

کاربردی از ارزیابی خطای اندازه گیری به روش تحلیل رده نهان

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

تاریخ دریافت: ۱۳۹۴/۷/۳

تاریخ پذیرش: ۱۳۹۶/۱/۳۰

چکیده:

تحلیل رده نهان (LCA) روشی برای ارزیابی خطاهای غیر نمونه گیری، به خصوص خطای اندازه گیری داده های رسته ای است. [۱]، چهار رهیافت مدل بندی رده نهان، یعنی پارامتری سازی مدل احتمالاتی، مدل لگ خطی، مدل مسیر تعدیل یافته و مدل نموداری را با استفاده از نمودارهای مسیر معرفی کرده است. این مدل ها قابل تبدیل به یکدیگرند. مدل های احتمالاتی رده نهان، درست نمای جدول رده بندی تقاطعی متغیرها را بر حسب احتمال های شرطی و حاشیه ای مربوط به هر خانه این جدول بیان می کند. در این رهیافت پارامترهای مدل با استفاده از الگوریتم EM برآورد می شوند. برای آزمون مدل رده نهان، آماره خی دو به عنوان ملاک نیکویی برازش معرفی شده است. در این مقاله از LCA و داده های یک آمارگیری کوچک مقیاس برای محاسبه خطای بدرده بندی (که یک نوع خطای اندازه گیری است) نسبت دانشجویانی که دست کم در یک درس مردود شده اند و نیز خطای بدرده بندی نسبت دانشجویانی که دست کم یک بار مشروط شده اند، استفاده شده است.

واژه های کلیدی: خطای کل آمارگیری، خطای اندازه گیری، مدل احتمالاتی، تحلیل رده نهان، استاندارد طلایی، خطای بدرده بندی.

۱ مقدمه

را ارزیابی کند؛ اما در این مقاله توجه ما بر اندازه گیری های رسته ای و خطاهای رده بندی، یعنی بدرده بندی کردن یک رسته واقعی به یکی از رسته های ممکن آن متغیر (خطای اندازه گیری) است. ایده کلی مدل های رده نهان به کارهای اولیه لازارسفلد [۹ و ۱۰] برمی گردد. [۳ و ۴] استفاده گسترده از این روش تا زمانی که روش ما کسیم درست نمایی را برای برآورد پارامترهای مدل پیشنهاد داد به تأخیر افتاد. [۵ و ۱۱] دانش مدل های رده نهان (LC) و تطبیق پذیری شان را در تحلیل داده های رسته ای گسترش دادند. کاربردهای جدید LCA نیز کشف و به نوشته های جدیدی منتهی شده است که می توان به [۲، ۶، ۸ و ۱۲] اشاره کرد. در بخش ۲ در ابتدا ساده ترین حالت متغیر رسته ای که دارای دو رده است و کم ترین تعداد تکرار مورد نیاز برای برآورد خطای اندازه گیری که دو اندازه گیری روی هر واحد نمونه ای است، در نظر گرفته می شود و بر اساس LCA، مدل خطای اندازه گیری و احتمال خطای پاسخ آن مورد بحث قرار می گیرد. سپس این حالت به حالت های کلی تر مثل دو اندازه گیری روی متغیر

از جمله روش هایی که برای تحلیل خطاهای غیر نمونه گیری به ویژه خطای اندازه گیری وجود دارد، رهیافت تحلیل رده نهان (LCA) است. زمانی که برخی از متغیرها نهان یا مشاهده نشده باشند، توزیع متغیر نهان بر اساس چند متغیر مشاهده شده نشان گر که برای اندازه گیری متغیر نهان ساخته شده اند، بررسی می شود. زمانی LCA به کار گرفته می شود که متغیر نهان و متغیرهای نشان گر، همگی متغیرهای رسته ای باشند. تحلیل رده نهان برای برآورد پارامترهای مدل خطای آمارگیری به کار می رود. این برآوردها، در ارزیابی میانگین توان دوم خطاها (MSE) و خطای کل آمارگیری (TSE) مفید هستند. همچنین این روش تحلیل می تواند با برآورد احتمال خطای اندازه گیری مربوط به یک پرسش، طراح پرسش نامه را در انتخاب پرسش های بهتر یاری کند. در کل، LCA می تواند به عنوان ابزاری مفید برای همه هدف های ارزیابی خطا تلقی شود و منبع های خطای آمارگیری

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

^۲ دانش آموخته کارشناسی ارشد آمار، دانشگاه علامه طباطبائی، ایران

^۳ دانش آموخته کارشناسی ارشد آمار، دانشگاه علامه طباطبائی، ایران

اندازه‌گیری، مدل زیر در نظر گرفته می‌شود:

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

که در آن ϵ_i مولفه خطای اندازه‌گیری 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 π طبق $\frac{1}{N} \sum_{i=1}^N \mu_i$ در مدل پرداخت. او با استفاده از احتمال‌های خطای θ_i و ϕ_i ، مفهوم‌های اعتبار، قابلیت اعتماد، واریانس پاسخ ساده، ارزیابی اندازه‌گیری 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:

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

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).

