آماره‌های آزمون و برآورد مؤلفه‌های واریانس مدل خطای پاسخ در آمارگیری‌های حضوری

روشنک علی‌محمدی

گروه آمار– دانشگاه الزهرا تهران

پست الکترونیکی: r_alimohammadi@alzahra.ac.ir

چکیده

در آمارگیری‌ها اعم از سرشماری‌ها و آمارگیری‌های نمونه‌ای امکان بروز انواع خطای غیرنمونه‌گیری وجود دارد. خطای پاسخ بخش مهمی از خطاهای غیرنمونه‌گیری را تشکیل می‌دهد و مدل‌سازی آن دارای اهمیت و کاربردهای فراوانی است. یکی از کاربردهای مدل خطای پاسخ، به‌کارگیری آن در برآورد میزان دقت است. در این مقاله مدل خطای پاسخ در آمارگیری‌ها به شیوهٔ مصاحبهٔ حضوری مورد نظر قرار گرفته و آماره‌هایی برای آزمون اثرهای مدل بر اساس تشکیل جدول تحلیل واریانس ارائه شده و همچنین برآوردگرهایی برای مؤلفه‌های واریانس به‌روش ماکسیمم درستنمایی مقید (در دو حالت) حاصل شده‌اند. به‌عنوان ارائهٔ کاربردی از نتایج حاصل، مجموعه‌هایی از داده‌ها به‌کار گرفته شده و مؤلفه‌های واریانس مدل خطای پاسخ با استفاده از برآوردگرهای ارائه‌شده به‌روش ماکسیمم درستنمایی مقید، محاسبه شده‌اند. یک مسأله در مورد برآوردهای حاصل از روش ماکسیمم درستنمایی مقید، لزوم قرار گرفتن آن‌ها در فضای پارامتر است. لذا برای اطمینان از اعتبار برآوردهای حاصل، از برنامه‌ریزی غیرخطی استفاده شده و نتایج نشان داده است که بهینه‌سازی غیرخطی برآوردهای حاصل از روش ماکسیمم درستنمایی مقید را کاملاً تأیید می‌کند.

واژه‌های کلیدی: *ماکسیمم درستنمایی مقید، برنامه‌ریزی غیرخطی، خطای غیرنمونه‌گیری، خطای پاسخ*

مقدمه

تعیین میزان کیفیت داده‌های حاصل از آمارگیری‌ها به‌دلیل کاربرد آن‌ها در برنامه‌ریزی‌ها و تصمیم‌گیری‌ها از اهمیت زیادی برخوردار است، لذا بررسی انواع خطا در آمارگیری‌ها مورد توجه قرار گرفته است. در اجرای آمارگیری‌های نمونه‌ای امکان بروز انواع خطا اعم از خطای نمونه‌گیری و

خطای غیرنمونه‌گیری وجود دارد. برخلاف خطای نمونه‌گیری که تنها در آمارگیری‌های نمونه‌ای روی می‌دهد، خطای غیرنمونه‌گیری در سرشماری‌ها نیز قابل وقوع است. دقت یکی از مفاهیم اساسی در ارتباط با کیفیت داده‌ها است. به منظور برآورد میزان دقت، لازم است که واریانس نمونه‌گیری و واریانس غیرنمونه‌گیری برآورد شود. گراوز [۱] به معرفی و بررسی انواع خطای غیرنمونه‌گیری پرداخته است. خطای پاسخ بخشی از خطاهای غیرنمونه‌گیری در آمارگیری‌ها است. مدل‌بندی خطای پاسخ دارای کاربردهای متعددی است و به‌همین دلیل مورد توجه خاصی قرار دارد. از جملهٔ این کاربردها به‌کارگیری مدل در برآورد میزان دقت است. در مدل‌بندی به شیوهٔ تحلیل واریانس، مدل خطای پاسخ بر اساس منابع ایجاد این نوع از خطای غیرنمونه‌گیری طراحی می‌شود. باوداز [۲] به بررسی منابع خطای پاسخ بر اساس تکرار اندازه‌گیری‌ها پرداخته و توصیه‌هایی در این مورد ارائه داده است.

کیش [۳] مدلی برای خطای پاسخ شامل دو مؤلفهٔ اثر پرسشگر و اثر سایر عوامل ارائه کرد. هارتلی و رائو [۴] مدل خطای پاسخ با مؤلفه‌های خطای پرسشگر، خطای کدگذار و اثر متقابل بین این عوامل را پیشنهاد کردند.

بی‌مر و تروین [۵] دو مدل عمومی برای خطای اندازه‌گیری در ارتباط با داده‌های پیوسته و داده‌های دوحالتی را بررسی کردند.

آیهان [۶] مدلی برای خطای پاسخ در آمارگیری‌های با دو مرحلهٔ مصاحبه و مصاحبهٔ مجدد را پیشنهاد کرد. در این مدل اثر بازبین، اثر پرسشگر و اثر پاسخگو در نظر گرفته شده است.

علی‌محمدی [۷] با بررسی نحوهٔ اجرای آمارگیری‌ها به شیوهٔ مصاحبهٔ حضوری در ایران، مدلی برای خطای پاسخ ارائه کرده است و بر اساس تشکیل جدول تحلیل واریانس، آماره‌هایی برای آزمون اثرهای مدل خطای پاسخ محاسبه کرده است. در این مقاله مسألهٔ برآورد مؤلفه‌های واریانس مدل خطای پاسخ به‌روش ماکسیمم درستنمایی مقید مورد تحقیق قرار گرفته و برآوردگرهای مؤلفه‌های واریانس در دو حالت محاسبه شده است.

