گزیده‌مطالب آماری
سال ۲۰، شماره‌ی ۱، بهار و تابستان ۱۳۸۸، صص ۱۵۳-۱۹۱

جانهی چندگانه‌ی غیر بیزی

جان جورن‌استد[†]

اداره‌ی آمار نروژ

مترجم: زهرا رضایی قهرودی

پژوهشکده‌ی آمار

چکیده: جانهی چندگانه روشی است که به طور مشخص برای براورد واریانس در حضور داده‌های گم‌شده طراحی شده است. فرمول ترکیبی روبین نیازمند آن است که روش جانهی «مناسب» باشد، که اساساً به این معنی است که جانهی‌ها، استخراج تصادفی از یک توزیع پسین در یک چارچوب بیزی باشند. در مؤسسات ملی آمار (NSI's National Statistical Institutes) مانند مرکز آمار نروژ، روش‌هایی که برای جانهی بی‌پاسخی استفاده می‌شوند، مانند بعضی از انواع جانهی‌های بی‌درنگ طبقه‌بندی‌شده، نوعاً غیر بیزی هستند. بنا بر این روش جانهی چندگانه‌ی روبین در مؤسسات ملی آمار معتبر نیست و نمی‌تواند استفاده شود. این مقاله به مسئله‌ی تعیین یک فرمول ترکیبی دیگر می‌پردازد که بتواند برای روش‌های جانهی که نوعاً در مؤسسات ملی آمار استفاده می‌شود، به کار رود و رهیافتی برای مطالعه‌ی این مسئله پیشنهاد می‌کند. فرمول‌های ترکیبی دیگری برای مکانیسم‌های خاص پاسخ و روش‌های جانهی از نوع بی‌درنگ تعیین شده‌اند.

[†] Bjϕrnstad, J. F. (2007). Non-Bayesian multiple imputation. *Journal of Official Statistics* **23**, 433-452.

واژگان کلیدی: براورد وایانس؛ آمارگیری نمونه‌ای؛ نمونه‌گیری طبقه‌بندی‌شده؛ رگرسیون لوژستیک؛ بی‌پاسخی، جانهی بی‌درنگ.
دریافت: ۱۳۸۸/۲/۲، پذیرش: ۱۳۸۸/۵/۱۷

۱- مقدمه

جانهی چندگانه روشی است که به طور مشخص برای براورد واریانس در حضور داده‌های گم‌شده طراحی شده و توسط روبین [۲] بسط یافته است. دو مرجع جدید دیگر با بحث‌ها و بررسی‌های بیش‌تر، روبین [۳] و شیفر [۴] هستند. ایده‌ی اصلی این روش ایجاد m مقدار جانهی‌شده برای هر مقدار گم‌شده و ترکیب m مجموعه‌ی داده‌های کامل‌شده به‌وسیله‌ی فرمول ترکیبی روبین برای براورد واریانس است. برای این‌که براوردگر معتبر باشد، جانهی باید سطح مناسبی از تغییرپذیری را نمایش دهد. در اصطلاح روبین، یکی از الزامات روش جانهی «مناسب بودن» است. در مؤسسات ملی آمار روش‌هایی که برای جانهی بی‌پاسخی استفاده می‌شود، حتی اگر الزام «مناسب بودن» را داشته باشند، خیلی کم این نیاز را تأمین می‌کنند. اما ایده‌ی ایجاد جانهی چندگانه برای اندازه‌گیری عدم حتمیت جانهی و استفاده از آن برای براورد واریانس و محاسبه‌ی بازه‌ی اطمینان هنوز مورد علاقه است. مشکل این است که فرمول ترکیبی روبین حالا دیگر برای جانهی‌های نامناسب معمول که توسط مؤسسات ملی آمار استفاده می‌شود، معتبر نیست. دلیل آن این است که تغییرپذیری در جانهی‌های نامناسب خیلی کم است و باید برای مؤلفه‌ی بین جانهی، وزن بیش‌تری در براورد واریانس در نظر گرفته شود. مشکل دیگر این‌است که مشخص شود این وزن چقدر باید باشد تا استنباط آماری معتبری ارائه دهد و همچنین برای چه نوع مکانیسم‌های بی‌پاسخی و مسایل براورد، امکان تعیین فرمول ترکیب ساده که به پارامترهای نامعلوم بستگی نداشته باشد، وجود دارد. این مقاله رهیافتی برای مطالعه‌ی این مسئله پیشنهاد می‌کند.

در بخش دوم رهیافتی برای تعیین ترکیب مجموعه‌ی داده‌های کامل‌شده از طریق جانهی پیشنهاد می‌شود. بخش سوم سه کاربرد با بی‌پاسخی تصادفی را بیان می‌کند:

الف) براورد میانگین جامعه از روی نمونه‌های تصادفی ساده با استفاده از روش جانهی بی‌درنگ؛

ب) براورد ضریب رگرسیونی در مدل نسبت، با استفاده از جانهی رگرسیونی مانده‌ها؛

پ) براورد ضریب رگرسیونی در رگرسیون خطی ساده با جانهی رگرسیونی مانده‌ها.

بخش چهارم به مسئله‌ی کلی جانهی چندگانه برای نمونه‌های طبقه‌بندی‌شده می‌پردازد. در بخش پنجم از نظریه‌ی بخش چهارم برای نمونه‌های طبقه‌بندی شده با بی‌پاسخی تصادفی درون طبقات استفاده می‌شود که موارد زیر را پوشش می‌دهد:

الف) براورد میانگین جامعه با استفاده از جانهی بی‌درنگ طبقه‌بندی‌شده؛

ب) براورد لگاریتم (نسبت بخت‌ها) در رگرسیون لوژستیک با وجود گم‌شدگی هم در متغیرهای وابسته و هم در متغیرهای تبیینی.

بخش ششم مسئله‌ی استفاده از قاعده‌ی ترکیبی یکسان برای همه‌ی مسائل براورد با یک روش جانهی مشخص و داده‌ها و مدل پاسخ معلوم را بیان می‌کند. یک نتیجه‌ی کلی برای جانهی بی‌درنگ و براورد خطی نیز ارائه شده است.

۲- رهیافتی برای تعیین یک فرمول ترکیبی دیگر برای براورد واریانس در جانهی چندگانه

فرض کنید $s = (1,\ldots,n)$ معرف کل‌نمونه با داده‌های کل‌نمونه‌ی $y = (y_1,\ldots,y_n)$ از مقادیر متغیرهای تصادفی $Y_1,\ldots,Y_n$ باشند. در نمونه‌گیری از جامعه‌ی متناهی تحت یک مدل طرح، شماره‌گذاری مجدد واحدهای انتخاب‌شده، انجام می‌شود و ماهیت تصادفی‌بودن y به‌وسیله‌ی طرح نمونه‌گیری تعیین می‌شود. هدف براورد برخی از پارامترهای θ است. داده‌های مشاهده‌شده که به‌وسیله‌ی $y_{\text{obs}} = \{(y_i : i \in s_r), s_r\}$ نمایش داده می‌شوند، بخش مشاهده‌شده‌ی y و نمونه‌ی پاسخ s_r از اندازه‌ی n_r است.

فرض کنید $\hat{\theta}$ براوردگری براساس داده‌های کل‌نمونه‌ی y باشد که با $\text{var}(\hat{\theta})$ به‌وسیله‌ی $\hat{\text{V}}(y)$ براورد شده است. برای $i \in s - s_r$، y_i^* را به‌وسیله‌ی بعضی روش‌ها جانهی می‌کنیم و y^* را که معرف داده‌های کامل شده است، در نظر می‌گیریم $(y_i : i \in s_r, y_i^* : i \in s - s_r)$. براساس y^*، داریم $\hat{\theta}^* = \hat{\theta}(y^*)$ و $\hat{\text{V}}^* = \hat{\text{V}}(y^*)$.

جانهی چندگانه‌ی m جانهی تکرار شده، منجر به m مجموعه‌ی داده‌های کامل‌شده با m براورد $i = 1,\ldots,m$ و $\hat{\theta}_i^*$، و براوردهای واریانس مربوط به آن‌ها، $\hat{\text{V}}_i^*$، برای $i = 1,\ldots,m$ می‌شود. براورد ترکیب‌شده به‌وسیله‌ی $\bar{\theta}^* = \sum_{i=1}^{m} \hat{\theta}_i^* / m$ تعیین می‌شود.

واریانس داخل جانهی‌ها به‌صورت $\bar{V}^* = \sum_{i=1}^{m} \hat{V}_i^* / m$ و مؤلفه‌ی بین جانهی‌ها به‌وسیله‌ی $B^* = \sum_{i=1}^{m} (\hat{\theta}_i^* - \bar{\theta}^*)^2 / (m-1)$ تعریف می‌شوند. واریانس براوردشده‌ی کل $\bar{\theta}^*$ به‌وسیله‌ی

$$W = \bar{V}^* + (k + \frac{1}{m}) B^*. \quad (1)$$

پیشنهاد می‌شود. همچنین لازم است k را به‌گونه‌ای تعیین کنیم که

$$E(W) = \mathrm{var}(\bar{\theta}^*). \quad (2)$$

روبین [۲] نشان داد که $k = 1$ می‌تواند با جانهی‌های مناسب استفاده شود، که اساساً به این معنی است که مقادیر جانهی‌شده از یک توزیع پسین در یک چارچوب بیزی قرعه‌کشی شوند.