رسته‌ای چندحالتی و همچنین حالت سه اندازه‌گیری روی متغیر رسته‌ای چندحالتی، تعمیم داده می‌شود. نمادگذاری‌ها بر اساس نمادگذاری [۱] صورت گرفته است. در بخش‌های ۳ و ۴ برای توصیف مدل رده‌نهن، 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 اندازه‌گیری 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 اندازه‌گیری می‌شود، ϕ_i احتمال خطای مثبت کاذب^۴ و θ_i احتمال خطای منفی کاذب^۵ در اندازه‌گیری Y_i باشند، این احتمال‌ها به صورت زیر تعریف می‌شوند:

$$\begin{aligned} \phi_i &= P(Y_i = 1 | \mu_i = 0), \\ \theta_i &= P(Y_i = 0 | \mu_i = 1). \end{aligned} \quad (۱)$$

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

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

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

احتمال‌های خطا:

برآوردگرهای ناریب ϕ و θ در حالت معلوم بودن مقدارهای

واقعی به صورت زیر به دست می‌آیند.

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

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

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

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

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

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

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

است به طوری که $g, a, b = 1, 2$ 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 ماکسیم درست‌نمایی استفاده می‌شود که هسته لگاریتم درست‌نمایی در این حالت به صورت زیر است:

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

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

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

$$\pi_{1|21}^{B|XG} \pi_{1|22}^{B|XG} \pi_{2|21}^{B|XG} \pi_{2|22}^{B|XG}$$

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

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

دو فرض زیر را به مدل تحمیل کردند:

[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.

فرض کنید X_i مقدار واقعی (مثبت واقعی با مقدار $X=1$ و منفی واقعی با $X=2$ نشان داده می‌شود)، 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.

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

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

نسبتی از جامعه که متعلق به زیرگروه ۱ است:

نسبتی از جامعه که متعلق به زیرگروه 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:

۱. احتمال‌های خطای اندازه‌گیری گروه‌های مختلف با هم که تحت شرط استقلال موضعی خطاهای اندازه‌گیری 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 خانه‌های جدول رده‌بندی تقاطعی GAB برابر با LK^2 و در نتیجه درجه آزادی آن برابر با $LK^2 - 1$ است زیرا مجموع شمارهای همه خانه‌ها برابر با تعداد کل مشاهده‌ها (n) است. برای برآورد $LK(2K - 1) - 1$ پارامتر مدل از فرض برابری احتمال‌های خطا در زیرگروه‌ها استفاده می‌شود، بنا بر این تعداد پارامترها $LK + 2k(k - 1) - 1$ است.

برای حالت $L = 2$ تعداد پارامترهای مدل برابر با درجه آزادی جدول، یعنی $2K^2 - 1$ می‌شود و بنا بر این این مدل همواره اشباع شده است؛ در نتیجه صورت بسته‌ای برای برآورد ماکسیمم درست‌نمایی (MLE) پارامترهای مدل به دست می‌آید. ولی برای حالت‌هایی که $L \geq 3$ ، مدل اشباع نشده بوده و برای تعیین MLEها باید از روش‌های عددی مانند الگوریتم EM استفاده کرد.

حالت کلی‌تر، وجود بیش از دو اندازه‌گیری برای متغیر رسته‌ای چندحالتی نیز قابل تصور است. به‌عنوان مثال اگر سه اندازه‌گیری A و B و C برای متغیر نهان X با K رسته وجود داشته باشد، فرض استقلال موضعی به‌صورت زیر نوشته می‌شود:

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

و در هسته تابع درست‌نمایی، احتمال زیر قرار داده می‌شود:

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

درجه آزادی جدول ABC برابر با $K^3 - 1$ و تعداد پارامترهای مدل نامعقد برابر با $(3K + 1)(K - 1)$ است. در بخش بعدی، حالت کلی مدل‌بندی خطای رده‌بندی، فرض‌هایی که برای شناسایی‌پذیری مدل‌های خطای رده‌بندی نیاز است و همچنین روش‌های برآورد آنها با فرض سه اندازه‌گیری A و B و C روی متغیر X ارائه می‌شوند.