یکی از ویژگی‌های برآوردیابی به روش ماکسیمم درستنمایی مقید، لزوم قرار گرفتن برآوردهای مؤلفه‌های واریانس در فضای پارامتر است. به منظور بررسی عملی این ویژگی، سه مجموعه از داده‌ها را به‌کار گرفته و مؤلفه‌های واریانس مدل خطای پاسخ به روش ماکسیمم درستنمایی مقید برآورد شده است. برای حصول اطمینان از ماکسیمم بودن تابع درستنمایی مورد نظر به‌ازای برآوردهای حاصل، از بهینه‌سازی غیرخطی استفاده شده و ماکسیمم تابع درستنمایی مورد نظر تحت قیود مطلوب محاسبه شده است. برای مطالعه در مورد روش‌های مختلف برآورد مؤلفه‌های واریانس از جمله روش ماکسیمم درستنمایی مقید می‌توان به مراجع [۹ و ۸] و در مورد روش‌های بهینه‌سازی به مراجع [۱۱ و ۱۰] مراجعه کرد.

مدل خطای پاسخ در آمارگیری‌ها

برای ارائهٔ مدل مناسب خطای پاسخ لازم است شرایط اجرای آمارگیری‌ها مورد بررسی قرار گرفته و بر اساس وضعیت اجرای آمارگیری‌ها به مدل‌بندی خطای پاسخ پرداخت. با بررسی نحوهٔ اجرای آمارگیری‌های حضوری مدل خطای پاسخ به‌صورت زیر ارائه شده است:

$$y_{ijkls} = \mu_{ijkls} + A_i + B_{ij} + C_{ijk} + D_{ijkl} + R_{ijkls} \quad (1)$$

که در آن μ_{ijkls} مقدار واقعی، y_{ijkls} مقدار مشاهده‌شدهٔ sامین واحد (در استان iام مربوط به کارشناس مسئول jام و بازبین (و کدگذار) kام و پرسشگر lام)، A_i اثر استان iام، B_{ij} اثر کارشناس مسئول jام در استان iام، C_{ijk} اثر بازبین (و کدگذار) kام در استان iام برای کارشناس مسئول jام، D_{ijkl} اثر پرسشگر lام مربوط به بازبین kام و کارشناس مسئول jام در استان i ام، و R_{ijkls} اثر sامین پاسخگوی منتسب به پرسشگر lام مربوط به بازبین kام و کارشناس مسئول jام در استان iام است.

اندیس $i = 1,...,I \ (=30)$ برای استان‌ها، اندیس $j = 1,...,J_i$ مربوط به کارشناس مسئول، اندیس $k = 1,...,K_{ij}$ مربوط به بازبین (و کدگذار)، اندیس $l = 1,...,L_{ijk}$ مربوط به پرسشگر و اندیس $s = 1,...,S_{ijkl}$ مربوط به پاسخگو است.

در مدل (۱)، اثر پاسخگو در اثر پرسشگر، اثر پرسشگر در اثر بازبین، اثر بازبین در اثر کارشناس مسئول، و تمامی این اثرها در اثر استان به‌صورت آشیانه‌ای هستند. در طرح‌های آماری ایران، کارشناسان مسئول و بازبین‌ها اثرهایی ثابت در مدل دارند و به‌دلیل اجرای این طرح‌ها در کلیهٔ استان‌ها، اثر استان نیز در مدل ثابت است و پرسشگرها و پاسخگویان اثرهایی تصادفی در مدل دارند. بنابراین مدل مورد نظر، مدل تحلیل واریانس آمیختهٔ آشیانی است.

آماره‌های آزمون اثرهای مدل خطای پاسخ

در این بخش آماره‌های آزمون اثرهای مدل (۱) به‌روش تحلیل واریانس ارائه می‌شوند و در بخش بعد برآوردگر مؤلفه‌های واریانس به‌روش ماکسیمم درستنمایی مقید[1] حاصل می‌شوند. در این مدل اثرهای A، B و C ثابت و D و R اثرهای تصادفی هستند. با فرض متعادل بودن داده‌ها در تمامی سطوح (یعنی $J_i = J$ ،$K_{ij} = K$ ،$L_{ijk} = L$ و $S_{ijkl} = S$) و $D_{ijkl} \sim N(0,\sigma_D^2)$ و $R_{ijkls} \sim N(0,\sigma_R^2)$ و برای مقادیر واقعی $\mu_{ijkls} = \mu + M_{ijkls}$ که $M_{ijkls} \sim N(0,\sigma_M^2)$ و همچنین برای اثرهای ثابت A، B و C فرض می‌شود:

$$i = 1,...,I ،\ \sum_k C_{ijk} = 0 ،\ i = 1,...,I, j = 1,...,J$$

$$\sum_{i=1}^{I} A_i = 0 ،\ \sum_{j=1} B_{ij} = 0 ،$$

به‌طور کلی (با توجه به صفر شدن جملات حاصل‌ضرب‌های متقابل)، مجموع توان‌های دوم کل (SST) را می‌توان به‌صورت زیر نوشت:

