تعداد پارامترهای مدل برابر با درجه آزادی جدول، یعنی $2K^2 - 1$ می شود و بنا بر این این مدل همواره اشباع شده است؛ در نتیجه صورت بسته ای برای برآورد ماکسیمم درست نمایی (MLE) پارامترهای مدل به دست می آید. ولی برای حالت هایی که $L \geq 3$ ، مدل اشباع نشده بوده و برای تعیین MLE باید از روش های عددی مانند الگوریتم EM استفاده کرد.

حالت کلی تر، وجود بیش از دو اندازه گیری برای متغیر رسته ای چندحالتی نیز قابل تصور است. به عنوان مثال اگر سه اندازه گیری A و B و C برای متغیر نهان X با K رسته وجود داشته باشد، فرض استقلال موضعی به صورت زیر نوشته می شود:

$$\pi_{abc|x}^{ABC|X} = \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X}$$

و در هسته تابع درست نمایی، احتمال زیر قرار داده می شود:

$$\pi_{abc}^{ABC} = \sum_x \pi_x^X \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X}$$

درجه آزادی جدول ABC برابر با $K^3 - 1$ و تعداد پارامترهای مدل نامقید برابر با $(3K + 1)(K - 1)$ است. در بخش بعدی، حالت کلی مدل بندی خطای رده بندی، فرض هایی که برای شناسایی پذیری مدل های خطای رده بندی نیاز است و همچنین روش های برآورد آنها با فرض سه اندازه گیری A و B و C روی متغیر X ارائه می شوند.

۳ مدل رده نهان استاندارد

در این قسمت، حالت سه اندازه گیری یا بیش تر را بررسی می کنیم و به بیان یک ساختار اندازه پذیر توسط چند نشان گر که برای اندازه گیری آن ساختار طراحی شده اند، می پردازیم. گفته شد که

رده بندی تقاطعی ABC بر حسب X به صورت زیر نوشته می شود:

$$\pi_{abc}^{ABC} = \sum_x \pi_x^X \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X} \quad (۴-)$$

مدل (۵) یکی از ساده ترین مدل های رده نهنان است که مدل LC استاندارد نامیده می شود.

۴ پارامتری سازی مدل رده نهنان

برای توصیف یک مدل رده نهنان، چهار رهیافت مدل بندی یا پارامتری سازی را معرفی کرده است [۱]، که عبارت اند از: رهیافت پارامتری سازی مدل احتمالاتی،^۶ رهیافت مدل لگ خطی^۷ با یک متغیر نهنان، رهیافت مدل مسیر تعدیل یافته^۸ و رهیافت مدل نموداری^۹ با استفاده از نمودارهای مسیر. در این بخش فقط به رهیافت پارامتری سازی مدل احتمالاتی اشاره می شود.

۱.۴ پارامتری سازی مدل احتمالاتی LC مدل استاندارد

مدل احتمالاتی، درست نمایی جدول رده بندی تقاطعی مشاهده شده را بر حسب احتمال های شرطی و حاشیه ای مربوط به هر خانه این جدول بیان می کند. مدل احتمالاتی رده نهنان، مدل احتمالاتی ای است که شامل یک یا چند متغیر نهنان باشد.

به عنوان مثال، فرض کنید LCA شامل سه نشان گر A و B و C برای متغیر نهنان X به ترتیب با تعداد رده های k_A, k_B, k_C و k_x باشد (در بیش تر مدل ها فرض می شود $k_A = k_B = k_C = k_x$). یک مدل درست تعیین شده برای جدول رده بندی تقاطعی ABC باید بدون ابهام این، احتمال این که واحد انتخاب شده i ام درون خانه (a, b, c) جدول قرار گیرد را بر حسب پارامترهای $\pi_x^X, \pi_{a|x}^{A|X}, \pi_{b|x}^{B|X}, \pi_{c|x}^{C|X}$ که $x = 1, \dots, k_x - 1$ تعداد پارامترها در این مدل برابر با $k_x(k_A + k_B + k_C - 2)$ است (که از جمع $k_x - 1$ پارامتر مستقل متناظر با $\pi_x^X, k_x(k_A - 1)$ پارامتر مستقل متناظر با $\pi_{a|x}^{A|X}, k_x(k_B - 1)$ پارامتر مستقل متناظر با $\pi_{b|x}^{B|X}$ و $k_x(k_C - 1)$ پارامتر مستقل متناظر با $\pi_{c|x}^{C|X}$ حاصل می شود). احتمال خانه (a, b, c) در مدل LC استاندارد، برابر است

LCA چارچوبی برای بررسی رابطه میان متغیر واقعی و متغیرهایی که برای اندازه گیری آن طراحی شده اند، فراهم می آورد. یعنی اگر X متغیر نهنان (متغیر واقعی) و A و B و C نشان گرهای X (که برای اندازه گیری X طراحی شده اند) باشند، LCA به بررسی رابطه بین X و A و B و C می پردازد.

[۱] با تعمیم روش [۷]، رهیافت LCA را به تفصیل، معرفی کرده است. او با تعریف متغیرهای زیر و تحت شرطهایی که اصل های رده بندی هستند، ساده ترین مدل رده نهنان را که LC استاندارد نامیده می شود، ارائه کرد.