به‌طور کلی، باید عبارت (۲) را تعیین کرد. یک راه آزمایش و انجام آن، استفاده از امید ریاضی دوگانه به شرط y_{obs}، است به‌طوری که $E(W) = E\left\{E(W \mid Y_{\mathrm{obs}})\right\}$ و $\mathrm{var}(\bar{\theta}^*) = E\left\{\mathrm{var}(\bar{\theta}^* \mid Y_{\mathrm{obs}})\right\} + \mathrm{var}\left\{E(\bar{\theta}^* \mid Y_{\mathrm{obs}})\right\}$. نوعاً

$$E(\bar{V}^*) \approx \mathrm{var}(\hat{\theta}) \quad (3)$$

و $E(B^* \mid y_{\mathrm{obs}}) = \mathrm{var}(\hat{\theta}^* \mid y_{\mathrm{obs}})$. بنا بر این به طور تقریبی

$$E(W) = \mathrm{var}(\hat{\theta}) + \left(E(k) + \frac{1}{m} \right) E\ \mathrm{var}(\hat{\theta}^* \mid Y_{\mathrm{obs}}) \quad (4)$$

است. علاوه بر این، $\mathrm{var}\left(\bar{\theta}^* \mid y_{\mathrm{obs}}\right) = \frac{\mathrm{var}(\hat{\theta}^* \mid y_{\mathrm{obs}})}{m}$ و $E\left(\bar{\theta}^* \mid y_{\mathrm{obs}}\right) = E\left(\hat{\theta}^* \mid y_{\mathrm{obs}}\right)$. این حاکی از آن است که $\mathrm{var}(\bar{\theta}^*) = m^{-1} E\{\mathrm{var}(\hat{\theta}^* \mid Y_{\mathrm{obs}})\} + \mathrm{var}\{E(\hat{\theta}^* \mid Y_{\mathrm{obs}})\}$. با استفاده از روابط (۳) و (۴)، معادله‌ی (۲) به‌صورت

$$\mathrm{var}(\hat{\theta}) + E(k)E[\mathrm{var}(\hat{\theta}^* \mid Y_{\mathrm{obs}})] = \mathrm{var}\{E(\hat{\theta}^* \mid Y_{\mathrm{obs}})\}$$

در می‌آید و عبارت کلی زیر را ارائه می‌دهد.

$$E(k) = \frac{\mathrm{var}\,E(\hat{\theta}^* \mid Y_{\mathrm{obs}}) - \mathrm{var}(\hat{\theta})}{E\ \mathrm{var}(\hat{\theta}^* \mid Y_{\mathrm{obs}})} \qquad (۵)$$

اگر مقدار k مورد علاقه‌ی ما باشد باید حد اقل به طور تقریبی، مستقل از پارامترهای نامعلوم تعیین شود. علاوه بر این، لازم است که رابطه‌ی (۳) برقرار باشد. برای تشریح این‌که چگونه رابطه‌ی (۵) می‌تواند استفاده شود، در بخش بعدی سه حالت خاص با بی‌پاسخی تصادفی را مورد بررسی قرار می‌دهیم.

۳- سه کاربرد برای بی‌پاسخی تصادفی

۳-۱- براورد میانگین جامعه با جانهی بی‌درنگ

یک نمونه‌ی تصادفی ساده از جامعه‌ی متناهی به اندازه‌ی N را در نظر می‌گیریم، که هدف آن براورد میانگین جامعه، μ از متغیر y است. همچنین فرض بی‌پاسخی کاملاً تصادفی نیز درنظر گرفته شده است. در اصطلاح‌شناسی روبین [۲] و لیتل و روبین [۱]، مکانیسم گم‌شدگی، گم شدن کاملاً تصادفی (Missing Completely at Random) MCAR نامیده می‌شود. گم‌شدن کاملاً تصادفی به این معنا است که نماگرهای پاسخ $R_1,\ldots,R_N$ با احتمال‌های پاسخ مشابه $\Pr = P(R_i = 1)$ مستقل هستند. روش جانهی، روش بی‌درنگ است که y_i^* به تصادف از y_{obs} با جایگذاری قرعه‌کشی شده است و براورد، میانگین نمونه است. فرض کنید $\bar{y}_r$ میانگین نمونه‌ی مشاهده‌شده و $\hat{\sigma}_r^2 = \frac{1}{n_r - 1}\sum_{i \in s_r}(y_i - \bar{y}_r)^2$ واریانس نمونه‌ی مشاهده‌شده باشد. پس $\bar{Y}^*$ میانگین نمونه بر اساس جانهی برای نمونه‌ی کامل‌شده است، و براوردگر ترکیب‌شده به‌وسیله‌ی $\bar{\bar{Y}}^* = \sum_{i=1}^{m} \bar{Y}_i^* / m$ تعیین می‌شود. فرض کنید $\bar{Y}_s$ معرف میانگین نمونه بر اساس کل‌نمونه باشد. بنا بر این $\mathrm{var}(\bar{Y}_s) = \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right)$ است که در آن

$\sigma^2 = (N-1)^{-1}\sum_{i=1}^{N}(y_i-\mu)^2$ به‌عنوان واریانس جامعه است. همچنین با توجه به این‌که $E(Y_i^*|y_{obs}) = \bar{y}_r$ و $\mathrm{var}(Y_i^*|y_{obs}) = \hat{\sigma}_r^2(n_r-1)/n_r$ است، $E(\bar{Y}^*|y_{obs}) = \bar{y}_r$ و $\mathrm{var}(\bar{Y}^*|y_{obs}) = \{(n-n_r)/n^2\}\{(n_r-1)/n_r\}\hat{\sigma}_r^2$ خواهد بود. در این حالت، $\hat{V}^* = \hat{\sigma}_*^2(\frac{1}{n}-\frac{1}{N})$ است که در آن $\hat{\sigma}_*^2 = \frac{1}{n-1}(\sum_{s_r}(y_i-\bar{y}^*)^2 + \sum_{s-s_r}(y_i^*-\bar{y}^*)^2)$. می‌توان نشان داد که $E(\hat{\sigma}_*^2|y_{obs}) = \hat{\sigma}_r^2\left(1-\frac{1}{n_r}\right)\left(1+\frac{n_r}{n(n-1)}\right) \approx \hat{\sigma}_r^2$ و رابطه‌ی (۳) برقرار است. با استفاده از رابطه‌ی (۵) داریم

$$E(k) = \frac{\mathrm{var}(\bar{Y}_r) - \sigma^2\left(\frac{1}{n}-\frac{1}{N}\right)}{E\left(\frac{n-n_r}{n^2}\cdot\frac{n_r-1}{n_r}\right)E\left(\hat{\sigma}_r^2|n_r\right)}$$

$$= \frac{\sigma^2\left(E\left(\frac{1}{n_r}\right)-\frac{1}{N}\right)-\sigma^2\left(\frac{1}{n}-\frac{1}{N}\right)}{E\left(\frac{n-n_r}{n^2}\cdot\frac{n_r-1}{n_r}\right)\sigma^2} \approx \frac{(1-p_r)/p_r}{1-p_r} = \frac{1}{p_r}$$

که با در نظر گرفتن $f = (n-n_r)/n$ به‌عنوان نرخ بی‌پاسخی و قرار دادن $k = 1/(1-f)$ به طور تقریبی روابط مورد نظر برقرار است.

۲-۳- براورد ضریب رگرسیون در مدل نسبت با جانهی مانده‌ها

فرض می‌کنیم بی‌پاسخی همانند بخش ۱-۳ کاملاً تصادفی است. یک مدل نسبت، یعنی رگرسیونی که از مبدأ می‌گذرد، $Y_i = \beta x_i + \varepsilon_i$، با $\mathrm{var}(\varepsilon_i) = \sigma^2 x_i; i = 1,...,n$ را درنظر می‌گیریم. فرض شده است که همه‌ی x_iها، همچنین در نمونه‌ی بی‌پاسخ

معلوم‌اند . براوردگر داده‌های کامل $\hat{\beta} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} x_i$ به‌وسیله‌ی β، مشخص می‌شود. براوردگر ناریب $\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{1}{x_i}(y_i - \hat{\beta} x_i)^2/(n-1)$ به‌وسیله‌ی σ^2 مشخص می‌شود.

روش جانهی رگرسیون مانده‌ها در نظر گرفته شده است. فرض کنید $\hat{\beta}_r$، براورد $\hat{\beta}$ بر اساس نمونه‌ی مشاهده‌شده‌ی s_r باشد. مانده‌های استاندارد شده را به‌صورت $e_i = (y_i - \hat{\beta}_r x_i)/\sqrt{x_i}$ برای $i \in s_r$ تعریف می‌کنیم. برای $i \in s - s_r$ مقدار e_i^* را به‌تصادف و با جایگذاری از روی مجموعه‌ی مانده‌های مشاهده‌شده‌ی e_i، $i \in s_r$ قرعه‌کشی می‌کنیم. مقدار yی جانهی‌شده به‌وسیله‌ی $y_i^* = \hat{\beta}_r x_i + e_i^* \sqrt{x_i}$ مشخص می‌شود.