$$\begin{aligned}
SST &= \sum_i\sum_j\sum_k\sum_l\sum_s (y_{ijkls} - \bar{y}_{.....})^2 \\
&= \sum_i\sum_j\sum_k\sum_l\sum_s ((y_{ijkls} - \bar{y}_{ijkl.}) + (\bar{y}_{ijkl.} - \bar{y}_{ijk..}) + (\bar{y}_{ijk..} - \bar{y}_{ij...}) \\
&\quad + (\bar{y}_{ij...} - \bar{y}_{i....}) + (\bar{y}_{i....} - \bar{y}_{.....}))^2 \\
&= \sum_i\sum_j\sum_k\sum_l\sum_s (y_{ijkls} - \bar{y}_{ijkl.})^2 + \sum_i\sum_j\sum_k\sum_l\sum_s (\bar{y}_{ijkl.} - \bar{y}_{ijk..})^2 \\
&\quad + \sum_i\sum_j\sum_k\sum_l\sum_s (\bar{y}_{ijk..} - \bar{y}_{ij...})^2 + \sum_i\sum_j\sum_k\sum_l\sum_s (\bar{y}_{ij...} - \bar{y}_{i....})^2 \\
&\quad + \sum_i\sum_j\sum_k\sum_l\sum_s (\bar{y}_{i....} - \bar{y}_{.....})^2 = \sum_i\sum_j\sum_k\sum_l\sum_s (y_{ijkls} - \bar{y}_{ijkl.})^2 \\
&\quad + S\sum_i\sum_j\sum_k\sum_l (\bar{y}_{ijkl.} - \bar{y}_{ijk..})^2 + LS\sum_i\sum_j\sum_k (\bar{y}_{ijk..} - \bar{y}_{ij...})^2 \\
&\quad + LSK\sum_i\sum_j (\bar{y}_{ij...} - \bar{y}_{i....})^2 + LSKJ\sum_i (\bar{y}_{i....} - \bar{y}_{.....})^2 \\
&= SSR + SSD + SSC + SSB + SSA
\end{aligned}$$

که در آن $\bar{y}_{.....}$ میانگین کل مقادیر مشاهده‌شده، $\bar{y}_{i....}$ میانگین مقادیر مشاهده‌شده در استان iام، $\bar{y}_{ij...}$ میانگین

1- REML

مشاهدات در استان iام مربوط به کارشناس مسئول jام، $\bar{y}_{ijk..}$ میانگین مقادیر مشاهده‌شده در استان iام مربوط به کارشناس مسئول jام و بازبین (کدگذار) kام، $\bar{y}_{ijkl.}$ میانگین مشاهدات در استان iام مربوط به کارشناس مسئول jام، بازبین (کدگذار) kام و پرسشگر lام است.

تعریف می‌کنیم: $MSA = \dfrac{SSA}{I-1}$، در این صورت با بسط جملات SSA بر اساس مدل (۱) و صرف‌نظر از جزییات اثبات، امید ریاضی MSA به صورت زیر حاصل می‌شود:

$$E(MSA) = \frac{JKLS}{I-1}\sum_{i=1}^{I} A_i^2 + \sigma_M^2 + S\sigma_D^2 + \sigma_R^2$$

$MSB = \dfrac{SSB}{I(J-1)}$ میانگین توان‌های دوم عامل B است و محاسبهٔ امید ریاضی آن نتیجهٔ زیر را به‌دست می‌دهد:

$$E(MSB) = \frac{LSK}{I(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J} B_{ij}^2 + \sigma_M^2 + S\sigma_D^2 + \sigma_R^2$$

و $MSC = \dfrac{SSC}{IJ(K-1)}$، که با صرف‌نظر از جزییات محاسبات، امید ریاضی آن به‌صورت زیر حاصل می‌شود:

$$E(MSC) = \frac{LS}{IJ(K-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} C_{ijk}^2 + \sigma_M^2 + S\sigma_D^2 + \sigma_R^2$$

$MSD = \dfrac{SSD}{IJK(L-1)}$ و ثابت شده است که امید ریاضی آن به‌صورت زیر است:

$$E(MSD) = \sigma_M^2 + S\sigma_D^2 + \sigma_R^2$$

$MSR = \dfrac{SSR}{IJKL(S-1)}$ و امید ریاضی آن به‌صورت زیر محاسبه می‌شود:

$$E(MSR) = \sigma_M^2 + \sigma_R^2$$

بر اساس این محاسبات، جدول تحلیل واریانس مدل خطای پاسخ به‌صورت جدول ۱ حاصل می‌شود:

جدول ۱- جدول تحلیل واریانس مدل (۱)

منبع تغییرات	درجهٔ آزادی	مجموع توان‌های دوم SS	امید ریاضی MS
A	$I-1$	$LSKJ\sum_i(\bar{y}_{i....}-\bar{y}_{.....})^2$	$\frac{JKLS}{I-1}\sum_{i=1}^{I}A_i^2+\sigma_M^2+S\sigma_D^2+\sigma_R^2$
B	$I(J-1)$	$LSK\sum_i\sum_j(\bar{y}_{ij...}-\bar{y}_{i....})^2$	$\frac{LSK}{I(J-1)}\sum_i\sum_j B_{ij}^2+\sigma_M^2+S\sigma_D^2+\sigma_R^2$
C	$IJ(K-1)$	$LS\sum_i\sum_j\sum_k(\bar{y}_{ijk..}-\bar{y}_{ij...})^2$	$\frac{LS}{IJ(K-1)}\sum_i\sum_j\sum_k C_{ijk}^2+\sigma_M^2+S\sigma_D^2+\sigma_R^2$
D	$IJK(L-1)$	$S\sum_i\sum_j\sum_k\sum_l(\bar{y}_{ijkl.}-\bar{y}_{ijk..})^2$	$\sigma_M^2+S\sigma_D^2+\sigma_R^2$
R	$IJKL(S-1)$	$\sum_i\sum_j\sum_k\sum_l\sum_s(y_{ijkls}-\bar{y}_{ijkl.})^2$	$\sigma_M^2+\sigma_R^2$

برای آزمون اثر A، فرض صفر $H_0: A=0$ به‌کار می‌رود. با توجه به جدول ۱ آمارهٔ این آزمون عبارت است از $F_0 = \dfrac{MSA}{MSD}$ و این آماره تحت فرض صفر دارای توزیع F با $((I-1),IJK(L-1))$ درجهٔ آزادی است.