متغیرهای مدل رده نهنان به صورت زیر در نظر گرفته می شوند: متغیر نهنان X : که فرض می شود در شرایط آرمانی، اندازه پذیر است، ولی در شرایط واقعی با خطا اندازه گیری می شود، رده های نهنان: رسته های X ، متغیر آشکار: متغیرهایی مثل A و B و C و G که مشاهده می شوند،

متغیر نشان گر: متغیر آشکاری که به طور ویژه برای اندازه گیری X طراحی شده است (مثل A و B و C)، متغیر گروه بندی: متغیر آشکار دیگری که برای شرح پراکنندگی متغیر نهنان یا احتمال های خطا در مدل ظاهر می شود (مثل G). این متغیر بر خلاف متغیر نشان گر برای اندازه گیری X طراحی نشده است.

احتمال پاسخ مثل $\pi_{a|x}^{A|X} = P(A = a | X = x)$ احتمال شرطی نشان گرها به شرط متغیر نهنان،

احتمال خطا $\pi_{a|x}^{A|X}, \forall a \neq x$ نوعی احتمال پاسخ است که موردهایی را که در آنها X و نشان گر آن موافق نیستند را در نظر می گیرد، و

احتمال درستی $\pi_{a|x}^{A|X}, \forall a = x$ نوعی احتمال پاسخ است که موردهایی که در آن رسته متغیر نشان گر و رده نهنان یکسان اند در نظر می گیرد.

رابطه زیر برای $\pi_{a|x}^{A|X}$ برقرار است:

$$\sum_{a \neq x} \pi_{a|x}^{A|X} = 1 - \pi_{a|x}^{A|X}$$

تحت شرط استقلال موضعی، هسته تابع درست نمایی برای جدول

^۶ Proability Model

^۷ Log-Linear Model

^۸ Modified Path Model

^۹ Graphical Model

با:

جدول رده‌بندی تقاطعی XABC به صورت زیر است:

$$\ell = \sum_x \sum_a \sum_b \sum_c n_{xabc} \log(\pi_{xabc}^{XABC}) \quad (2-)$$

$$\pi_{abc}^{ABC} = \sum_x \pi_x^X \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X}$$

که در آن n_{xabc} فراوانی خانه (x,a,b,c) جدول تقاطعی XABC است و π_{xabc}^{XABC} احتمال قرارگیری واحد نام نمونه در خانه (x,a,b,c) است و از مدلی که باید تعیین کنیم، پیروی می‌کند. برای محاسبه برآورد ماکسیمم درست‌نمایی در این شرایط، نرم‌افزارهای آماری بسیاری موجودند.

تابع درست‌نمایی چندجمله‌ای کامل در مدل LC استاندارد به صورت زیر نوشته می‌شود (برای جدول رده‌بندی تقاطعی ABC که n_{abc} فراوانی خانه (a,b,c) است):

$$n = \sum_{abc} n_{abc}$$

$$\ell(\pi | n) = \frac{n!}{\prod_a \prod_b \prod_c n_{abc}!} \times \prod_a \prod_b \prod_c \left(\sum_x \pi_x^X \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X} \right)^{n_{abc}}$$

و لگاریتم تابع درست‌نمایی برابر است با:

$$\ell(\pi | n) = const + \sum_a \sum_b \sum_c n_{abc} \log \left(\sum_x \pi_x^X \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X} \right)$$

بنا بر این هسته لگ تابع درست‌نمایی مدل LC استاندارد به صورت زیر نوشته می‌شود:

$$\ell = \sum_a \sum_b \sum_c n_{abc} \log \left(\sum_x \pi_x^X \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X} \right) \quad (3-)$$

طبق تعریف، مجموع احتمال‌ها روی رسته‌های دوه‌به‌دو ناسازگار، برابر با ۱ است:

$$\sum_x \pi_x^X = \sum_a \pi_{a|x}^{A|X} = \sum_b \pi_{b|x}^{B|X} = \sum_c \pi_{c|x}^{C|X} = 1$$

تنها شرط مدل LC استاندارد، همین است و به همین دلیل، به آن مدل رده‌نهاد نامقید هم گفته می‌شود.

در مدل‌های LC مقید، قیدهای اضافی بر اساس ملاحظه‌های عملی یا نظری به مدل نامقید تحمیل می‌شوند. معمول‌ترین قیدها عبارت‌اند از:

۱. یک احتمال را برابر با برخی مقادیر ثابت فرض می‌کنند.
۲. دو یا چند احتمال را با هم برابر می‌گیرند.

در حالتی که X مشاهده نشده است و فقط $n_{abc} = \sum_x n_{xabc}$ در دست است، داریم:

$$m_{abc} = E(n_{abc})$$

$$\pi_{abc}^{ABC} = \sum_x \pi_{xabc}^{XABC}$$

در نتیجه:

$$m_{abc} = n \pi_{abc}^{ABC} = n \sum_x \pi_{xabc}^{XABC}$$

$$\ell = \sum_a \sum_b \sum_c n_{abc} \log \left(\sum_x \pi_{xabc}^{XABC} \right)$$

وجود تابع مجموع، درون لگاریتم در برآورد ماکسیمم درست‌نمایی، مشکل‌هایی به همراه دارد که برای رهایی از آنها می‌توان از روش‌های عددی مثل الگوریتم EM و نیوتن-رافسون کمک گرفت. الگوریتم EM با استفاده از بسیاری نرم‌افزارها از جمله ℓ_{EM} [۱۳] قابل دسترس است.

۲.۱.۴ الگوریتم EM برای پارامتری‌سازی احتمالاتی

در جدول ABC شمارش در خانه (a,b,c) با نماد n_{abc} نشان داده می‌شود. فرض می‌شود X متغیر نهان و سه نشان‌گر آن (A,B,C) دو حالتی هستند. پس تعداد پارامترها $7=3-1-2$ می‌کند تا ملاک همگرایی تعیین شده، برقرار شود. مقادیر پارامترها در $\pi_{a|x}^{A|X}$ ، $\pi_{b|x}^{B|X}$ و $\pi_{c|x}^{C|X}$ تکرار با π_x^X تکرار می‌شوند. نشان داده می‌شوند.