فرض کنید $X = \sum_{i=1}^{n} x_i$، $X_r = \sum_{i \in s_r} x_i$ و $X_{nr} = \sum_{i \in s - s_r} x_i = X - X_r$ باشد. همه‌ی فرضیات از اکنون به بعد به شرط n_r و X_r است و هدف ما تعیین مقدار k به طور مستقیم از رابطه‌ی (۵) می‌باشد. نسبت x-کل در گروه بی‌پاسخی به‌وسیله‌ی رابطه‌ی $f_X = X_{nr}/X$ مشخص شده است. بنابر این $\hat{\beta}^* = (\sum_{s_r} y_i + \sum_{s-s_r} y_i^*)/X$ و

$$\hat{\sigma}_*^2 = \frac{1}{n-1}\left(\sum_{s_r} \frac{1}{x_i}\left(y_i - \hat{\beta}^* x_i\right)^2 + \sum_{s-s_r} \frac{1}{x_i}\left(y_i^* - \hat{\beta}^* x_i\right)^2\right).$$

به‌منظور تعیین k از رابطه‌ی (۵)، نیاز به بررسی اعتبار رابطه‌ی (۳) و استخراج $\operatorname{var} E(\hat{\beta}^* | y_{obs})$، $E\operatorname{var}(\hat{\beta}^* | y_{obs})$ و $\operatorname{var}(\hat{\beta})$ است. یادآور می‌شود که $\operatorname{var}(\hat{\beta}) = \sigma^2/X$. در پیوست ۱ نشان داده شده است که شرط (۳) برای مقادیر بزرگ و متوسط n_r برقرار است و این‌که

$$\operatorname{var} E(\hat{\beta}^* | y_{obs}) = \frac{\sigma^2}{X_r} + \frac{(1-d_1)d_2 n_{nr} X_{nr}}{X^2} \cdot \frac{\sigma^2}{n_r} \qquad (۶)$$

$$E\operatorname{var}(\hat{\beta}^* | y_{obs}) = \frac{X_{nr}}{X^2} \cdot \frac{\sigma^2}{n_r}(n_r + d_1 - 2) \qquad (۷)$$

که در آن $0 \le d_1$ ، $d_2 \le 1$. با استفاده از روابط (۶) و (۷) و قرار دادن آن در رابطه‌ی (۵) خواهیم داشت

$$k = \frac{n_r X^2 - n_r X\, X_r + (1-d_1)d_2 n_{nr} X_{nr} X_r}{X_r X_{nr}(n_r + d_1 - 2)} \approx \frac{X}{X_r} + (1-d_1)d_2 \frac{n_{nr}}{n_r}$$

لازم به ذکر است که اگر همه‌ی $x_i = 1$ باشد، بنا بر این $d_1 = d_2 = 1$. اکنون با درنظر گرفتن $f_X = X_{nr}/X$ که نسبت $x-$کل در گروه بی‌پاسخی است و $f = n_{nr}/n$ که نرخ بی‌پاسخی است، چون معمولاً $(1-d_1)d_2 \approx 0$ است، سرانجام رابطه‌ی زیر برای مقادیر معمول x و نرخ‌های بی‌پاسخی به دست می‌آید.

$$k \approx \frac{1}{1-f_X} + (1-d_1)d_2 \frac{f}{1-f} \approx \frac{1}{1-f_X}$$

۳-۳- براورد ضریب رگرسیون در رگرسیون خطی ساده با جانهی مانده‌ها

همانند بخش ۱-۳ و ۲-۳ مکانیسم بی‌پاسخی با $\Pr = P(R_i = 1)$ گم شدن کاملاً تصادفی در نظر گرفته شده است. مدل رگرسیون خطی ساده با $\mathrm{var}(\varepsilon_i) = \sigma^2$ برای $i = 1,\ldots,n$ به‌صورت $Y_i = \alpha + \beta x_i + \varepsilon_i$ فرض شده است. همه‌ی x_iها معلوم فرض شده‌اند. همچنین فرض شده است که $\bar{x} = \sum_{i=1}^{n} x_i / n = 0$. بنا بر این براوردهای بر اساس کل داده‌ها به‌وسیله‌ی $\hat{\beta} = \sum_{i=1}^{n} x_i y_i / SS_x$ مشخص شده است که در آن $SS_x = \sum_{i=1}^{n} x_i^2$ و $\hat{\alpha} = \bar{y} = \sum_{i=1}^{n} y_i / n$. براوردگر ناریب σ^2 به‌وسیله‌ی $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ مشخص می‌شود.

فرض کنید $\hat{\alpha}_r$، $\hat{\beta}_r$ براوردهای براساس نمونه‌ی پاسخ باشند که از رابطه‌ی $\hat{\alpha}_r = \bar{y}_r - \hat{\beta}_r \bar{x}_r$ و $\hat{\beta}_r = \sum_{i \in s_r}(x_i - \bar{x}_r) y_i / SS_{x,r}$ به دست می‌آیند. همچنین $\bar{y}_r = \sum_{i \in s_r} y_i / n_r$، $\bar{x}_r = \sum_{i \in s_r} x_i / n_r$ و $SS_{x,r} = \sum_{i \in s_r}(x_i - \bar{x}_r)^2$.

جانهی مانده‌های ساده به‌صورت زیر تعریف می‌شود؛ مانده‌های مشاهده‌شده برای $j \in s_r$ ، $e_j = (y_j - \hat{\alpha}_r - \hat{\beta}_r x_j)$ است. برای $i \in s - s_r$ به تصادف و با

جایگذاری از $(e_j, j \in s_r)$ قرعه‌کشی شده است. مقدار yی جانهی‌شده به‌وسیله‌ی $y_i^* = \hat{\alpha}_r + \hat{\beta}_r \bar{x}_i + e_i^*$ مشخص می‌شود.

براوردهای بر اساس جانهی عبارت‌اند از $\hat{\beta}^* = (\sum_{i \in s_r} x_i y_i + \sum_{i \in s-s_r} x_i y_i^*)/SS_x$ و $\bar{y}_{nr}{}^* = \sum_{s-s_r} y_i^*/(n-n_r)$ که در آن $\hat{\alpha}^* = (n_r \bar{y}_r + (n-n_r)\bar{y}_{nr}^*)/n$ و $\hat{\sigma}_*^2 = \frac{1}{n-2}\left\{\sum_{s_r}(y_i - \hat{\alpha}^* - \hat{\beta}^* x_i)^2 + \sum_{s-s_r}(y_i^* - \hat{\alpha}^* - \hat{\beta}^* x_i)^2\right\}$. می‌توان نشان داد که (برای خلاصه‌ی اثبات به پیوست ۲ نگاه کنید) $E(\hat{\sigma}_*^2) = \sigma^2 E\left(\frac{n_r - 2}{n_r} \cdot \frac{n-2f}{n-2}\right) \approx \sigma^2$ که همان‌طور که در بخش ۱-۳ نیز اشاره شد $f = (n-n_r)/n$ است. چون $\text{var}(\hat{\beta}) = \sigma^2/SS_x$، بنا بر این رابطه‌ی (۳) برقرار است. به‌سهولت دیده می‌شود که $E(\hat{\beta}^* | y_{\text{obs}}) = \hat{\beta}_r$ و $\text{var}(\hat{\beta}^* | y_{\text{obs}}) = s_e^2 c_r / SS_x$ که در آن $s_e^2 = \sum_{s_r} e_i^2/n_r$ و $c_r = \sum_{s-s_r} x_i^2/SS_x \in <0,1>$. می‌توان نشان داد که $E(s_e^2 | s_r) = \frac{n_r - 2}{n_r}\sigma^2$. همچنین به‌وضوح $E(c_r) = 1 - p_r$ و $\text{var}(\hat{\beta}_r | s_r) = \sigma^2/SS_{x,r}$. از رابطه‌ی (۵) نتیجه می‌شود که

$$E(k) = \frac{E(1/SS_{x,r}) - 1/SS_x}{(1-p_r)E\{(n_r-2)/n_r\}/SS_x} \approx \frac{1/E(SS_{x,r}) - 1/SS_x}{(1-p_r)/SS_x}$$

با استفاده از این حقیقت که به شرط n_r، s_r نمونه‌ای تصادفی ساده است به‌گونه‌ای که نماگرهای پاسخ آن با $\text{cov}(R_i, R_j) = -f(1-f)/(n-1)$ هم‌بسته‌اند، درمی‌یابیم که $E(SS_{x,r}) = \left(p_r - \frac{1-p_r}{n-1}\right)SS_x$. نتیجه می‌شود که به طور تقریبی

$$k = 1/(1-f), \qquad E(k) = \frac{1}{p_r - \frac{1}{n}} \approx \frac{1}{p_r}.$$