اثر B توسط فرض صفر $H_0: B=0$ و آمارهٔ $F_0=\frac{MSB}{MSD}$ آزمون می‌شود. این آماره تحت H_0 دارای توزیع F با $(I(J-1),IJK(L-1))$ درجه آزادی است.

آزمون اثر ثابت C توسط فرض صفر $H_0: C=0$ و آمارهٔ آزمون $F_0=\frac{MSC}{MSD}$ صورت می‌گیرد. این آماره تحت H_0 دارای توزیع F با $(IJ(K-1),IJK(L-1))$ درجهٔ آزادی است.

آزمون اثر تصادفی D توسط فرض صفر $H_0: \sigma_D^2=0$ و آمارهٔ آزمون $F_0=\frac{MSD}{MSR}$ صورت می‌گیرد. این آماره تحت H_0 دارای توزیع F با $(IJK(L-1),IJKL(S-1))$ درجهٔ آزادی است.

در صورتی می‌توان اثر پاسخگو را آزمون کرد که تکرار اندازه‌گیری‌ها قابل انجام باشد.

برآورد مؤلفه‌های واریانس مدل خطای پاسخ

در این بخش برآوردگرهای مؤلفه‌های واریانس مدل خطای پاسخ به‌روش ماکسیمم درستنمایی مقید ارائه می‌شود. دلیل استفاده از این روش، اجتناب از حصول برآوردهای منفی برای واریانس است.

در صورتی که داده‌ها متعادل باشند، تابع درستنمایی مقید مربوط به مدل (۱) را می‌توان به‌صورت زیر نوشت:

$$L(\sigma_R^2,\sigma_D^2 \mid SSR,SSD)=\frac{\exp(-1/2(SSR/t+SSD/h))}{(2\pi)^{(N-IJK)/2}\,t^{IJKL(S-1)/2}\,h^{IJK(L-1)/2}\,(SL)^{IJK/2}} \tag{2}$$

که در آن $t=\sigma_R^2+\sigma_M^2$ و $h=t+S\sigma_D^2$ هستند و N اندازهٔ نمونه است. لگاریتم این تابع به‌صورت زیر حاصل می‌شود:

$$l=\frac{-(N-IJK)}{2}\log(2\pi)-\frac{1}{2}IJKL(S-1)\log(t)-\frac{1}{2}IJK(L-1)\log(h)-\frac{1}{2}(SSR/t+SSD/h)-\frac{1}{2}IJK\log(SL) \tag{3}$$

و مشتق‌های مرتبهٔ اول l نسبت به t و h را می‌توان به‌صورت زیر نوشت:

$$d(l)/d(t)=\frac{-IJKL(S-1)}{2t}+\frac{SSR}{2t^2}$$

$$d(l)/d(h)=\frac{-IJK(L-1)}{2h}+\frac{SSD}{2h^2}$$

با برابر صفر قرار دادن این مشتق‌های جزیی، ریشه‌های آن به‌صورت $\hat{t}=SSR/(IJKL(S-1))=MSR$ و $\hat{h}=SSD/(IJK(L-1))=MSD$ حاصل می‌شود و در نتیجه $\hat{\sigma}_D^2=(MSD-MSR)/S$ به‌دست می‌آید. برای آن که این مقادیر برآوردگرهای ماکسیمم درستنمایی مقید مؤلفه‌های واریانس σ_D^2 و t باشند، لازم است در فضای پارامتر قرار داشته باشند، یعنی $\hat{\sigma}_D^2 \geq 0$ و $\hat{t}>0$ باشد. بدین منظور علامت برآوردگرهای حاصل برای مؤلفه‌های واریانس $(\hat{t},\hat{\sigma}_D^2)$ را بررسی می‌کنیم، با توجه به نتایج قبل داریم:

$$\hat{t}=MSR$$

که همواره مثبت است و در صورتی که $MSD \geq MSR$ باشد:

$$\hat{\sigma}_D^2=(MSD-MSR)/S$$

نیز نامنفی است.

یک راه مواجهه با حالتی که $MSD<MSR$ باشد، قرار دادن $\hat{\sigma}_D^2=0$ است، بنابراین $h=t$ می‌شود و به‌جای h در رابطهٔ (۲)، t را قرار می‌دهیم. در این صورت لگاریتم تابع درستنمایی حاصل عبارت است از:

$$l' = \frac{-(N-IJK)}{2}\log(2\pi) - \frac{1}{2}IJKL(S-1)\log(t)$$
$$-\frac{1}{2}IJK(L-1)\log(t) - \frac{1}{2}(SSR/t + SSD/t) - \frac{1}{2}IJK\log(SL)$$

مشتق جزیی l' نسبت به t را می‌توان به‌صورت زیر نوشت:

$$d(l')/d(t) = \frac{-IJKL(S-1)}{2t} - \frac{IJK(L-1)}{2t} + \frac{SSR+SSD}{2t^2}$$

بنابراین ریشهٔ این مشتق، $\hat{t} = \dfrac{SSR+SSD}{IJK\,(LS-1)}$ است.