در ابتدا مقادیرهای آغازین پارامترها در گام E وارد می‌شوند (این مقادیرها توسط کاربر یا به طور تصادفی توسط نرم‌افزار اختیار می‌شوند). در این گام برای π_x^X تکرار، برآوردی از جدول رده‌بندی تقاطعی داده‌های کامل که با $XABC^{(t)}$ نشان داده

۱.۱.۴ برآورد پارامترهای مدل LC

اگر X به جای این که متغیر نهان باشد، یک متغیر مشاهده شده (آشکار) باشد، لگاریتم تابع درست‌نمایی برای خانه (x,a,b,c) از

پارامترها نیازمند هستند، که این باعث کند شدن سرعت همگرایی آنها نسبت به الگوریتم EM می شود که به مشتق دوم (مگر برای برآوردهای مدل مبنای خطای استاندارد برآوردهای پارامترهای مدل) نیاز ندارد. همچنین این روشها بر خلاف الگوریتم EM به بسیار وابسته مقادیرهای آغازین هستند.

۵ ارزیابی مدل

مدل های رده نهنان می توانند با استفاده از آماره نیکویی برازش خردی دو مثل مدل های لگ خطی سنتی، آزمون شوند. فرض کنید n_{abc} شمارش مشاهده شده برای خانه (a,b,c) از جدول رده بندی تقاطعی ABC، و \hat{m}_{abc} شمارش برآورد شده تحت مدل مفروض باشند. در این صورت آماره خردی دو پیروسون به صورت زیر به دست می آید:

$$X^2 = \sum_{abc} \frac{(n_{abc} - \hat{m}_{abc})^2}{\hat{m}_{abc}}$$

و آماره انحراف عبارت است از:

$$G^2 = 2 \sum_{abc} n_{abc} \log \left(\frac{n_{abc}}{\hat{m}_{abc}} \right)$$

شرط پذیرفته شدن مدل این است که مقدار محاسبه شده X^2 یا G^2 کوچک تر از مقدار χ^2_{df} معیار با درجه آزادی

$$df = k - np - 1 \quad (1-)$$

باشد، که k تعداد خانه های جدول و np تعداد پارامترهای مدل می باشند. همچنین اگر p مقدار آماره X^2 کوچک باشد (به عنوان مثال، کم تر از ۰/۰۵) مدل رد می شود و قضاوت می شود که برازش خوبی به داده ها صورت نگرفته است.

اگر $df = 0$ ، مدل اشباع شده است. اما اگر $df < 0$ ، مدل شناسایی ناپذیر است و نمی توان پارامترها را برآورد کرد. در این حالت باید یا قید به مدل تحمیل شود یا تعداد خانه های جدول افزایش یابد؛ به عنوان مثال، متغیر گروه بندی وارد مدل شود.

۶ کاربرد

به منظور تحلیل خطاها، یک آمارگیری در سطح یکی از دانشکده های دانشگاه های تهران اجرا شد. با توجه به نیاز به وجود مقادیرهای واقعی متغیرها برای برآورد اریبی ها، موضوع آمارگیری باید طوری انتخاب و اجرا می شد که مقدار به طور

می شود، قابل محاسبه است. شمار خانه (x,a,b,c) در جدول $XABC^{(t)}$ به صورت زیر به دست می آید:

$${}^{(t)}n_{xabc} = {}^{(t)}\pi_{x|abc}^{X|ABC} n_{abc}$$

با استفاده از قاعده بیز داریم:

$${}^{(t)}\pi_{x|abc}^{X|ABC} = \frac{{}^{(t)}\pi_{xabc}^{XABC}}{{}^{(t)}\pi_{xabc}^{XABC} + {}^{(t)}\pi_{xabc}^{XABC}}$$

بر اساس فرض های مدل LC، هر قید ثابتی که بر پارامترها تحمیل شود، باید در این گام با مقادیرهای متناظرش، جایگزین شود. در گام M، از این برآوردهای جدول XABC برای محاسبه MLE پارامترها با استفاده از درست نمایی داده های کامل استفاده می شود.

$$\ell_{full} = \log(\ell_{full}) = \sum_x \sum_a \sum_b \sum_c n_{xabc} \log \left(\sum_x \pi_x^X \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X} \right)$$

با هفت بار مشتق گرفتن از این تابع، نسبت به هفت پارامتر مورد نظر، به هفت معادله درست نمایی می رسیم که با برابر با صفر قرار دادن این معادله ها، برآورد پارامترها به صورت زیر به دست می آید.

$$\begin{aligned} {}^{(t+1)}\pi_1^X &= \frac{{}^{(t)}n_{x+++}}{n}, \\ {}^{(t+1)}\pi_{a|x}^{A|X} &= \frac{{}^{(t)}n_{ax+++}}{{}^{(t)}n_{x+++}}, \\ {}^{(t+1)}\pi_{b|x}^{B|X} &= \frac{{}^{(t)}n_{x+b+}}{{}^{(t)}n_{x+++}}, \\ {}^{(t+1)}\pi_{c|x}^{C|X} &= \frac{{}^{(t)}n_{x+++c}}{{}^{(t)}n_{x+++}}. \end{aligned}$$

از این برآوردها استفاده شده، تکرار بعدی الگوریتم EM شروع می شود و تا جایی پیش می رود که تفاوت حل های برآوردهای پارامترها بین دو تکرار، کوچک تر از مقدار همگرایی شود. در این جا گفته می شود الگوریتم به طور کامل همگرا شده است و بنا بر این احتمال شیوع و احتمال های خطا برآورد می شوند.