۴- جانهی چندگانه برای نمونه‌های طبقه‌بندی‌شده

۴-۱- ترکیب‌های جداگانه

یک روش ترکیب m مجموعه‌ی داده‌های کامل این است که برای هر طبقه این کار به طور جداگانه انجام شود یعنی یک مقدار kی مجزا برای هر طبقه تعیین شود. پس ساختارکلی به‌صورت زیر است؛ نمونه‌ی s به H طبقه‌ی نمونه به‌صورت $s_1,\ldots,s_H$ تقسیم می‌شود. فرض کنید y_h کل داده‌های طرح ریزی شده از زیرنمونه‌ی s_h به‌اندازه‌ی n_h باشد. فرض شده است که $y_1,\ldots,y_H$ مستقل هستند. بخش مشاهده‌شده‌ی y_h با نماد $y_{h,\mathrm{obs}}$ نمایش داده می‌شود که s_{hr} نمونه‌ی پاسخ از s_h به اندازه‌ی n_{hr} است. براوردگر مبتنی بر مجموعه‌ی داده‌های کامل، جمع عبارات مستقل $\hat\theta=\sum_{h=1}^{H}\hat\theta_h$ است که در آن $\hat\theta_h$ بر اساس y_h به دست آمده است. $\mathrm{var}(\hat\theta)=\sum_{h=1}^{H}\mathrm{var}(\hat\theta_h)$ $\hat{\mathrm{V}}(\hat\theta)=\sum_{h=1}^{H}\hat{\mathrm{V}}_h(y_h)$ براورد می‌شود که در آن $\hat{\mathrm{V}}_h(y_h)$ براورد واریانس $\hat\theta_h$ براساس y_h است. برای $i\in s_h-s_{hr}$، y_i^* با استفاده از برخی روش‌ها براساس $y_{h,\mathrm{obs}}$ جانهی می‌شود و y_h^* را به‌عنوان داده‌های کامل درنظر می‌گیریم $(y_{h,\mathrm{obs}},y_i^*,i\in s_h-s_{hr})$. براساس y_h^*، داریم $\hat\theta_h^*=\hat\theta_h(y_h^*)$ و $\hat{\mathrm{V}}_h^*=\hat{\mathrm{V}}_h(y_h^*)$. سپس براوردگر مبتنی بر جانهی به‌وسیله‌ی $\hat\theta^*=\sum_{h=1}^{H}\hat\theta_h^*$ و $\hat{\mathrm{V}}^*=\sum_{h=1}^{H}\hat{\mathrm{V}}_h^*$ مشخص می‌شوند. جانهی چندگانه‌ی m جانهی تکرار شده به m مجموعه‌ی داده‌های کامل‌شده با m براورد برای هر طبقه‌ی h، به‌صورت $\hat\theta_{h,i}$ و براورد واریانس مربوط $\hat{\mathrm{V}}_{h,i}^*$ برای $i=1,\ldots,m$ منجر می‌شود. براوردهای کل و واریانس‌های مربوط به‌صورت $\hat\theta_i^*=\sum_{h=1}^{H}\hat\theta_{h,i}^*$ برای $i=1,\ldots,m$ و $\hat{\mathrm{V}}_i^*=\sum_{h=1}^{H}\hat{\mathrm{V}}_{h,i}^*$ برای $i=1,\ldots,m$ است. براورد ترکیب‌شده برای طبقه‌ی hام به‌وسیله‌ی $\bar\theta_h^*=\sum_{i=1}^{m}\hat\theta_{h,i}^*/m$ مشخص می‌شود. واریانس داخل جانهی برای طبقه‌ی hام عبارت است از $\bar{\mathrm{V}}_h^*=\sum_{i=1}^{m}\hat{\mathrm{V}}_{h,i}^*/m$ و مؤلفه‌ی بین جانهی به‌وسیله‌ی رابطه‌ی $\mathrm{B}_h^*=\sum_{i=1}^{m}(\hat\theta_{h,i}^*-\bar\theta_h^*)^2/(m-1)$ مشخص می‌شود. با دنبال کردن همان ایده‌ی بخش ۲،

رابطه‌ی (۱)، واریانس براوردشده‌ی کل $\bar{\theta}_h^*$، به‌صورت $W_h = \bar{V}_h^* + (k_h + \frac{۱}{m})B_h^*$ پیشنهاد می‌شود. براورد کل ترکیب‌شده نیز به‌وسیله‌ی $\bar{\theta}^* = \sum_{i=۱}^{m} \hat{\theta}_i^* / m = \sum_{h=۱}^{H} \bar{\theta}_h^*$ مشخص می‌شود. نتیجه می‌گیریم که واریانس براوردشده‌ی کل $\bar{\theta}^*$ نیز می‌تواند به‌صورت رابطه‌ی زیر بیان شود.

$$W_{\text{sep}} = \sum_{h=۱}^{H} W_h = \bar{\mathrm{V}}^* + \sum_{h=۱}^{H} (k_h + \frac{۱}{m})B_h^* \qquad (۸)$$

که در آن $\bar{\mathrm{V}}^* = \sum_{i=۱}^{m} \hat{\mathrm{V}}_i^* / m = \sum_{h=۱}^{H} \bar{\mathrm{V}}_h^*$ است. اگر رابطه‌ی (۳) برای هر طبقه‌ی h برقرار باشد

$$E(\bar{V}_h^*) \approx \mathrm{var}(\hat{\theta}_h) \qquad (۹)$$

و با استفاده از رابطه‌ی (۵)، k_h باید در رابطه‌ی زیر صدق کند

$$E(k_h) = \frac{\mathrm{var}\, E(\hat{\theta}_h^* \,|\, Y_{h,\text{obs}}) - \mathrm{var}(\hat{\theta}_h)}{E\ \mathrm{var}(\hat{\theta}_h^* \,|\, Y_{h,\text{obs}})} \qquad (۱۰)$$

فرمول ترکیبی (۸) یک شق دیگر برای فرمول ترکیبی معمول (۱) است که به‌خصوص زمانی سودمند است که به عبارت‌هایی ساده برای k_h و نه برای k دست یافته باشیم. بخش بعدی یک عبارت برای k در این حالت بسط داده است.

۴-۲ - یک فرمول ترکیبی کلی

اکنون فرض کنید W با استفاده از رابطه‌ی (۱) مشخص شده باشد. هدف تعیین فاکتور بین جانهی k است. چون $E(W) = E(W_{\text{sep}})$ است، بنا بر این داریم

$$E\left\{\sum_{h=۱}^{H}\left(k_h + \frac{۱}{m}\right)B_h^*\right\} = E\left(k + \frac{۱}{m}\right)B^* \qquad (۱۱)$$

که در آن، $B^* = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i^* - \bar{\theta}^*)^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left\{\sum_h (\hat{\theta}_{h,i}^* - \bar{\theta}_h^*)\right\}^2$. توجه داشته باشید که چون $E(B^* \mid y_{obs}) = \mathrm{var}(\hat{\theta}^* \mid y_{\mathrm{obs}}) = \sum_{h=1}^{H} \mathrm{var}\left(\hat{\theta}_h^* \mid y_{\mathrm{obs}}\right)$ و همچنین $E(B_h^* \mid y_{\mathrm{obs}}) = \mathrm{var}(\hat{\theta}_h^* \mid y_{\mathrm{obs}})$، بنا بر این

$$E(B^* \mid y_{\mathrm{obs}}) = E(\sum_{h=1}^{H} B_h^* \mid y_{\mathrm{obs}}).$$

بنا بر این اتحاد (۱۱)، به‌صورت

$$E\{\sum_{h=1}^{H} k_h E(B_h^* \mid Y_{\mathrm{obs}})\} = E\{kE(B^* \mid Y_{\mathrm{obs}})\}$$

می‌شود. در صورتی که بخواهیم از فرمول ترکیبی معمول (۱) استفاده کنیم، به حل $k = \sum_{h=1}^{H} k_h E(B_h^* \mid y_{\mathrm{obs}}) / E(B^* \mid y_{\mathrm{obs}})$ دست می‌یابیم و لذا

$$k = \frac{\sum_{h=1}^{H} k_h \,\mathrm{var}(\hat{\theta}_h^* \mid y_{\mathrm{obs}})}{\mathrm{var}(\hat{\theta}^* \mid y_{\mathrm{obs}})} = \sum_{h=1}^{H} k_h \cdot \frac{\mathrm{var}(\hat{\theta}_h^* \mid y_{\mathrm{obs}})}{\mathrm{var}(\hat{\theta}^* \mid y_{\mathrm{obs}})} \qquad (۱۲)$$

که این رابطه یک میانگین موزون از k_h است. زمانی که همه‌ی k_hها برابر باشند، مثلاً $k_h = k_\circ$، به عبارت ساده‌ای برای k یعنی $k = k_\circ$ دست می‌یابیم.

۵- چهار کاربرد برای نمونه‌های طبقه‌بندی‌شده و بی‌پاسخی تصادفی درون طبقات

۵-۱- براورد میانگین جامعه از نمونه‌ی طبقه‌بندی‌شده با جانهی بی‌درنگ طبقه‌بندی‌شده

نمونه‌های تصادفی ساده‌ی طبقه‌بندی‌شده از یک جامعه‌ی متناهی به‌اندازه‌ی N، با H طبقه‌ی به‌اندازه‌ی N_h برای $h = 1,\ldots,H$ را درنظر بگیرید. هدف براورد میانگین جامعه μ از متغیر y است. فرض کنید بی‌پاسخی کاملاً تصادفی، که توسط روبین [۲] و لیتل و روبین [۱] با MAR (گم‌شدگی‌تصادفی) نشان داده شده است، در هر طبقه وجود

داشته باشد. این مطلب به این معنا است که نماگرهای پاسخ در طبقه‌ی hام، $R_{h,1}$، ...، R_{h,N_h} با احتمال‌های $p_{hr} = P(R_{h,i} = ۱)$ از هم مستقل هستند. روش جانهی، روش بی‌درنگ طبقه‌بندی‌شده است. فرض کنید $y_{h,\text{obs}}$ بخش مشاهده‌شده از نمونه‌ی پاسخ s_{hr} به‌اندازه‌ی n_{hr} از طبقه‌ی h باشد، $y_{h,\text{obs}} = (y_i : i \in s_{hr})$. سپس مقدار جانهی‌شده‌ی y_i^* در طبقه‌ی hام به‌تصادف از $y_{h,\text{obs}}$ قرعه‌کشی می‌شود. براوردگر مبتنی بر داده‌های کل نمونه، میانگین موزون طبقه‌بندی‌شده‌ی معمول $\bar{Y}_{\text{strat}} = \sum_{h=۱}^{H} N_h \bar{y}_h / N = \sum_{h=۱}^{H} v_h \bar{y}_h$ است. در این‌جا، $v_h = N_h / N$ و $\bar{y}_h = \sum_{s_h} y_i / n_h$ است که در آن s_h نمونه‌ی طبقه‌ی hام و $n_h = |s_h|$ است. پس $\text{var}(\bar{Y}_{\text{strat}}) = \sum_{h=۱}^{H} v_h^۲ \sigma_h^۲ \left(\frac{۱}{n_h} - \frac{۱}{N_h} \right)$ با $\sigma_h^۲ = \sum_{i \in U_h} (y_i - \mu_h)^۲ / (N_h - ۱)$، واریانس جامعه در طبقه‌ی hام می‌باشد. در این‌جا U_h، جمعیت طبقه‌ی hام و μ_h میانگین در U_h است.