اگر $(\hat{t},\hat{h})$ مختصات نقطهٔ ماکسیمم تابع درستنمایی مورد بررسی باشد، همان برآوردگرها به‌روش ماکسیمم درستنمایی مقید برای (t,h) هستند. لذا مسألهٔ ماکسیمم بودن تابع درستنمایی به‌ازای ریشهٔ مشتق‌های جزیی یعنی $(\hat{t},\hat{h})$، را به‌صورت تحلیلی بررسی می‌کنیم.

رابطهٔ (۲) تابع درستنمایی مورد استفاده در روش ماکسیمم درستنمایی مقید را به‌دست می‌دهد و رابطهٔ (۳) لگاریتم آن را نشان می‌دهد. طبق قضیهٔ آزمون مشتق دوم توابع دومتغیره در حسابان برای آن که $(\hat{t},\hat{h})$ مختصات نقطهٔ ماکسیمم باشد، لازم است دو شرط زیر برقرار باشد.

۱- حداقل یکی از مشتق‌های مرتبهٔ دوم تابع l به‌ازای $(\hat{t},\hat{h})$ منفی شود. برای تابع مورد نظر داریم:

$$d^2(l)/d(t^2)\Big|_{(\hat{t},\hat{h})} = \frac{-IJKL(S-1)}{2t^2} - \frac{SSR}{t^3}\Big|_{(\hat{t},\hat{h})} = \frac{-IJKL(S-1)}{2(MSR)^2} \tag{4}$$

و

$$d^2(l)/d(h^2)\Big|_{(\hat{t},\hat{h})} = \frac{IJK(L-1)}{2h^2} - \frac{SSD}{h^3}\Big|_{(\hat{t},\hat{h})} = \frac{-IJK(L-1)}{2(MSD)^2} \tag{5}$$

و همواره $1 < S$، بنابراین شرط ۱ برقرار است.

۲- دترمینان ماتریس هسین l (رابطهٔ زیر) به ازای $(\hat{t},\hat{h})$ اکیداً مثبت شود:

$$((d^2(l)/d(t^2))(d^2(l)/d(h^2)) - (d^2(l)/d(t)d(h))^2\Big|_{(\hat{t},\hat{h})} = \left(\frac{-IJKL(S-1)}{2(MSR)^2}\right)\left(\frac{-IJK(L-1)}{2(MSD)^2}\right)$$

که این شرط نیز برقرار است. بنابراین l در نقطه $(\hat{t},\hat{h})$ ماکسیمم است، یعنی $(\hat{t},\hat{h})$ برآوردگرهای ماکسیمم درستنمایی مقید برای (t,h) است و بنا بر خاصیت ناوردایی برآوردگرهای ماکسیمم درستنمایی مقید، $(\hat{t},\hat{\sigma}_D^2)$ برآوردگرهای ماکسیمم درستنمایی مقید برای (t,σ_D^2) حاصل می‌شوند.

در صورتی که داده‌ها نامتعادل باشند و تنها در سطح آخر متعادل باشند، یعنی اندیس‌ها مانند مدل (۱) باشد و تنها برای سطح آخر رابطهٔ $S_{ijkl} = S$ برقرار باشد، نتایجی همانند روابط بالا حاصل می‌شود. در این صورت رابطه (۲) به‌صورت زیر درمی‌آید:

$$L(\sigma_R^2,\sigma_D^2\,|SSR,SSD) = \frac{\exp(-1/2(SSR/t+SSD/h))}{2\pi^{(N-K_{..})/2}\,t^{L_{...}(S-1)/2}\,h^{(L_{...}-K_{..})/2}\,c} \tag{6}$$

که در آن $L_{...}$ تعداد کل پرسشگرها مربوط به کلیهٔ بازبین‌ها و کارشناسان مسئول در تمامی استان‌ها است و $K_{..}$ تعداد کل بازبین‌ها در آمارگیری مورد نظر، و c مقدار ثابت است. لگاریتم این تابع به‌صورت زیر است:

$$l_1 = \frac{-(N-K_{..})}{2}\log(2\pi) - \frac{1}{2}L_{...}(S-1)\log(t) - \frac{1}{2}(L_{...}-K_{..})\log(h) - \frac{1}{2}(SSR/t+SSD/h) - \log(c) \tag{7}$$

مشتق مرتبه‌ی اول l_1 نسبت به t و h عبارتند از:

$$d(l_1)/d(t) = \frac{-L_{...}(S-1)}{2t} + \frac{SSR}{2t^2}$$
$$d(l_1)/d(h) = \frac{-(L_{...}-K_{..})}{2h} + \frac{SSD}{2h^2}$$

بنابراین ریشهٔ مشتق‌های جزیی رابطه (۷) به‌صورت

$\hat{t} = SSR/(L_{...}(S-1)) = MSR$

و $\hat{h} = SSD/(L_{...}-K_{..}) = MSD$ است. در نتیجه $\hat{\sigma}_D^2 = (MSD - MSR)/S$ به‌دست می‌آید. (بدیهی است که در این حالت درجات آزادی SSD و SSR متفاوت با حالت متعادل است، لذا MSD و MSR نیز متفاوتند.)