بیشترین فایده الگوریتم EM زمانی است که هیچ صورت بسته ای برای معادله های درست نمایی داده های ناکامل موجود نباشد، در حالی که معادله های درست نمایی داده های کامل به راحتی قابل حل باشند. پس الگوریتم EM برای بسیاری از مدل های LC آرمانی است. روش های نیوتن-رافسون و امتیازبندی فیشر به مشتق دوم معادله های درست نمایی بر حسب

تقریبی بدون خطای متغیرها قابل دسترسی باشند. در قسمت آموزش این دانشکده برخی اطلاعات در مورد وضعیت تحصیلی دانشجویان ثبت می‌شود که می‌توان این اندازه‌ها را اندازه‌های استاندارد طلایی تلقی کرد. بنا بر این در این بررسی، هدف‌ها و جامعه هدف به صورت زیر در نظر گرفته شدند.

۱.۶ طرح مطالعه

جامعه هدف در این کاربرد، همه دانشجویان مقطع کارشناسی ارشد ورودی سال‌های ۱۳۸۹ و ۱۳۹۰ دانشکده مورد نظر بوده است که در نیم‌سال اول سال تحصیلی ۱۳۹۰-۱۳۹۱ در رشته‌های موجود در این دانشکده اشتغال به تحصیل داشته‌اند. در این مطالعه پرسش‌نامه‌ای طراحی و واحدهای نمونه‌ای به تصادف به سه پرسشگر مورد استفاده منتسب شدند و به صورت تلفنی مورد مصاحبه قرار گرفتند.

چارچوب نمونه‌گیری شامل فهرست همه شماره‌های دانشجویی دانشجویان کارشناسی ارشد ورودی سال‌های ۱۳۸۹ و ۱۳۹۰ همه رشته‌های این دانشکده به همراه با معدل کل، تعداد واحدهای گذرانیده شده، تعداد ترم‌های مشروطی تا انتهای نیم‌سال اول سال تحصیلی ۱۳۹۰-۱۳۹۱ و نیز شماره تماس آنها بود که از طریق اداره تحصیلات تکمیلی این دانشکده در اختیار پژوهشگر قرار گرفت.

از آن‌جا که به نظر می‌رسید وضعیت نمره‌ها و شرایط درسی برای رشته‌های مختلف متفاوت باشد، برای گزینش نمونه از روش نمونه‌گیری طبقه‌ای استفاده شد. چارچوب بر اساس رشته‌های تحصیلی طبقه‌بندی شد و سه طبقه آمار اقتصادی-اجتماعی (طبقه اول)، آمار ریاضی (طبقه دوم) و ریاضیات مالی (طبقه سوم) تشکیل شدند.

واحد نمونه‌گیری در این آمارگیری، هر یک از دانشجویان کارشناسی ارشد رشته‌های آمار اقتصادی-اجتماعی، آمار ریاضی، و ریاضیات مالی ورودی سال‌های ۱۳۸۹ و ۱۳۹۰ دانشکده مورد نظر بوده است که در نیم‌سال اول سال تحصیلی ۱۳۹۰-۱۳۹۱ اشتغال به تحصیل داشته‌اند. جامعه هدف در این مطالعه شامل $N = 94$ دانشجو بوده است. اندازه‌های جامعه هر یک از طبقه‌ها در جدول ۲ آمده است.

برای تعیین تعداد واحدهای نمونه‌ای در هر طبقه، m_h یک نمونه مقدماتی از هر طبقه به اندازه n_{hpre} به صورت تصادفی انتخاب و اندازه نمونه نهایی برای هر طبقه از فرمول زیر محاسبه

$$n_h = \frac{Z_{\frac{\alpha}{2}}^2 \frac{s_h^2}{\bar{y}_h}}{e_h^2 + \frac{Z_{\frac{\alpha}{2}}^2 \frac{s_h^2}{N_h}}{N_h}} \quad (۰)$$

که در آن $s_h^2 = \frac{n_{hpre}}{n_{hpre}-1} \hat{p}_{hpre}(1 - \hat{p}_{hpre})$ و N_h اندازه جامعه طبقه h ، n_{hpre} اندازه نمونه مقدماتی از طبقه h ، \bar{y}_h برآورد نسبت متناظر با طبقه h ، e_h حاشیه خطای طبقه h و $Z_{\frac{\alpha}{2}}$ صدک $(1 - \frac{\alpha}{2})$ ام توزیع نرمال استاندارد می‌باشند. در این مطالعه $\alpha = 0.1$ و $e_h = 0.2$ ، $h = 1, 2, 3$ در نظر گرفته شدند.