۳ مدل رده‌نهای استاندارد

در این قسمت، حالت سه اندازه‌گیری یا بیش‌تر را بررسی می‌کنیم و به بیان یک ساختار اندازه‌پذیر توسط چند نشان‌گر که برای اندازه‌گیری آن ساختار طراحی شده‌اند، می‌پردازیم. گفته شد که

۱. احتمال‌های خطای اندازه‌گیری گروه‌های مختلف با هم برابرند؛ یعنی

$$\pi_{a|x1}^{A|XG} = \pi_{a|x2}^{A|XG} = \pi_{a|x}^{A|X}, \quad a \neq x$$

$$\pi_{b|x1}^{B|XG} = \pi_{b|x2}^{B|XG} = \pi_{b|x}^{B|X}, \quad b \neq x$$

و

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

$$\pi_{11}^{X|G} \neq \pi_{12}^{X|G} \phi.$$

با توجه به این فرض‌ها، پارامترهای مدل مقید عبارت‌اند از:

$$\pi_1^G, \pi_{11}^{X|G}, \pi_{12}^{X|G}, \pi_{11}^{A|X}, \pi_{12}^{A|X}, \pi_{11}^{B|X}, \pi_{12}^{B|X}$$

همچنین با تحمیل این قیدها، احتمال رده‌بندی یک واحد نمونه‌ای به خانه (g, a, b) به‌شکل زیر درمی‌آید:

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

با جایگذاری این احتمال‌ها در فرمول هسته لگاریتم درست‌نمایی (۴) خواهیم داشت:

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

برآوردهای ماکسیمم درست‌نمایی پارامترها (احتمال‌های خانه‌ای و به‌خصوص احتمال‌های خطا و احتمال شیوع) با استفاده از روش‌های عددی مثل الگوریتم EM به دست می‌آیند.

این‌گونه تحلیل خطای رده‌بندی به‌راحتی به حالتی که متغیر مورد نظر، متغیری چندحالتی است (با فرض دو اندازه‌گیری) تعمیم داده می‌شود. فرض کنید X و در نتیجه A و B هر یک K رسته دارند و متغیر گروه‌بندی، جامعه را به L گروه تقسیم کند. همچنین فرض کنید $\pi_{x|g}^{X|G}$ ، $x = 1, \dots, K$ شیوع واقعی هر رسته در هر یک از زیرگروه‌های جامعه ($\sum_{x=1}^K \pi_{x|g}^{X|G} = 1$) و $\pi_{a|xg}^{A|XG}$ ، $\pi_{b|xg}^{B|XG}$ ، $a \neq x$ ، $b \neq x$ و باشند. در این صورت، هسته لگاریتم درست‌نمایی برای جدول GAB به‌صورت زیر است:

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

رده‌بندی تقاطعی ABC بر حسب X به صورت زیر نوشته می‌شود:

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

مدل (۵) یکی از ساده‌ترین مدل‌های رده‌نهادن است که مدل LC استاندارد نامیده می‌شود.

۴ پارامتری‌سازی مدل رده‌نهادن

برای توصیف یک مدل رده‌نهادن، چهار رهیافت مدل‌بندی یا پارامتری‌سازی را معرفی کرده است [۱]، که عبارت‌اند از: رهیافت پارامتری‌سازی مدل احتمالاتی،^۶ رهیافت مدل لگ خطی^۷ با یک متغیر نهان، رهیافت مدل مسیر تعدیل یافته^۸ و رهیافت مدل نموداری^۹ با استفاده از نمودارهای مسیر. در این بخش فقط به رهیافت پارامتری‌سازی مدل احتمالاتی اشاره می‌شود.

۱.۴ پارامتری‌سازی مدل احتمالاتی LC مدل استاندارد

مدل احتمالاتی، درست‌نمایی جدول رده‌بندی تقاطعی مشاهده‌شده را بر حسب احتمال‌های شرطی و حاشیه‌ای مربوط به هر خانه این جدول بیان می‌کند. مدل احتمالاتی رده‌نهادن، مدل احتمالاتی‌ای است که شامل یک یا چند متغیر نهان باشد.