در این حالت نیز برای اطمینان از این‌که ریشهٔ مشتق‌های جزیی، برآوردگرهای ماکسیمم درستنمایی مقید هستند، شرایط آزمون مشتق دوم توابع دو متغیره را در نقطه $(\hat{t},\hat{h})$ بررسی نموده‌ایم. نتایج حاصل نشان می‌دهد که شرایط مورد نظر برقرارند و $(\hat{t},\hat{h})$ مختصات نقطهٔ ماکسیمم است. زیرا:

$$d^2(l)/d(t^2)\Big|_{(\hat{t},\hat{h})} = \frac{L_{...}(S-1)}{2t^2} - \frac{SSR}{t^3}\Big|_{(\hat{t},\hat{h})}$$
$$= \frac{L_{...}(S-1)}{2(MSR)^2} - \frac{SSR}{(MSR)^3} = -\frac{L_{...}(S-1)}{2(MSR)^2}$$

و

$$d^2(l)/d(h^2)\Big|_{(\hat{t},\hat{h})} = \frac{L_{...}-K_{..}}{2h^2} - \frac{SSD}{h^3}\Big|_{(\hat{t},\hat{h})}$$
$$= \frac{L_{...}-K_{..}}{2(MSD)^2} - \frac{SSD}{(MSR)^3} = -\frac{L_{...}-K_{..}}{2(MSD)^2}$$

که شرط اول برقرار است. برای بررسی شرط دوم، دترمینان ماتریس هسین به‌صورت زیر است:

$$((d^2(l)/d(t^2))(d^2(l)/d(h^2)) - (d^2(l)/d(t)d(h))^2\Big|_{(\hat{t},\hat{h})}$$
$$= (\frac{-L_{...}(S-1)}{2(MSR)^2})(\frac{-(L_{...}-K_{..})}{2(MSD)^2})$$

که همواره مقداری مثبت است. بنابراین نتیجه می‌شود $(\hat{t},\hat{h})$ برآوردگرهای ماکسیمم درستنمایی مقید برای (t,h) هستند و بنا بر خاصیت ناوردایی برآوردگرهای ماکسیمم درستنمایی مقید، $(\hat{t},\hat{\sigma}_D^2)$ برآوردگرهای ماکسیمم درستنمایی مقید برای (t,σ_D^2) هستند.

چنان‌که ملاحظه می‌شود، تنها شرط برای به‌کارگیری نتایج حالت اخیر متعادل بودن داده‌ها در سطح آخر است. به‌عبارت دیگر با اختصاص تعداد برابر پاسخگو به پرسشگران می‌توان از نتایج حاصل استفاده کرد. این فرض محدودکننده‌ای نیست و اکثر اوقات در آمارگیری‌ها برقرار است.

بدیهی است که برآوردگرهای مؤلفه‌های واریانس به‌روش ماکسیمم درست‌مایی مقید باید در فضای پارامتر قرار داشته باشند، یعنی نامنفی باشند. چنان‌که ملاحظه می‌شود $\hat{t}$ همواره مثبت است و در صورتی که $MSD \geq MSR$ باشد، $\hat{\sigma}_D^2$ نیز نامنفی است. در صورتی که $MSD < MSR$ باشد، با استفاده از برنامه‌ریزی غیرخطی مسألهٔ یافتن ماکسیمم تابع درستنمایی مورد نظر در فضای پارامتر را بررسی می‌کنیم. هدف از استفاده از بهینه‌سازی غیرخطی، ماکسیمم‌یابی تابع درستنمایی مورد نظر تحت قیود $t > 0$ و $\sigma_D^2 \geq 0$ است، در این صورت برآورد ماکسیمم درستنمایی مقید در ناحیهٔ $t > 0$ و $\sigma_D^2 \geq 0$ حاصل می‌شود که همان محدودیت‌های فضای پارامتر است. تابع هدف همان لگاریتم تابع درستنمایی مقید (رابطهٔ (۳)) است که مختصات نقطهٔ ماکسیمم آن در ناحیهٔ مورد نظر ($t > 0$ و $\sigma_D^2 \geq 0$) برآوردگرهای ماکسیمم درستنمایی مقید مؤلفه‌های واریانس را به‌دست می‌دهد.

در بخش بعد، برآورد ماکسیمم درستنمایی مقید مؤلفه‌های واریانس برای چند مجموعه از داده‌ها با استفاده از برنامه‌ریزی غیرخطی محاسبه می‌شود.

بر اساس نتایج حاصل در این بخش، روش ماکسیمم درستنمایی مقید منجر به جواب‌های تحلیلی برای برآورد مؤلفه‌های واریانس مدل خطای پاسخ شد. به‌طور کلی در صورتی که جواب تحلیلی به‌روش ماکسیمم درستنمایی مقید برای برآورد مؤلفه‌های واریانس قابل حصول نباشد، می‌توان از روش‌های عددی استفاده کرد. یکی از متداول‌ترین روش‌های عددی روش نیوتن- رافسون است. در این روش معادلات تکرار از رابطه‌ی زیر به‌دست می‌آیند:

$$\theta^{(m+1)} = \theta^{(m)} - (H^{(m)})^{-1} \nabla l^{(m)} \qquad (۸)$$

که در آن θ بردار پارامترهای مورد برآورد، $H^{(m)}$ ماتریس هسین و $\nabla l^{(m)}$ بردار گرادیان تابع هدف هستند که در آن‌ها به‌جای θ، مقدار $\theta^{(m)}$ قرار گرفته است. (m نشان‌دهندهٔ شمارهٔ گام است.)

برای اجتناب از برآوردهای منفی برای مؤلفه‌های واریانس در روش نیوتن- رافسون و یا هر روش دیگر که بر اساس تکرار باشد، می‌توان از روش جنریچ و شلاکتر (۱۹۸۶) استفاده کرد [۱۲]. در این روش در صورتی که در یک گام از تکرار، مقادیر برآورد منفی باشد از آن‌ها برای یافتن برآوردهای گام جدید استفاده نمی‌شود و گام جدید را طوری در نظر می‌گیرند که دارای نصف طول گام مورد نظر باشد و پس از آن اگر تمام مقادیر (برآوردهای) حاصل از گام اخیر، مثبت شوند از آن‌ها در معادلات تکرار برای یافتن مقادیر برآورد گام بعد استفاده می‌شود و در غیر این صورت طول گام مجدداً نصف می‌شود و مانند قبل عمل می‌شود و به همین ترتیب ادامه می‌یابد. به‌کارگیری این روش به برآوردهای نامنفی برای مؤلفه‌های واریانس می‌انجامد.