متغیرهای مورد بررسی، از طریق پرسش‌های زیر اندازه‌گیری شدند:

Q۲: آیا تا پایان ترم اول سال تحصیلی ۱۳۹۰-۱۳۹۱ نمره ضعیف و ناخوشایندی از هیچ‌یک از درس‌های کارشناسی ارشدتان گرفته‌اید؟

بله خیر

Q۳: کم‌ترین معدل ترمی شما تا پایان نیم‌سال اول سال تحصیلی ۱۳۹۰-۱۳۹۱ در کدام یک از رشته‌های زیر قرار داشته است؟

کم‌تر از ۱۴ ۱۴-۱۶ ۱۶-۱۸

بیش‌تر از ۱۸

Q۴: آیا تا پایان نیم‌سال اول سال تحصیلی ۱۳۹۰-۱۳۹۱ از هیچ‌یک از درس‌های کارشناسی ارشدتان افتاده‌اید؟

بله خیر

Q۷: آیا در هیچ ترمی تا پایان نیم‌سال اول سال تحصیلی ۱۳۹۰-۱۳۹۱ مشروط شده‌اید؟

بله خیر

متغیرهای دوحالتی Y_2, Y_3, Y_4, Y_7 به ترتیب متناظر با چهار پرسش Q_2, Q_3, Q_4, Q_7 تعریف می‌شوند، به طوری که متغیرهای Y_2, Y_4, Y_7 در صورتی که دانشجو پاسخ بله را انتخاب کند، مقدار ۱ و در صورت انتخاب گزینه «خیر»، مقدار «صفر» را اختیار می‌کنند. متغیر Y_3 نیز در صورتی که فرد رسته کم‌تر از ۱۴ را انتخاب کرده باشد، مقدار ۱ و در صورت انتخاب بقیه گزینه‌ها مقدار صفر را اختیار می‌کند.

برای محاسبه اندازه نمونه هر طبقه، بزرگ‌ترین واریانس متناظر با این چهار متغیر در نظر گرفته شد. با قرار دادن $\alpha = 0.1$ و $e_h = 0.2$ در فرمول (۹)، اندازه نمونه برای هر طبقه برآورد شد. جدول ۳ اندازه نمونه محاسبه شده برای هر طبقه را نشان می‌دهد.

جدول ۲. اندازه جامعه هر یک از طبقه‌ها

| طبقه | آمار اقتصادی-اجتماعی | آمار ریاضی | ریاضیات مالی |
|-----------------|----------------------|------------|--------------|
| اندازه جامعه‌ای | $N_1 = 31$ | $N_2 = 34$ | $N_3 = 29$ |

جدول ۳. اندازه نمونه هر یک از طبقه‌ها

| طبقه | آمار اقتصادی-اجتماعی | آمار ریاضی | ریاضیات مالی |
|----------------------|----------------------|------------|--------------|
| اندازه نمونه مقدماتی | ۳ | ۷ | ۴ |
| S^2 بزرگ‌ترین | ۰/۳۳۳ | ۰/۲۸۵ | ۰/۳۳۳ |
| مقدار فرمول (۹) | ۱۳/۰۰۷ | ۱۲/۲۵۵۷ | ۱۲/۶۳۵۲ |
| اندازه نمونه | $n_1 = 13$ | $n_2 = 13$ | $n_3 = 13$ |
| تعداد نمونه مانده | ۱۰ | ۶ | ۹ |

اما جدول ۴ برای انتخاب پرسش مناسب‌تر برای نیل به هدف تعیین نسبت مشروطی اطلاعی به دست نمی‌دهد. با مقایسه پاسخ‌های پرسش‌ها با داده‌های ثبتي (داده‌های کارنامه‌ای (Y_M) ، پرسشی که مطابقت بیشتری با رده‌بندی داده‌های ثبتي داشته باشد، مناسب‌تر است (جدول ۵).

۲.۶ برآورد خطاها با استفاده از LCA

در این قسمت، خطای رده‌بندی مربوط به هر پرسش، در ابتدا در حالتی که مقدارهای واقعی در دست است و سپس در حالت عدم دسترسی به اندازه‌های واقعی با کمک LCA برآورد می‌شوند.

۱.۲.۶ برآورد ϕ و θ با استفاده از داده‌های ثبتي

جدول ۵. جدول رده‌بندی تقاطعی پاسخ‌های دو پرسش و داده‌های

ثبتي Q_3 و Q_7

| | | Y_M | | کل |
|-------|-----|-------|-----|----|
| | | بله | خیر | |
| Y_3 | بله | ۵ | ۱ | ۶ |
| | خیر | ۴ | ۲۹ | ۳۳ |
| | کل | ۹ | ۳۰ | ۳۹ |
| Y_7 | بله | ۸ | ۰ | ۸ |
| | خیر | ۱ | ۳۰ | ۳۱ |
| | کل | ۹ | ۳۰ | ۳۹ |

با توجه به جدول ۵ (که در آن داده‌های ثبتي کارنامه‌ای، اندازه‌های استاندارد طلائی، Y_M ، فرض شده‌اند)، یک نفر وجود دارد که پاسخ او به پرسش Q_3 مثبت بوده، یعنی اعلام کرده که ترم مشروطی داشته، در حالی که طبق کارنامه وی درس افتاده نداشته است. همچنین چهار نفری که اعلام نکرده‌اند ترم مشروطی داشته‌اند، در معدل‌های ترمی آنها معدل زیر ۱۴ یافت شده است. بنا بر این بین پاسخ‌های افراد به Q_3 و اندازه‌های استاندارد طلائی (Y_M) به‌طور کامل مطابقت وجود ندارد. با نظر به پرسش Q_7 نیز مشاهده می‌شود که یک نفری که ترم مشروطی داشته، آن را اظهار نکرده و پاسخ منفی به این پرسش داده است؛ پس هیچ‌یک از دو پرسش Q_3 و Q_7 بدون خطا نبوده‌اند.