به عنوان مثال، فرض کنید LCA شامل سه نشان‌گر A و B و C برای متغیر نهان X به ترتیب با تعداد رده‌های k_A, k_B, k_C و k_x باشد (در بیش‌تر مدل‌ها فرض می‌شود $k_A = k_B = k_C = k_x$). یک مدل درست تعیین‌شده برای جدول رده‌بندی تقاطعی ABC باید بدون ابهام این، احتمال این که واحد انتخاب شده نام 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 می‌شود). احتمال خانه (a, b, c) در مدل LC استاندارد، برابر است

LCA چارچوبی برای بررسی رابطه میان متغیر واقعی و متغیرهایی که برای اندازه‌گیری آن طراحی شده‌اند، فراهم می‌آورد. یعنی اگر X متغیر نهان (متغیر واقعی) و A و B و C نشان‌گرهای X (که برای اندازه‌گیری X طراحی شده‌اند) باشند، LCA به بررسی رابطه بین X و A و B و C می‌پردازد.

[۱] با تعمیم روش [۷]، رهیافت LCA را به تفصیل، معرفی کرده است. او با تعریف متغیرهای زیر و تحت شرط‌هایی که اصل‌های رده‌بندی هستند، ساده‌ترین مدل رده‌نهادن را که LC استاندارد نامیده می‌شود، ارائه کرد.

متغیرهای مدل رده‌نهادن به صورت زیر در نظر گرفته می‌شوند: متغیر نهان X : که فرض می‌شود در شرایط آرمانی، اندازه‌پذیر است، ولی در شرایط واقعی با خطا اندازه‌گیری می‌شود، رده‌های نهان: رسته‌های X ،

متغیر آشکار: متغیرهایی مثل A و B و C و G که مشاهده می‌شوند،

متغیر نشان‌گر: متغیر آشکاری که به طور ویژه برای اندازه‌گیری X طراحی شده است (مثل A و B و C).

متغیر گروه‌بندی: متغیر آشکار دیگری که برای شرح پراکندگی متغیر نهان یا احتمال‌های خطا در مدل ظاهر می‌شود (مثل G). این متغیر بر خلاف متغیر نشان‌گر برای اندازه‌گیری X طراحی نشده است.

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

احتمال خطا $\pi_{a|x}^{A|X}, \forall a \neq x$: نوعی احتمال پاسخ است که موردی را که در آنها X و نشان‌گر آن موافق نیستند را در نظر می‌گیرد، و

احتمال درستی $\pi_{a|x}^{A|X}, \forall a = x$: نوعی احتمال پاسخ است که موردی را که در آن رسته متغیر نشان‌گر و رده‌نهادن یکسان‌اند در نظر می‌گیرد.

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

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

تحت شرط استقلال موضعی، هسته تابع درست‌نمایی 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:

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

که در آن n_{xabc} فراوانی خانه (x, a, b, c) جدول تقاطعی XABC است و π_{xabc}^{XABC} احتمال قرارگیری واحد i ام 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.

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 $7 = 2^3 - 1$ است. الگوریتم EM بین گام E و گام M حرکت می‌کند تا ملاک همگرایی تعیین‌شده، برقرار شود. مقدارهای پارامترها در t امین تکرار با $\pi_1^{(t)}$ ، $\pi_{\backslash 1}^{A|X(t)}$ ، $\pi_{\backslash 1}^{B|X(t)}$ و $\pi_{\backslash 1}^{C|X(t)}$ نشان داده می‌شوند.

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

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

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:

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

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 رده نهان 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 مقدارهای ثابت فرض می‌کنند.

۲. دو یا چند احتمال را با هم برابر می‌گیرند.

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 نهان می توانند با استفاده از آماره نیکویی برازش خرد 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 خرد 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:

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

با استفاده از قاعده بیز داریم:

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

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.

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

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

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

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 ام، n_{hpre} اندازه نمونه مقدماتی از طبقه h ، $\hat{p}_{hpre}$ برآورد نسبت متناظر با طبقه h ، e_h حاشیه خطای طبقه h و $Z_{\frac{\alpha}{2}}$ صدک $(1 - \frac{\alpha}{2})$ م 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_3, Y_4 در صورتی که دانشجو پاسخ بله را انتخاب کند، مقدار ۱ و در صورت انتخاب گزینه «خیر»، مقدار «صفر» را اختیار می‌کنند. متغیر Y_7 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

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

جدول ۳. اندازه نمونه هر یک از طبقه‌ها

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

۲.۶. برآورد خطاها با استفاده از LCA

اما جدول ۴ برای انتخاب پرسش مناسب‌تر برای نیل به

هدف تعیین نسبت مشروطی اطلاعاتی به دست نمی‌دهد. با مقایسه پاسخ‌های پرسش‌ها با داده‌های ثبتی (داده‌های کارنامه‌ای (Y_M) ، پرسشی که مطابقت بیش‌تری با رده‌بندی داده‌های ثبتی داشته باشد، مناسب‌تر است (جدول ۵).