فرض کنید $\bar{y}_{hr}$ میانگین نمونه‌ی مشاهده‌شده از طبقه‌ی hام و $\hat{\sigma}_{hr}^۲ = \frac{۱}{n_{hr} - ۱} \sum_{i \in s_{hr}} (y_i - \bar{y}_{hr})^۲$ واریانس نمونه‌ی مشاهده‌شده باشد. براوردگر مبتنی بر جانهی به‌وسیله‌ی $\bar{Y}_{\text{strat}}^* = \sum_{h=۱}^{H} N_h \bar{y}_h^* / N$ تعیین می‌شود که در آن $\bar{y}_h^* = \left(\sum_{s_{hr}} y_i + \sum_{s_h - s_{hr}} y_i^* \right) / n_h = \left(n_{hr} \bar{y}_{hr} + \sum_{s_h - s_{hr}} y_i^* \right) / n_h$ است. فرض کنید m تکرار جانهی $\bar{Y}_{\text{strat}}^*$ به‌وسیله‌ی $\bar{Y}_{\text{strat},i}^*$ برای $i = ۱, \ldots, m$ نمایش داده شود. براوردگر ترکیب‌شده به‌وسیله‌ی $\bar{\bar{Y}}_{\text{strat}}^* = \sum_{i=۱}^{m} \bar{Y}_{\text{strat},i}^* / m$ مشخص می‌شود.

۱-۱-۵- ترکیبات طبقات مجزا

از بخش ۱-۳ نتیجه شد که $k_h = ۱/(۱ - f_h)$، که در آن $f_h = (n_h - n_{hr}) / n_h$ نرخ بی‌پاسخی در طبقه‌ی hام است. با استفاده از رابطه‌ی (۸)، فرمول ترکیبی براورد واریانس $\bar{\bar{Y}}_{\text{strat}}^*$ به‌صورت

$$W_{\text{sep}} = \bar{\mathrm{V}}^{*} + \sum_{h=1}^{H}\left(\frac{1}{1-f_h}+\frac{1}{m}\right)B_h^{*}$$

است. در این‌جا، $\bar{\mathrm{V}}^{*} = \sum_{h=1}^{H}\bar{\mathrm{V}}_h^{*}$ و $\bar{\mathrm{V}}_h^{*}$ میانگین m مقدار از براورد واریانس مبتنی بر جانهی $\hat{V}_h^{*} = v_h^{2}\hat{\sigma}_{h*}^{2}(\frac{1}{n_h}-\frac{1}{N_h})$ است که در آن

$$\hat{\sigma}_{h*}^{2} = \frac{1}{n_h-1}\left(\sum_{s_{hr}}(y_i-\bar{y}_h^{*})^{2}+\sum_{s_h-s_{hr}}(y_i^{*}-\bar{y}_h^{*})^{2}\right)$$

۲-۱-۵- فرمول ترکیبی کلی. تعیین k در رابطه‌ی (۱)

از رابطه‌ی (۱۲) ، لازم است که $\mathrm{var}(v_h\bar{\mathrm{Y}}_h^{*}|y_{\text{obs}})$ و $\mathrm{var}(\bar{\mathrm{Y}}_{\text{strat}}^{*}|y_{\text{obs}}) = \sum_{h=1}^{H}\mathrm{var}(v_h\bar{\mathrm{Y}}_h^{*}|y_{\text{obs}})$ تعیین شوند. پس از آن

$$k = \sum_{h=1}^{H}\frac{1}{1-f_h}\cdot\frac{\mathrm{var}(v_h\bar{Y}_h^{*}|y_{\text{obs}})}{\mathrm{var}(\bar{Y}_{\text{strat}}^{*}|y_{\text{obs}})}$$

اکنون،

$$E(\bar{\mathrm{Y}}_h^{*}|y_{h,\text{obs}}) = \bar{y}_{hr}$$

و

$$\mathrm{var}(\bar{\mathrm{Y}}_h^{*}|y_{h,\text{obs}}) = \left\{\frac{n_h-n_{hr}}{n_h^{2}}\right\}\cdot\left\{\frac{(n_{hr}-1)}{n_{hr}}\right\}\hat{\sigma}_{hr}^{2} \approx \frac{f_h\hat{\sigma}_{hr}^{2}}{n_h}.$$

بنابراین می‌توانیم مقدار k را به‌صورت زیر تعیین کنیم.

$$k = \sum_{h=1}^{H}\frac{1}{1-f_h}\cdot\frac{f_h v_h^{2}\hat{\sigma}_{hr}^{2}/n_h}{\sum_{k=1}^{H}f_k v_k^{2}\hat{\sigma}_{kr}^{2}/n_h}$$

اگر اندازه‌های طبقه‌ی N_h بزرگ باشد، $\hat{\mathrm{V}}(v_h\overline{\mathrm{Y}}_h)=v_h^{۲}\hat{\sigma}_{hr}^{۲}/n_h$ درنظر گرفته می‌شود. همچنین اگر فرض کنیم $b_h=f_h\hat{\mathrm{V}}(v_h\overline{\mathrm{Y}}_h)/\sum_{k=۱}^{H}f_k\hat{\mathrm{V}}(v_k\overline{\mathrm{Y}}_k)$ باشد آن‌گاه

$$k=\frac{\sum_{h=۱}^{H}\hat{\mathrm{V}}(v_h\overline{\mathrm{Y}}_h)f_h\frac{۱}{۱-f_h}}{\sum_{h=۱}^{H}\hat{\mathrm{V}}(v_h\overline{\mathrm{Y}}_h)f_h}=\sum_{h=۱}^{H}b_h\cdot\frac{۱}{۱-f_h} \qquad (۱۳)$$

چون $\sum_{h=۱}^{H}b_h=۱$ است، نتیجه می‌شود که k میانگین موزون عکس نرخ پاسخ است. اگر همه‌ی $f_h=f$، نرخ بی‌پاسخی کلی باشد، آن‌گاه مقدار k مانند نمونه‌ی تصادفی ساده به‌صورت $k=۱/(۱-f)$ درنظر گرفته می‌شود. درغیر این‌صورت، اگر یا نرخ بی‌پاسخی زیاد باشد و یا واریانس براوردشده‌ی $v_h\overline{\mathrm{Y}}_h$ زیاد باشد، نرخ پاسخ طبقه $۱-f_h$، وزن زیادی دارد.

۳-۱-۵- یک عبارت دیگر برای k در رابطه‌ی (۱)

با استفاده از رابطه‌ی (۵) به‌طور مستقیم، می‌توان فرمول دیگری برای k به‌دست آورد. به شرط y_{obs}، میانگین‌های نمونه‌ی جانهی‌شده‌ی $\overline{\mathrm{Y}}_h^*$ مستقل هستند، که این مطلب دلالت بر این دارد که $E(\overline{\mathrm{Y}}_{\mathrm{strat}}^*|y_{\mathrm{obs}})=\sum_{h=۱}^{H}N_h\bar{y}_{hr}/N=\bar{y}_{\mathrm{strat},r}$ و $\mathrm{var}(\overline{\mathrm{Y}}_{\mathrm{strat}}^*|y_{\mathrm{obs}})\approx\sum_{h=۱}^{H}v_h^{۲}f_h\hat{\sigma}_{hr}^{۲}/n_h$. درست مانند بخش ۱-۳ ، رابطه‌ی (۳) برقرار است و با استفاده از رابطه‌ی (۵)، رابطه‌ی زیر به‌دست می‌آید.

$$E(k)\approx\frac{\mathrm{var}(\overline{\mathrm{Y}}_{\mathrm{strat},r})-\mathrm{var}(\overline{\mathrm{Y}}_{\mathrm{strat}})}{E(\sum_h v_h^{۲}f_h\hat{\sigma}_{hr}^{۲}/n_h)}$$
$$=\frac{\sum_{h=۱}^{H}v_h^{۲}\sigma_h^{۲}\left(E\left(\frac{۱}{n_{hr}}\right)-\frac{۱}{N_h}\right)-\sum_{h=۱}^{H}v_h^{۲}\sigma_h^{۲}\left(\frac{۱}{n_h}-\frac{۱}{N_h}\right)}{\sum_{h=۱}^{H}v_h^{۲}.E\left\{\frac{f_h}{n_h}E\left(\hat{\sigma}_{hr}^{۲}|n_{hr}\right)\right\}}$$

$$\approx \frac{\sum_{h=1}^{H} v_h^2 \sigma_h^2 \frac{1-p_{hr}}{n_h} \cdot \frac{1}{p_{hr}}}{\sum_{h=1}^{H} v_h^2 \sigma_h^2 \frac{1-p_{hr}}{n_h}} = \frac{\sum_{h=1}^{H} v_h^2 \frac{\sigma_h^2}{n_{hr}} E(f_h) \frac{1-f_h}{E(1-f_h)}}{\sum_{h=1}^{H} v_h^2 \frac{\sigma_h^2}{n_{hr}} E(f_h)(1-f_h)} \qquad (۱۴)$$

اکنون، $\text{var}(\bar{Y}_{hr}) = E\ \text{var}(\bar{Y}_{hr} | n_{hr}) = \sigma_h^2 E(1/n_{hr})$. با فرض $\hat{V}(v_h \bar{Y}_{hr}) = v_h^2 \hat{\sigma}_{hr}^2 / n_{hr}$ می‌توان نتیجه گرفت که اگر اندازه‌های طبقه‌ی N_h بزرگ باشد، با درنظر گرفتن

$$\frac{1}{k} = \frac{\sum_{h=1}^{H} (1-f_h) f_h \hat{V}(v_h \bar{Y}_{hr})}{\sum_{h=1}^{H} f_h \hat{V}(v_h \bar{Y}_{hr})} = \sum_{h=1}^{H} a_h (1-f_h) \qquad (۱۵)$$