کاربرد

در این بخش مؤلفه‌های واریانس مدل خطای پاسخ به‌روش ماکسیمم درستنمایی مقید و با استفاده از بهینه‌سازی غیرخطی برآورد می‌شوند. در بهینه‌سازی غیرخطی مقدار ماکسیمم تابع تحت قیود مطلوب محاسبه می‌شود، به‌عبارت دیگر ماکسیمم تابع درستنمایی تحت قیود مورد نظر (یعنی در ناحیهٔ $t > ۰$ و $\sigma_D^۲ \geq ۰$) حاصل می‌شود. بدین منظور می‌توان از نرم‌افزار ریاضی ***Maple*** و به‌کارگیری برنامه ***Optimization*** و دستور ***NLPsovle*** که برای بهینه‌سازی توابع غیرخطی است استفاده کرد. با انتخاب گزینهٔ ***assume=nonnegative*** و ***maximize*** مقدار ماکسیمم تابع مورد نظر در فضای نامنفی برای پارامترها محاسبه می‌شود.

در مورد مجموعه داده‌های مورد بررسی، بهینه‌سازی به‌روش برنامه‌ریزی درجهٔ دوم مکرر[۱] انجام شده است. این روش شباهت‌هایی به‌روش نیوتن دارد. در این روش تابع هدف در هر گام با یک تابع درجهٔ دوم تقریب می‌شود و سپس به روش برنامه‌ریزی درجهٔ دوم مقدار بهینه در آن گام محاسبه می‌شود.

به‌منظور بررسی نحوهٔ برآورد مؤلفه‌های واریانس با استفاده از برنامه‌ریزی غیرخطی در حالت $MSD < MSR$، سه مجموعه از داده‌های نوعی مربوط به درآمد کل، هزینهٔ خالص و هزینه‌ی ناخالص خانوار که در آن‌ها

1- Sequential Quadratic Programming

$MSD < MSR$ است را در نظر گرفته‌ایم. (در این داده‌ها $n=۱۱۸۰$، $I=۱$، $J=۱$، $K=۲$، $L=۲$ و $S=۲۹۵$ که این مقادیر در محاسبات مربوط به برآورد مؤلفه‌های واریانس مورد استفاده قرار می‌گیرند).

رابطه (۲) تابع درستنمایی مورد نظر را نشان می‌دهد که در بهینه‌سازی غیرخطی همان تابع هدف است و مختصات نقطه‌ی ماکسیمم این تابع در ناحیه‌ی $t > 0$ و $\sigma_D^۲ \geq ۰$، برآوردهای ماکسیمم درستنمایی مقید برای مؤلفه‌های واریانس مدل خطای پاسخ را به‌دست می‌دهند.

برای داده‌های مربوط به درآمد کل خانوار، $SSR = ۲۲۲۶۶۴۱/۲۳۴$، $MSR = ۱۸۹۳/۳۶۷$ و $SSD = ۲۲۳/۹۴۸$، $MSD = ۱۱۱/۹۷۴$، لذا رابطه‌ی $MSD < MSR$ برقرار است و محاسبه‌ی برآورد مؤلفه‌های واریانس به‌روش ماکسیمم درستنمایی مقید با استفاده از بهینه‌سازی غیرخطی، $\hat{t} = ۱۸۹۰/۳۷۷۹$ و $\hat{\sigma}_D^۲ = ۰$ را نتیجه می‌دهد. به‌عبارت دیگر با استفاده از برنامه‌ریزی غیرخطی، مختصات نقطه‌ی ماکسیمم لگاریتم تابع درستنمایی مورد نظر در ناحیه‌ی $t > ۰$ و $\sigma_D^۲ \geq ۰$ محاسبه شده است که همان برآورد مؤلفه‌های واریانس هستند.

برای مجموعه داده‌های نوعی مربوط به هزینه‌ی خالص خانوار، $SSR = ۲۶۹۹۳۰۴/۴۱$، $MSR = ۲۲۹۵/۳۲۷$ و $SSD = ۲۹۲۸/۳۲۲$، $MSD = ۱۴۶۴/۱۶$ است. بنابراین با توجه به این‌که $MSD < MSR$ و با به‌کارگیری بهینه‌سازی غیرخطی، برآورد ماکسیمم درست‌نمایی مقید مؤلفه‌های واریانس به‌صورت $\hat{t} = ۲۲۹۳/۹۱۵۷$ و $\hat{\sigma}_D^۲ = ۰$ محاسبه می‌شوند.

در داده‌های نوعی مربوط به هزینهٔ ناخالص خانوار، $SSR = ۳۰۴۲۰۹۷/۴۵۱$، $MSR = ۲۵۸۶/۸۱۷$ و $SSD = ۳۶۲۹/۶۱۷$، $MSD = ۱۸۱۴/۸۰۸$ و رابطهٔ $MSD < MSR$ برقرار است. برآورد ماکسیمم درستنمایی مقید مؤلفه‌های واریانس با استفاده از بهینه‌سازی غیرخطی، $\hat{t} = ۲۵۸۵/۵۰۶۸$ و $\hat{\sigma}_D^۲ = ۰$ حاصل می‌شوند.