گفته شد که دو پرسش Q_3 و Q_7 پرسش‌نامه برای اندازه‌گیری خصیصه‌ای یکسان یعنی نسبت مشروطی دانشجویان طراحی شده‌اند و باید به پاسخ‌های مشابهی منتهی شوند (البته پرسش Q_3 نیز به متغیر دو حالتی تبدیل شده است، در صورتی که پاسخگو رسته معدل کم‌تر از ۱۴ را انتخاب کرده باشد، متغیر Y_3 مقدار ۱ و در صورت انتخاب بقیه رسته‌ها، مقدار صفر را می‌گیرد). با رسم جدول رده‌بندی تقاطعی این دو پرسش، خانه‌های غیر قطر اصلی جدول نشان‌دهنده عدم توافق بین پاسخ فرد به این دو پرسش و نشان‌دهنده وجود خطای رده‌بندی توسط دست کم یکی از پرسش‌ها می‌باشند. این جدول برای پرسش‌های Q_3 و Q_7 در جدول ۴ نشان داده شده است. در این جدول دیده می‌شود که یک نفر که به پرسش Q_3 پاسخ مثبت داده، پاسخش به Q_7 منفی بوده است و سه نفر که به Q_7 پاسخ مثبت داده‌اند، به Q_3 پاسخ منفی داده‌اند.

جدول ۴. جدول رده‌بندی تقاطعی پاسخ‌های دو پرسش Q_3 و Q_7

| | | Y_7 | | کل |
|-------|-----|-------|-----|----|
| | | بله | خیر | |
| Y_3 | بله | ۵ | ۱ | ۶ |
| | خیر | ۲ | ۳۰ | ۳۳ |
| | کل | ۸ | ۳۱ | ۳۹ |

در این کاربرد، دو پرسش Q۳ (کدبندی شده به متغیر دو حالتی) و Q۷ به عنوان دو نشان گر A و B برای متغیر نهن X (که دست کم یک بار مشروط شدن دانشجو را نشان می‌دهد) در نظر گرفته شدند. فرض کنید

$$X = \begin{cases} ۱, & \text{اگر دانشجو مشروط شده باشد} \\ ۲, & \text{اگر دانشجو مشروط نشده باشد} \end{cases}$$

$$A, B = \begin{cases} ۱, & \text{اگر پاسخ مثبت باشد} \\ ۲, & \text{اگر پاسخ منفی باشد} \end{cases}$$

با توجه به پاسخ‌های دانشجویان حاضر در نمونه، جدول رده‌بندی تقاطعی AB تشکیل داده شد (جدول ۴). در به کار گیری مدل رده نهن فرض شد که خطای اندازه‌گیری نشان‌گرها مستقل از هم هستند؛ یعنی مدل‌های اندازه‌گیری $y_{iA} = \mu_i + \epsilon_{iA}$ و $y_{iB} = \mu_i + \epsilon_{iB}$ برای دو نشان گر A و B در نظر گرفته شدند، که در آنها y_{iA} پاسخ فرد i ام به اندازه‌گیری A، y_{iB} پاسخ فرد i ام به اندازه‌گیری B، μ_i پاسخ واقعی (مقدار واقعی) فرد i ام، ϵ_{iA} خطای اندازه‌گیری در اندازه‌گیری اول، و ϵ_{iB} خطای اندازه‌گیری در باز اندازه‌گیری مجدد هستند.

جدول ۴. جدول رده‌بندی تقاطعی نشان‌گرها و نسبت مشروطی‌ها به تفکیک رشته تحصیلی

| G | | | B | | کل |
|---|---|----|---|----|----|
| | | | ۲ | ۱ | |
| ۱ | A | ۱ | ۱ | ۱ | ۲ |
| | | ۲ | ۱ | ۱۰ | ۱۱ |
| | | کل | ۲ | ۱۱ | ۱۳ |
| ۲ | A | ۱ | ۳ | ۰ | ۳ |
| | | ۲ | ۲ | ۸ | ۱۰ |
| | | کل | ۵ | ۸ | ۱۳ |
| ۳ | A | ۱ | ۱ | ۰ | ۱ |
| | | ۲ | ۰ | ۱۲ | ۱۲ |
| | | کل | ۱ | ۱۲ | ۱۳ |

فرض استقلال ϵ_{iA} و ϵ_{iB} با استفاده از آزمون خی دو (χ^2) بررسی شد و این فرض در سطح معنی‌داری ۰/۰۱ رد نشد. مدل‌های مختلفی به داده‌ها برازش داده شد و بر اساس ملاک خی دو، مدل احتمالاتی انتخاب شد. سپس با استفاده از نرم‌افزار ℓEM (که الگوریتم EM را به کار می‌گیرد) از رهیافت مدل‌بندی احتمالاتی

فرض کنید ϕ و θ به ترتیب نشان‌دهنده احتمال‌های خطای مثبت کاذب یعنی احتمال این که فردی که در واقعیت مشروط نداشته، در رسته دانشجویان مشروط شده قرار داده شده باشد) و احتمال خطای نادرست منفی (یعنی احتمال این که فردی که در واقعیت مشروطی داشته به اشتباه در رسته دانشجویان غیر مشروطی رسته‌بندی شده باشد) برای اندازه‌گیری y باشند، که در ۱ تعریف شده‌اند. با دانستن مقدار واقعی μ_i طبق رابطه ۲ می‌توان این احتمال‌ها را برآورد کرد. بنا بر این با استفاده از داده‌های جدول ۵ برآوردهای θ و ϕ برای Q۳ و Q۷ به صورت زیر به دست می‌آیند.

$$\hat{\phi}_3 = \frac{۱}{۱+۲۹} = ۰/۰۳۳, \quad \hat{\phi}_7 = \frac{۰}{۰+۳۰} = ۰,$$

$$\hat{\theta}_3 = \frac{۴}{۴+۵} = ۰/۴۴, \quad \hat{\theta}_7 = \frac{۱}{۱+۸} = ۰/۱۱.$$