که در آن وزن‌ها $a_h = f_h \hat{V}(v_h \bar{Y}_{hr}) / \sum_{k=1}^{H} f_k \hat{V}(v_k \bar{Y}_{kr})$ است، عبارت مربوط به $E(k)$ به‌طور تقریبی صدق می‌کند. چون $\sum_{h=1}^{H} a_h = 1$ است، در می‌یابیم که $1/k$ میانگین موزون نرخ‌های پاسخ است. اگر همه‌ی $f_h = f$، نرخ بی‌پاسخی کلی باشد، همان‌طور که در بخش۲-۱-۵ نشان داده شد، $k = 1/(1-f)$ خواهد بود. همان‌گونه که در بخش ۲-۱-۵ دیدیم، به این نکته نیز در رابطه‌ی (۱۵) اشاره کردیم که نرخ پاسخ طبقه $1-f_h$، حتی اگر نرخ بی‌پاسخی زیاد باشد و یا واریانس براورد شده‌ی $v_h \bar{Y}_{hr}$ زیاد باشد، وزن زیادی اخذ می‌کند. براورد مجموع براساس نمونه‌ی پاسخ به‌وسیله‌ی $\bar{Y}_{\text{strat},r} = \sum_h \nu_h \bar{Y}_{hr}$ مشخص می‌شود. با درنظر گرفتن این حقیقت از رابطه‌ی (۱۴) که $E(k) \approx \sum_{h=1}^{H} \text{var}(v_h \bar{Y}_h) E(f_h) \frac{1}{E(1-f_h)} / \sum_{h=1}^{H} \text{var}(v_h \bar{Y}_h) E(f_h)$، فرمول (۱۳) برای k به دست می‌آید. بنا بر این می‌توان نتیجه گرفت که اگر اندازه‌های طبقه‌ی N_h بزرگ باشد، با قرار دادن مقدار kای که در رابطه‌ی (۱۳) مشخص شده است، فرمول $E(k)$ به‌طور تقریبی برقرار است.

۲-۵-رگرسیون لوژستیک با متغیرهای تبیینی دودویی. براورد لگاریتم (نسبت بخت‌ها)

متغیرهای Y_1، ...، Y_n متغیرهای مستقلی هستند که مقادیر صفر و ۱ را اخذ می‌کنند و متغیر تبیینی x با مقادیر ثابت معلوم x_1، ...، x_n که مقادیر صفر و ۱ را اخذ می‌کند، درنظر گرفته شده است. احتمال رده‌ها با $x_i = 1$ $\pi_1 = P(Y_i = 1|x_i = 1)$ و $\pi_0 = P(Y_i = 1|x_i = 0)$ مشخص شده است. یک مدل گم‌شدن تصادفی برای متغیرهای پاسخ R_1، ...، R_n با احتمال‌های $P(R_i = 1|x_i = 1) = p_{1r}$ و $P(R_i = 1|x_i = 0) = p_{0r}$ فرض می‌شود. می‌توان مدل را در قالب مدل لوجیت به‌صورت $\log\{P(Y = 1|x)/P(Y = 0|x)\} = \alpha + \beta x$ بازپارامتری کرد که در آن $\alpha = \log\{\pi_0/(1-\pi_0)\}$ و $\beta = \log\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$ لگاریتم نسبت بخت‌ها است. هدف براورد β است. فرض کنید $s = (1,...,n)$ معرف کل نمونه‌ی با طبقات $s_1 = \{i \in s : x_i = 1\}$ و $s_0 = \{i \in s : x_i = 0\}$ باشد. اندازه‌های s_0 و s_1 با n_0 و n_1 مشخص شده‌اند که $n_1 = \sum_{i=1}^{n} x_i = X$ و $n_0 = n - X$ است. نمونه‌های پاسخ در طبقات، $s_{1r} = \{i \in s_1 : R_i = 1\}$ و $s_{0r} = \{i \in s_0 : R_i = 1\}$ با مجموع نمونه‌ی پاسخ s_r به اندازه‌ی n_r هستند. همچنین فرض کنید $n_{1r} = |s_{1r}|$ و $n_{0r} = |s_{0r}|$ باشد. می‌بینیم که $n_{1r} = \sum_{s_r} x_i = X_r$ و $n_{0r} = n_r - X_r$. می‌توان داده‌های s_r را که در آن n_{ijr} معرف تعداد مشاهدات با $x = i$ و $y = j$ است، به‌صورت زیر نمایش داد (جدول ۱ را ببینید).

پس براوردهای ماکسیمم درست‌نمایی برای π_{1r} و $\pi_{.r}$ به‌ترتیب برابر $\hat{\pi}_{1r} = n_{11r}/n_{1r}$ و $\hat{\pi}_{0r} = n_{01r}/n_{0r}$ است و در نتیجه براورد ماکسیمم درستنمایی β به‌صورت $\hat{\beta}_r = \log\frac{\hat{\pi}_{1r}/(1-\hat{\pi}_{1r})}{\hat{\pi}_{0r}/(1-\hat{\pi}_{0r})} = \log(n_{11r}n_{00r}/n_{10r}n_{01r})$ است. به همین ترتیب، براوردگر مبتنی بر کل نمونه به‌وسیله‌ی $\hat{\beta} = \log\frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_0/(1-\hat{\pi}_0)} = \log(n_{11}n_{00}/n_{10}n_{01})$ با نمادهایی آشکارا متناظر نمایش داده می‌شود. می‌توان این براورد را به‌صورت

جدول ۱– داده‌های مشاهده‌شده و مجموع بی‌پاسخی برای دو رده

y/x	$y=۰$	$y=۱$	مجموع	بی‌پاسخی
$x=۰$	$n_{۰۰r}$	$n_{۰۱r}$	$n_{۰r}$	$n_{۰}-n_{۰r}$
$x=۱$	$n_{۱۰r}$	$n_{۱۱r}$	$n_{۱r}$	$n_{۱}-n_{۱r}$

$\hat{\beta}=\log\{\hat{\pi}_{۱}/(۱-\hat{\pi}_{۱})\}-\log\{\hat{\pi}_{۰}/(۱-\hat{\pi}_{۰})\}=\hat{\beta}_{۱}-\hat{\beta}_{۰}$ به همان شکل بخش ۱-۴، بیان کرد. همچنین $\hat{\beta}_{۱}$ و $\hat{\beta}_{۰}$ که از طبقاتی با نمونه‌ی مجزای $s_{۱}$ و $s_{۰}$ به‌دست آمده‌اند، مستقل هستند. برای اندازه‌های نمونه‌ی $n_{۰}$ و $n_{۱}$ بزرگ، $\hat{\beta}$ به‌طور تقریبی دارای توزیع $N\left(\beta,\sigma^{۲}_{\hat{\beta}}\right)$ است که در آن $\sigma^{۲}_{\hat{\beta}}=\{n_{۱}\pi_{۱}(۱-\pi_{۱})\}^{-۱}+\{n_{۰}\pi_{۰}(۱-\pi_{۰})\}^{-۱}$ است. همچنین به‌طور تقریبی $\mathrm{var}\left(\hat{\beta}_{۱}\right)=۱/\{n_{۱}\pi_{۱}(۱-\pi_{۱})\}$ و $\mathrm{var}\left(\hat{\beta}_{۰}\right)=۱/\{n_{۰}\pi_{۰}(۱-\pi_{۰})\}$ و براورد $\mathrm{var}\left(\hat{\beta}\right)$ از طریق رابطه‌ی زیر مشخص می‌شود:

$$\hat{\mathrm{V}}\left(\hat{\beta}\right)=\frac{۱}{n_{۱}\hat{\pi}_{۱}\left(۱-\hat{\pi}_{۱}\right)}+\frac{۱}{n_{۰}\hat{\pi}_{۰}\left(۱-\hat{\pi}_{۰}\right)}=\left(\frac{۱}{n_{۱۱}}+\frac{۱}{n_{۱۰}}\right)+\left(\frac{۱}{n_{۰۱}}+\frac{۱}{n_{۰۰}}\right)$$

به‌گونه‌ای که $\hat{\mathrm{V}}\left(\hat{\beta}\right)=\hat{\mathrm{V}}_{۱}+\hat{\mathrm{V}}_{۰}$، که در آن $\hat{\mathrm{V}}_{۱}=\left(\frac{۱}{n_{۱۱}}+\frac{۱}{n_{۱۰}}\right)$ و $\hat{\mathrm{V}}_{۰}=\left(\frac{۱}{n_{۰۱}}+\frac{۱}{n_{۰۰}}\right)$ به ترتیب براورد واریانس $\hat{\beta}_{۱}$ و $\hat{\beta}_{۰}$ هستند.

اکنون روش جانهی زیر د رنظر گرفته خواهد شد؛ برای هر مقدار گم‌شده در $s_{۱}-s_{۱r}$، مقدار جانهی‌شده‌ی y^{*} به‌تصادف از توزیع براوردشده‌ی Y به شرط $x=۱$ به‌صورت زیر استخراج شده است؛

با احتمال $\hat{\pi}_{۱r}=n_{۱۱r}/n_{۱r}$، $y^{*}=۱$ و با احتمال $۱-\hat{\pi}_{۱r}$، $y^{*}=۰$.

همین روش جانهی برای $s_{۰}-s_{۰r}$ استفاده می‌شود به‌گونه‌ای که y^{*} به‌تصادف از توزیع براوردشده‌ی Y به شرط $x=۰$ استخراج می‌شود. این روش همان روش جانهی بی‌درنگ طبقه‌بندی‌شده است به‌گونه‌ای که مقادیر جانهی‌شده به‌تصادف و با جایگذاری از $y_{۱,\mathrm{obs}}=(y_{i}:i\in s_{۱r})$ و $y_{۰,\mathrm{obs}}=(y_{i}:i\in s_{۰r})$ استخراج می‌شوند.