محاسبات و بررسی‌های انجام شده نشان می‌دهد که در هر سه مورد، تابع درستنمایی مورد نظر ماکسیمم خود را دقیقاً در نقطه‌ای با مختصات ($\sigma_D^۲ = ۰$ و $t = \frac{SSR + SSD}{IJK\,(LS - ۱)}$) می‌گیرد. یعنی $\hat{\sigma}_D^۲ = ۰$ و $\hat{t} = \frac{SSR + SSD}{IJK\,(LS - ۱)}$ برآوردگرهای $\sigma_D^۲$ و t با استفاده از بهینه‌سازی غیرخطی هستند. چنان‌که ملاحظه می‌شود این نتایج با آنچه در بخش قبل برای مواجهه با حالت $MSD < MSR$ ارائه شد، کاملاً مطابقت دارد و نتایج حاصل از روش ماکسیمم درستنمایی مقید بخش ۴ را تأیید می‌کند. این محاسبات نشان می‌دهد که به‌کارگیری برآوردگرهای حاصل به‌روش ماکسیمم درستنمایی مقید که در بخش قبل ارائه شد، به جواب‌های کاملاً قابل قبولی برای برآورد مؤلفه‌های واریانس مدل خطای پاسخ می‌انجامد.

نتیجه‌گیری

مدل‌بندی خطای پاسخ دارای کاربردهای متعددی است و به همین دلیل مورد توجه خاصی قرار دارد. یکی از کاربردهای مدل خطای پاسخ در برآورد میزان دقت آماره‌های حاصل از آمارگیری است. در این مقاله با بررسی مدل خطای پاسخ، آماره‌های آزمون اثرهای مدل خطای پاسخ مورد بررسی قرار گرفته و ارائه شده‌اند و چنان‌که ملاحظه شد برای برآورد اثر

پاسخگو لازم است که مصاحبه‌ها با تکرار انجام شود. همچنین برآوردگرهای مؤلفه‌های واریانس مدل خطای پاسخ به‌روش ماکسیمم درستنمایی مقید محاسبه شده‌اند و در هر مورد جواب تحلیلی مورد نظر ارائه شده است. علاوه بر این ریشه‌های مشتق لگاریتم تابع درستنمایی مورد نظر بررسی شده و به‌طور تحلیلی نشان داده شده است که به‌ازای این ریشه‌ها تابع درستنمایی، ماکسیمم مقدار خود را می‌گیرد و بنابراین جواب‌های حاصل، برآوردگرهای مؤلفه‌های واریانس به‌روش ماکسیمم درستنمایی مقید هستند.

در صورتی که $MSD < MSR$ (یعنی تغییرپذیری بین پرسشگران کم‌تر از تغییرپذیری بین پاسخگویان باشد)، برآورد واریانس اثر پرسشگر منفی می‌شود، در این حالت برای یافتن مختصات نقطه‌ی ماکسیمم تابع درستنمایی (یا برآوردهای مؤلفه‌های واریانس)، از برنامه‌ریزی غیرخطی استفاده شده است و برای مجموعه داده‌های مورد بررسی ملاحظه می‌شود که نتایج حاصل از بهینه‌سازی غیرخطی با نتایج حاصل از روش ماکسیمم درستنمایی مقید کاملاً مطابقت دارد. لذا استفاده از برآوردگرهای حاصل از روش ماکسیمم درستنمایی مقید برای برآورد مؤلفه‌های واریانس مدل خطای پاسخ توصیه می‌شود.

مراجع

[1] Groves, R.M., Survey Errors and Survey Costs, Wiley, N.Y. (2004).

[2] Bavdaz, M., "The Response Process in Recurring Business Surveys", Proceeding of European Conference on Quality in Survey Statistics, (2006).

[3] Kish, L., "Studies of Interviewer Variance for Attitudinal Variables", Journal of American Statistical Association 57 (1962) 92-115.

[4] Hartley, H.O. and Rao, J.N.K., "Estimation of Nonsampling Variance Components in Sample Surveys", In: Current Topics in Surveys Sampling, (Eds: D. Krowski, R. Platek and J.N.K. Rao), Academic Press, N.Y. (1978).

[5] Biemer, P.P. and Trewin, D., "A Review of Measurement Error Effects on the Analysis of Survey Data", In: Survey Measurement and Process Quality, Edited by L. Lyberg, P. Biemer, M. Collins, E. Deleeuw, C. Dippo, N. Schwartz and D. Trewin, Wiley, N.Y. (1997).

[6] Ayhan, H.O., "Models of Response Error Components in Supervised Interview-Reinterview Surveys", Journal of Applied Statistics, 3, 9 (2003)1047-1054.

[۷] علی‌محمدی، روشنک، *مدل‌بندی خطای پاسخ در آمارگیری‌ها*، رسالهٔ دکتری، دانشگاه تربیت مدرس، ۱۳۸۴.

[8] Searle, S.R., Casella, G. and McCulloch, C. E., Variance Components, Wiley, N.Y. (1992).

[9] Rao, S.R.S., Variance Components: Mixed Models, Methodologies and Applications, Chapman and Hall, UK (1999).

[10]Nocedal, J. and Wright, S.J., Numerical Optimization, Springer, USA (2006).

[11]Raszczynski, A., Nonlinear Optimization, Princeton University Press, USA (2006).

[12]Jenrich, R.J. and Schluchter, M.D., “Unbiased Repeated Measures Models with Structural Covariance Matrices”,Biometrics 42 (1986) 805-820.