بر اساس پرسش اول (Q۳)، حدود ۳ درصد از دانشجویانی که مشروط نشده‌اند، اظهار داشته‌اند که مشروط شده‌اند و حدود ۴۴ درصد از دانشجویانی که دست کم در یک ترم از دوره کارشناسی ارشد تا نیم‌سال اول سال تحصیلی ۱۳۹۰-۱۳۹۱ مشروط شده‌اند، گفته‌اند که در هیچ ترمی مشروط نشده‌اند. بر اساس پرسش دوم (Q۷)، هیچ‌یک از دانشجویانی که تا نیم‌سال اول سال تحصیلی ۱۳۹۰-۱۳۹۱ مشروط نشده‌اند، پاسخ مثبت نداده‌اند (یعنی نگفته‌اند که مشروط شده‌ایم) و فقط حدود ۱۱ درصد از کسانی که مشروط شده‌اند، مشروط شدن‌شان را اعلام نکرده‌اند. به نظر می‌رسد که پرسش دوم به نتیجه‌های واقعی‌تری منتهی شده است. بنا بر این پرسش Q۷ برای هدف ما مناسب‌تر بوده است. دلیل این امر می‌تواند وجود عبارت «کم‌ترین معدل ترم» در پرسش اول باشد که باعث گمراه شدن پاسخگو و بد فهمیده شدن منظور پرسش شده باشد، که البته با انجام دادن مصاحبه تشخیصی با دانشجویان می‌توان مشکل بد تفسیر شدن پرسش را رفع کرد. علاوه بر این ممکن است افراد، معدل‌های ترم‌های تحصیلی‌شان را فراموش کرده باشند؛ ولی عبارت مشروط شدن که در پرسش دوم آمده است، کم‌تر از ذهن پاسخگو پاک شده باشد.

۲.۲.۶ برآورد ϕ و θ به روش LCA

در حالتی که داده‌های ثبتي (اندازه‌های استاندارد طلایی) در دسترس نباشند، از روش LCA برای انتخاب پرسش مناسب‌تر به کمک برآورد پارامترهای ϕ و θ ی متناظر با پرسش‌ها و مقایسه آنها استفاده می‌شود.

$\{G | X | G | A | G | B | G\}$ می‌باشند.

بخشی از خروجی برنامه برازش این مدل احتمالاتی که در شکل ۱ آمده است، برآورد احتمال‌های پاسخ شرطی دو نشان‌گر A و B را نشان می‌دهد. احتمال‌های خطا که در LCA به صورت رابطه (۳) تعریف می‌شوند، از این قسمت از خروجی برنامه استخراج می‌شوند. احتمال خطای نادرست منفی برای A برابر با . درصد برآورد شده است؛ یعنی پرسش اول در حدود ۳۷ درصد از پاسخگویانی را که در واقع تا انتهای نیم‌سال اول سال تحصیلی ۱۳۹۰-۱۳۹۱ دست کم یک ترم مشروط شده بودند، به‌عنوان پاسخگویانی که تا انتهای این نیم‌سال مشروط نشده بودند، بدرده‌بندی کرده است. ولی پرسش دوم در حدود ۱۷ درصد (کم‌تر از نصف نسبت مربوط به پرسش اول) از افرادی را که در واقع مشروطی داشته‌اند به‌عنوان دانشجویانی که تا آن نیم‌سال در هیچ ترمی مشروط نشده بودند، بدرده‌بندی کرده است. تفاوت احتمال خطای نادرست منفی دو پرسش در حدود ۲۰ درصد بوده است.

هیچ‌یک از پرسش‌های اول و دوم، افرادی را که در واقع تا انتهای نیم‌سال اول سال تحصیلی ۹۰-۹۱ مشروط نشده بودند بدرده‌بندی نکرده است و این افراد را در رسته دانشجویانی که دست کم یک‌بار مشروط شده بودند، قرار نداده است؛ یعنی احتمال خطای مثبت کاذب این دو پرسش، صفر برآورد شده است. البته منطقی هم است که افراد به‌اشتباه اظهار نکنند که دارای یک صفت نامطلوب از نظر اجتماعی هستند و صفت مشروط شدن، خصیصه‌ای نیست که افراد به‌دنبال داشتن آن باشند، در حالی که عکس آن محتمل است؛ یعنی در صورت دارا بودن خصیصه نامطلوب، افراد سعی می‌کنند اعلام کنند که این صفت را ندارند. به همین دلیل، احتمال خطای نادرست منفی برای این دو پرسش بیش‌تر از احتمال خطای مثبت کاذب می‌باشد. بنا بر این پرسش دوم، یعنی Q۷، بر پرسش اول ترجیح داده می‌شود.

۷ بحث و نتیجه‌گیری

تحلیل رده نهن (LCA) برای برآورد برخی پارامترهای مدل خطای آمارگیری به کار می‌رود. همچنین این روش تحلیل، می‌تواند با برآورد احتمال خطای مربوط به یک پرسش، طراح پرسش‌نامه را در انتخاب پرسش‌های بهتر یاری کند. در یک آمارگیری کوچک مقیاس خطای بدرده‌بندی نسبت دانشجویانی