مقادیر براورد شده در $s-s_r$ را می‌توان به همان شکل داده‌های اصلی نشان داد که در آن n^*_{ij} معرف تعداد مقادیر جانهی‌شده با $x=i$ و $y=j$ است (جدول ۲ را ببینید).

براورد π_1 براساس جانهی به‌وسیله‌ی $\hat{\pi}^*_1=\left(n_{11r}+n^*_{11}\right)/n_1$ مشخص می‌شود به‌گونه‌ای که براورد مبتنی بر جانهی

$$\hat{\beta}^*_1=\log\left\{\hat{\pi}^*_1/\left(1-\hat{\pi}^*_1\right)\right\}=\log\left\{\left(n_{11r}+n^*_{11}\right)/\left(n_1-n_{11r}-n^*_{11}\right)\right\}$$

است. به همین ترتیب، براوردهای β_0 و β براساس جانهی به‌صورت $\hat{\beta}^*_0=\log\left\{\left(n_{01r}+n^*_{01}\right)/\left(n_0-n_{01r}-n^*_{01}\right)\right\}$ و $\hat{\beta}^*=\hat{\beta}^*_1-\hat{\beta}^*_0$ تعیین می‌شوند.

m جانهی تکرار شده به m براورد $\hat{\beta}^*_{1,i}$، $\hat{\beta}^*_{0,i}$، $\hat{\beta}^*_i$ برای $i=1,\ldots,m$ منجر می‌شود. براورد ترکیب‌شده به وسیله‌ی

$$\overline{\beta}^*=\sum_{i=1}^m\hat{\beta}^*_i/m=\sum_{i=1}^m\hat{\beta}^*_{1,i}/m-\sum_{i=1}^m\hat{\beta}^*_{0,i}/m=\overline{\beta}^*_1-\overline{\beta}^*_0$$

بیان می‌شود. براورد واریانس جانهی‌شده‌ی $\hat{\mathrm{V}}^*$ برای $\hat{\beta}$ به‌صورت زیر مشخص می‌شود؛

$$\hat{\mathrm{V}}^*=\frac{1}{n_{11r}+n^*_{11}}+\frac{1}{n_{10r}+n^*_{10}}+\frac{1}{n_{01r}+n^*_{01}}+\frac{1}{n_{00r}+n^*_{00}} \qquad (16)$$

می‌بینیم که $E\left(\hat{\mathrm{V}}^*\middle|y_{\mathrm{obs}}\right)\approx\frac{1}{n_1\hat{\pi}_{1r}\left(1-\hat{\pi}_{1r}\right)}+\frac{1}{n_0\hat{\pi}_{0r}\left(1-\hat{\pi}_{0r}\right)}$ و رابطه‌ی (۳) برقرار است. همچنین توجه می‌کنیم که رابطه‌ی (۹) نیز به طور مجزا برای هر رده برقرار است.

۱-۲-۵- ترکیب رده‌های مجزا

ابتدا از رهیافت بخش ۱-۴ استفاده می‌کنیم و مقادیر مجزای k_1 و k_0 را برای دو رده تعیین می‌کنیم. طبقه‌ی اول $s_1=\{i\in s: x_i=1\}$ را در نظر می‌گیریم. در پیوست ۳ نشان داده شده است که $E\left(\hat{\beta}^*_1\middle|y_{1,\mathrm{obs}}\right)\approx\hat{\beta}_{1r}$ و $\mathrm{var}\left(\hat{\beta}^*_1\middle|y_{1,\mathrm{obs}}\right)\approx f_1\left(1-f_1\right)\hat{\mathrm{V}}\left(\hat{\beta}_{1r}\right)$. با استفاده از رابطه‌ی (۱۰)، به طور تقریبی داریم

جدول ۲- مجموع‌های جانهی‌شده برای دو رده

مجموع	$y=۱$	$y=۰$	y/x
$n_{۰}-n_{۰r}$	$n^{*}_{۰۱}$	$n^{*}_{۰۰}$	$x=۰$
$n_{۱}-n_{۱r}$	$n^{*}_{۱۱}$	$n^{*}_{۱۰}$	$x=۱$

$$E\left(k_{۱}\right)=\frac{\operatorname{var}(\hat{\beta}_{۱r})-\operatorname{var}(\hat{\beta}_{۱})}{E\left\{f_{۱}\left(۱-f_{۱}\right)\hat{\mathrm{V}}(\hat{\beta}_{۱r})\right\}}=\frac{E\ \operatorname{var}\left(\hat{\beta}_{۱r}\left|n_{۱r}\right.\right)-\operatorname{var}\left(\hat{\beta}_{۱}\right)}{E\left\{f_{۱}\left(۱-f_{۱}\right)E[\hat{\mathrm{V}}(\hat{\beta}_{۱r})\left|n_{۱r}\right.]\right\}}$$

$$\approx\frac{\frac{۱}{\pi_{۱}\left(۱-\pi_{۱}\right)}\left(E\left(\frac{۱}{n_{۱r}}\right)-\frac{۱}{n_{۱}}\right)}{Ef_{۱}\left(۱-f_{۱}\right)\frac{۱}{n_{۱r}\pi_{۱}\left(۱-\pi_{۱}\right)}}\approx\frac{\left(۱-p_{۱r}\right)/p_{۱r}}{۱-p_{۱r}}=\frac{۱}{p_{۱r}}$$

که با قرار دادن $k_{۱}=۱/\left(۱-f_{۱}\right)$ به طور تقریبی در رابطه صدق می‌کند. درست به همین ترتیب پی می‌بریم که $k_{۰}=۱/\left(۱-f_{۰}\right)$ است که در آن $f_{۰}=\left(n_{۰}-n_{۰r}\right)/n_{۰}$ نسبت بی‌پاسخی در طبقه‌ی $s_{۰}$ است. مؤلفه‌ی بین جانهی برای $\hat{\beta}^{*}_{۱}$ به‌وسیله‌ی $B^{*}_{۱}=\frac{۱}{m-۱}\sum_{i=۱}^{m}\left(\hat{\beta}^{*}_{۱,i}-\bar{\beta}^{*}_{۱}\right)^{۲}$ مشخص می‌شود و همچنان $B^{*}_{۰}$، مؤلفه‌ی بین جانهی برای $\hat{\beta}^{*}_{۰}$ است. بنا بر این یک واریانس براوردشده‌ی براورد ترکیبی $\bar{\beta}^{*}$ برای β بر اساس جانهی، با استفاده از رابطه‌ی (۸) به‌صورت زیر مشخص می‌شود

$$W_{\text{sep}}=\bar{\mathrm{V}}^{*}+\sum_{x=۰}^{۱}\left(\frac{۱}{۱-f_{x}}+\frac{۱}{m}\right)B^{*}_{x}$$

که در آن $\bar{\mathrm{V}}^{*}$ میانگین m تکرار براورد واریانس جانهی‌شده‌ی $\hat{\mathrm{V}}^{*}$ است که از طریق رابطه‌ی (۱۶) مشخص شده است.

۲-۲-۵- فرمول ترکیبی کلی. تعیین k در رابطه‌ی (۱)

چون $\operatorname{var}(\hat{\beta}_{۱}^{*}\,|\,y_{۱,\mathrm{obs}})=f_{۱}(۱-f_{۱})\hat{\mathrm{V}}(\hat{\beta}_{۱r})$ و $\operatorname{var}(\hat{\beta}_{۰}^{*}\,|\,y_{۰,\mathrm{obs}})=f_{۰}(۱-f_{۰})\hat{\mathrm{V}}(\hat{\beta}_{۰r})$ است، از رابطه‌ی (۱۲) داریم؛

$$k=\frac{۱}{۱-f_{۱}}\cdot\frac{f_{۱}(۱-f_{۱})\hat{\mathrm{V}}(\hat{\beta}_{۱r})}{\sum_{x=۰}^{۱}f_{x}(۱-f_{x})\hat{\mathrm{V}}\left(\hat{\beta}_{xr}\right)}+\frac{۱}{۱-f_{۰}}\cdot\frac{f_{۰}(۱-f_{۰})\hat{\mathrm{V}}(\hat{\beta}_{۰r})}{\sum_{x=۰}^{۱}f_{x}(۱-f_{x})\hat{\mathrm{V}}\left(\hat{\beta}_{xr}\right)} \qquad (۱۷)$$

$\operatorname{var}\left(\hat{\beta}_{۱}\right)\approx(n_{۱r}/n_{۱})\operatorname{var}\left(\hat{\beta}_{۱r}\,|\,n_{۱r}\right)=(۱-f_{۱})\operatorname{var}\left(\hat{\beta}_{۱r}\,|\,n_{۱r}\right)$. به همین ترتیب، $\operatorname{var}\left(\hat{\beta}_{۰}\right)\approx(۱-f_{۰})\operatorname{var}\left(\hat{\beta}_{۰r}\,|\,n_{۰r}\right)$ است. بنا بر این می‌توان واریانس براوردهای کل نمونه‌ی $\hat{\beta}_{۱}$ و $\hat{\beta}_{۰}$ را به ترتیب از طریق $\hat{\mathrm{V}}\left(\hat{\beta}_{۱}\right)=(۱-f_{۱})\hat{\mathrm{V}}\left(\hat{\beta}_{۱r}\right)$ و $\hat{\mathrm{V}}\left(\hat{\beta}_{۰}\right)=(۱-f_{۰})\hat{\mathrm{V}}\left(\hat{\beta}_{۰r}\right)$ براورد کرد. بنا بر این

$$k=\frac{۱}{۱-f_{۱}}\cdot\frac{f_{۱}\hat{\mathrm{V}}(\hat{\beta}_{۱})}{\sum_{x=۰}^{۱}f_{x}\hat{\mathrm{V}}\left(\hat{\beta}_{x}\right)}+\frac{۱}{۱-f_{۰}}\cdot\frac{f_{۰}\hat{\mathrm{V}}(\hat{\beta}_{۰})}{\sum_{x=۰}^{۱}f_{x}\hat{\mathrm{V}}\left(\hat{\beta}_{xr}\right)}=\frac{۱}{۱-f_{۱}}b_{۱}+\frac{۱}{۱-f_{۰}}\cdot(۱-b_{۱})$$