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 ثبتی

جدول ۵. جدول رده‌بندی تقاطعی پاسخ‌های دو پرسش و داده‌های

ثبتی Q_3 و Q_7

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

با توجه به جدول ۵ (که در آن داده‌های ثبتی کارنامه‌ای، اندازه‌های استاندارد طلایی، 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 طلایی (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_7 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 $Q3$ (کدبندی‌شده به متغیر دو‌حالتی) و $Q7$ به‌عنوان دو نشان‌گر A و B برای متغیر نهان X (که دست‌کم یک بار مشروط شدن دانشجو را نشان می‌دهد) در نظر گرفته شدند. فرض کنید

$$X = \begin{cases} 1, & \text{اگر دانشجو مشروط شده باشد} \\ 2, & \text{اگر دانشجو مشروط نشده باشد} \end{cases}$$

$$A, B = \begin{cases} 1, & \text{اگر پاسخ مثبت باشد} \\ 2, & \text{اگر پاسخ منفی باشد} \end{cases}$$

با توجه به پاسخ‌های دانشجویان حاضر در نمونه، جدول رده‌بندی تقاطعی AB تشکیل داده شد (جدول ۴). در به‌کارگیری مدل رده‌نهاد 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 نشان‌گر 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
			۱	۲	
۱	A	۱	۱	۱	۲
		۲	۱۰	۱	۱۱
		کل	۱۱	۲	۱۳
۲	A	۱	۰	۳	۳
		۲	۸	۲	۱۰
		کل	۸	۵	۱۳
۳	A	۱	۰	۱	۱
		۲	۱۲	۰	۱۲
		کل	۱۲	۱	۱۳

فرض استقلال ϵ_{iA} و ϵ_{iB} با استفاده از آزمون 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 χ^2 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 به‌ترتیب نشان‌دهنده احتمال‌های خطای مثبت کاذب یعنی احتمال این‌که فردی که در واقعیت مشروط نداشته، در رسته دانشجویان مشروط شده قرار داده شده باشد) و احتمال خطای نادرست‌منفی (یعنی احتمال این‌که فردی که در واقعیت مشروطی داشته باشد 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 $Q3$ و $Q7$ به‌صورت زیر به دست می‌آیند.

$$\hat{\phi}_3 = \frac{1}{1+29} = 0.033, \quad \hat{\phi}_7 = \frac{0}{0+30} = 0, \\ \hat{\theta}_3 = \frac{4}{4+5} = 0.44, \quad \hat{\theta}_7 = \frac{1}{1+8} = 0.11.$$

بر اساس پرسش اول ($Q3$)، حدود ۳ درصد از دانشجویانی که مشروط نشده‌اند، اظهار داشته‌اند که مشروط شده‌اند و حدود ۴۴ درصد از دانشجویانی که دست‌کم 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 ($Q7$)، 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 $Q7$ 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 به‌روش 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 متناظر با پرسش‌ها و مقایسه آنها استفاده می‌شود.

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

بخشی از خروجی برنامه برازش این مدل احتمالاتی که در شکل ۱ آمده است، برآورد احتمال‌های پاسخ شرطی دو نشان گر 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 ۰.۳۷ 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، به عنوان 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 به عنوان 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 به اشتباه اظهار نکنند که دارای یک صفت نامطلوب از نظر اجتماعی هستند و صفت مشروط شدن، خصیصه‌ای نیست که افراد به دنبال 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 Q_7 ، بر پرسش اول ترجیح داده می‌شود.

۷ بحث و نتیجه‌گیری

تحلیل رده نهان (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 نهان استاندارد (۵) (با این تفاوت که در این جا فقط دو نشان گر 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 برآوردی 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)	
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 نشان گر A و B وضعیت مشروطی

بنا بر این مدل شناسایی‌پذیر نیست. در این حالت می‌توان یک متغیر گروه‌بندی G مثل متغیر رشته تحصیلی را به منظور 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 دست‌کم یک درس را افتاده‌اند و نیز خطای رده‌بندی نسبت دانشجویانی که دست‌کم 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.