استفاده و مدل رده نهن استاندارد (۵) (با این تفاوت که در این‌جا فقط دو نشان‌گر A و B وجود دارد) به شمارش‌های خانه‌ای جدول رده‌بندی تقاطعی (۴) طبق روش بیان‌شده در زیربخش ۲.۱.۴ برازش داده شد. پس از ۱۹ بار تکرار الگوریتم EM و با ملاک همگرایی 0.0000008 برآوردهایی برای احتمال‌های پاسخ و نسبت‌های شیوع واقعی به دست آمد. درجه آزادی مدل برازش یافته بر اساس رابطه (۸) برابر است با:

$$df = k - np - 1 = 4 - 5 - 1 = -2 < 0$$

| P(G) | P(A X) | P(B X) |
|--------|-----------------|-----------------------|
| 1 | 0.3333 (0.0755) | 1 1 0.6250 (0.1712) |
| 2 | 0.3333 (0.0755) | 2 1 0.3750 (0.1712) |
| 3 | 0.3333 (0.0755) | 1 2 0.0000 (0.0000) |
| | | 2 2 1.0000 (0.0000) |
| P(X G) | P(B X) | |
| 1 1 | 0.2461 (0.1260) | 1 1 0.8334 (0.1522) |
| 1 2 | 0.4102 (0.1471) | 2 1 0.1666 (0.1522) |
| 1 3 | 0.0820 (0.0791) | 1 2 0.0000 (0.0000) |
| 2 1 | 0.7539 (0.1260) | 2 2 1.0000 (0.0000) |
| 2 2 | 0.5898 (0.1471) | |
| 2 3 | 0.9180 (0.0791) | |

شکل ۱. برآورد احتمال‌های پاسخ شرطی دو نشان‌گر A و B وضعیت مشروطی

بنا بر این مدل شناسایی‌پذیر نیست. در این حالت می‌توان یک متغیر گروه‌بندی G مثل متغیر رشته تحصیلی را به‌منظور افزایش درجه آزادی، وارد مدل کرد. این متغیر در صورتی که دانشجو در رشته آمار اقتصادی - اجتماعی باشد، مقدار ۱، در صورت عضویت دانشجو در طبقه آمار ریاضی مقدار ۲، و در صورتی که پاسخگو دانشجوی رشته ریاضیات مالی باشد، مقدار ۳ را اختیار می‌کند. جدول رده‌بندی تقاطعی نشان‌گرهای A و B (متناظر با Y_3, Y_7) به تفکیک رشته تحصیلی در جدول ۶ آمده است. شمارش‌های خانه‌ای این جدول، داده‌های ورودی برنامه برازش مدل احتمالاتی

که دست کم یک درس را افتاده‌اند و نیز خطای رده‌بندی نسبت دانشجویانی که دست کم یک بار مشروط شده‌اند، محاسبه شد. پرسش‌های Q۳ و Q۷ که برای اندازه‌گیری یک خصیصه یکسان (وضعیت مشروطی دانشجو) در پرسش‌نامه آورده شده‌اند، به‌عنوان نشان‌گرهایی برای خصیصه مورد بررسی (که متغیر نهان نامیده می‌شود) در نظر گرفته شدند. با به کارگیری روش تحلیل رده‌نهاد، برآوردی از خطای رده‌بندی متناظر با هر پرسش به دست آمد. احتمال خطای نادرست منفی متناظر با پرسش Q۳ برابر با ۰/۳۷۵۰ و برای پرسش Q۷ برابر با ۰/۱۶۶۶ برآورد شدند. برآورد احتمال خطای مثبت کاذب هر دو پرسش نیز صفر به دست آمد.

پرسشی که احتمال‌های خطاهای رده‌بندی متناظر با آن کوچک‌تر بودند، یعنی Q۷ به‌عنوان پرسش مناسب‌تر برگزیده شد.

مراجع

- [1] Biemer, P. P. (2011), *Latent Class Analysis of Survey Error*, New York: John Wiley & Sons Inc.
- [2] Collins, L. and Lanza, S. (2010), *Latent Class and Latent Transition Analysis*, Hoboken, NJ, New York: John Wiley & Sons Inc.
- [3] Goodman, L. A. (1974a), Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable models, *Biometrika*, **61**, 215-231.
- [4] Goodman, L. A. (1974b), The Analysis of systems of Qualitative Variables When Some of the Variables Are Unobservable. Part I: A modified Latent Structure Approach, *American Journal of Sociology*, **79**, 1179-1259.
- [5] Haberman, S. J. (1979), *Analysis of Qualitative Data*, New York: John Wiley & Sons Inc. Academic Press.
- [6] Hagenaars, J. A. and McCutcheon, A. L. (2002), *Applied Latent Class Analysis*, New York: John Wiley & Sons Inc. Cambridge University Press.
- [7] Hui, S. L. and Walter, S. D. (1980), Estimating the Error Rates of Diagnostic Tests, *Biometrics*, **36**, 167-171.
- [8] Langeheine, R. and Rost, J. (1988), *Sample Survey Methods and Theory*, New York: John Wiley & Sons Inc. Plenum Press.
- [9] Lazarsfeld, P. F. (1950a), *The Logical and Mathematical Foundations of Latent Structure Analysis*, in S. A. Stouffer et al., eds., *Measurement and prediction*, Princeton: University Press, Chapter 10, 362-412.
- [10] Lazarsfeld, P. F. (1960b), *Some Latent Structures*, in S. A. Stouffer et al., eds., *Measurement and Prediction*, Princeton University Press, Princeton, Chapter 11, 362-412.
- [11] McCutcheon, A. L. (1987), *Latent Class Analysis, Quantitative applications in the Social Science Series*, Vol. 64, Sage Publication, Thousand Oaks, CA.
- [12] Rost, J. and Langeheine, R. (1977), *Application of Latent Trait and Latent Class Models in the Social Sciences*, New York: John Wiley & Sons Inc. Waxmann Munster.
- [13] Vermunt, J. K. (1977), *LEM: A General Program for the Analysis of Categorical Data*, Tilburg, Department of Methodology and Statistics, Tilburg University, Netherland.