درست مانند بخش ۲-۱-۵ می‌بینیم که k میانگین موزون عکس نرخ‌های پاسخ است. اگر همه‌ی $f_{h}=f$ ، نرخ بی‌پاسخی کلی باشد، آن‌گاه $k=۱/(۱-f)$. در غیر این‌صورت، نرخ پاسخ طبقه‌ی $۱-f_{x}$ زمانی که یا نرخ بی‌پاسخی زیاد باشد و یا واریانس براورد شده‌ی $\hat{\beta}_{x}$ زیاد باشد، وزن زیادی خواهد داشت. یا این‌که از رابطه‌ی (۱۷)

$$\frac{۱}{k}=\frac{\sum_{x=۰}^{۱}(۱-f_{x})f_{x}\hat{\mathrm{V}}\left(\hat{\beta}_{xr}\right)}{\sum_{x=۰}^{۱}f_{x}\hat{\mathrm{V}}\left(\hat{\beta}_{xr}\right)}=\sum_{x=۰}^{۱}a_{x}(۱-f_{x})$$

که در آن وزن‌ها به‌صورت $a_{x}=f_{x}\hat{\mathrm{V}}(\hat{\beta}_{xr})/\{f_{۱}\hat{\mathrm{V}}(\hat{\beta}_{۱r})+f_{۰}\hat{\mathrm{V}}(\hat{\beta}_{۰r})\}$ می‌باشد. بنا بر این می‌توان $۱/k$ را به‌عنوان میانگین موزون از نرخ‌های پاسخ بیان کرد.

اگر هدف براورد π_1 و π_0 باشد، مقدار $k=1/(1-f_1)$ و $k=1/(1-f_0)$ به‌ترتیب برای π_1 و π_0 به‌دست می‌آید.

۳-۵- رگرسیون لوژستیک با متغیرهای تبیینی رسته‌ای گسسته. براورد لگاریتم (نسبت بخت‌ها)

اگر متغیر تبیینی x به‌صورت رسته‌ای است که مثلاً H رده را تعریف می‌کند، می‌توان نتایج را به‌صورت زیر تعمیم داد.
فرض کنید برای $h=0,\ldots,H-1$ ، $\pi_h=P(Y=1|x=h)$ است. رگرسیون لوژستیکی که برای این داده‌های رده‌ای تعریف می‌شود، از طریق معرفی $H-1$ متغیر تبیینی دودویی x_1، ...، x_{H-1} صورت می‌گیرد که در آن اگر مشاهدات به رده‌ی h تعلق داشته باشد، برای $h=0,\ldots,H-1$ ، $x_h=1$ و در غیر این صورت صفر خواهد بود. پس اگر $x_1=x_2=\cdots=x_{H-1}=0$ باشد، مشاهده به رده‌ی صفر تعلق دارد. مدل لوجیت با $x=(x_1,x_2,\ldots,x_{H-1})$ به‌صورت

$$\log\{P(Y=1|x)/P(Y=0|x)\}=\alpha+\beta_1x_1+\beta_2x_2+\ldots+x_{H-1}\beta_{H-1}$$

نمایش داده می‌شود و می‌توان نتیجه گرفت که برای رده‌ی h در برابر رده‌ی صفر $\alpha=\log\frac{\pi_0}{1-\pi_0}$ و $\beta_h=\log\frac{\pi_h/(1-\pi_h)}{\pi_0/(1-\pi_0)}$ لگاریتم نسبت بخت‌ها است. براورد β_h از طریق جانهی چندگانه دقیقاً مشابه متغیر دودویی x با جایگزینی رده‌ی h با رده‌ی ۱ انجام می‌شود.

۴-۵- رگرسیون لوژستیک با مقادیر گم‌شده در یک متغیر تبیینی دودویی

این وضعیت همانند بخش ۲-۵ است با این تفاوت که y به طور کامل در s مشاهده شده است، $y=(y_1,\ldots,y_n)$، و مقادیر گم‌شده برای متغیر تبیینی x وجود دارد. Y_1، ...، Y_n متغیرهای مستقل با مقادیر صفر و ۱ هستند و متغیرهای تبیینی صفر و ۱ با

جدول ۳- داده‌های مشاهده شده و مجموع‌های بی‌پاسخی برای طبقه‌ی y

y/x	$y=۰$	$y=۱$
$x=۰$	$n_{۰۰r}$	$n_{۰۱r}$
$x=۱$	$n_{۱۰r}$	$n_{۱۱r}$
مجموع	$n^{۰}_{۰r}$	$n^{۰}_{۱r}$
بی‌پاسخی	$n^{۰}_{۰}-n^{۰}_{۰r}$	$n^{۰}_{۱}-n^{۰}_{۱r}$

مقادیر ثابت $x_۱$، $x_۲$، ...، x_n که برخی از آن‌ها گم‌شده است نیز وجود دارد. متغیرهای پاسخ نشان‌دهنده‌ی گم‌شدگی x_iها است اما اکنون با مدل گم‌شدن تصادفی، $P\left(R_i=۱\middle|y_i=۱\right)=q_{۱r}$ و $P\left(R_i=۱\middle|y_i=۰\right)=q_{۰r}$ روبه‌رو هستیم.

در غیر این‌صورت، مدل مشابه بخش ۲-۵ با احتمال‌های رده‌ای $\pi_۱=P\left(Y_i=۱\middle|x_i=۱\right)$ و $\pi_۰=P\left(Y_i=۱\middle|x_i=۰\right)$ و مدل لوجیت $\log\left\{P\left(Y=۱\middle|x\right)/P\left(Y=۰\middle|x\right)\right\}=\alpha+\beta x$ با $\beta=\log\frac{\pi_۱/\left(۱-\pi_۱\right)}{\pi_۰/\left(۱-\pi_۰\right)}$ است. هنوز هدف براورد β است.

اکنون فرض کنید $s^{۱}=\{i\in s: y_i=۱\}$ و $s^{۰}=\{i\in s: y_i=۰\}$ با اندازه‌های $n^{۱}_{۱}$ و $n^{۰}_{۰}$ باشد. نمونه‌های پاسخ در طبقات، $s^{۱}_{r}=\{i\in s^{۱}: R_i=۱\}$ و $s^{۰}_{r}=\{i\in s^{۰}: R_i=۱\}$ با نمونه‌ی پاسخ کل $s_r=\{i\in s: R_i=۱\}=s^{۱}_{r}\cup s^{۰}_{r}$ است. حال، داده‌ها می‌توانند همانند قبل معرفی شوند با این تفاوت که مجموع‌های بی‌پاسخی به هر طبقه‌ی y تعلق دارد (جدول ۳ را ببینید).

براورد ماکسیمم درستنمایی $\hat{\pi}_{۱r}$، $\hat{\pi}_{۰r}$، $\hat{\beta}_r$، براساس s_r مشابه قبل است که برای براورد کل نمونه‌ی $\hat{\beta}$ بوده است. روش جانهی، روش بی‌درنگ طبقه‌بندی‌شده برای طبقه‌های y است. برای هر مقدار گم‌شده‌ی x در $s^{۱}-s^{۱}_{r}$، مقدار جانهی‌شده‌ی x^* به‌تصادف از $x_{۱,\mathrm{obs}}=\left(x_i: i\in s^{۱}_{r}\right)$ قرعه‌کشی می‌شود. به همین ترتیب، مقادیر جانهی شده در $s^{۰}-s^{۰}_{r}$ نیز به تصادف از $x_{۰,\mathrm{obs}}=\left(x_i: i\in s^{۰}_{r}\right)$ قرعه‌کشی می‌شود. مقادیر

جدول ۴- مجموع‌های جانهی‌شده برای طبقه‌های y

y/x	$y=0$	$y=1$
$x=0$	n^*_{00}	n^*_{01}
$x=1$	n^*_{10}	n^*_{11}
مجموع	$n^{(0)}_0 - n^{(0)}_{0r}$	$n^{(0)}_1 - n^{(0)}_{1r}$

جانهی شده در $s - s_r$ می‌تواند به همان صورت داده‌های اصلی بیان شود به‌طوری که اکنون n^*_{ij} معرف تعداد مقادیر جانهی‌شده با $x=i$ و $y=j$ است (جدول ۴ را ببینید).

لازم است یک عبارت تقریبی برای امید ریاضی و واریانس $\hat{\beta}^*$، که اکنون با $\hat{\beta}_*$ نمایش داده می‌شود، به شرط داده‌های مشاهده شده، پیدا کنیم. این مطلب در پیوست ۴ نشان داده شده است که

$$\mathrm{var}\left(\hat{\beta}_* | y, x_{\mathrm{obs}}\right) \approx f^{(1)}\left(1-f^{(1)}\right)\left(\frac{1}{n_{11r}}+\frac{1}{n_{01r}}\right)+f^{(0)}\left(1-f^{(0)}\right)\left(\frac{1}{n_{10r}}+\frac{1}{n_{00r}}\right)$$

و

$$E\left(\hat{\beta}_* | y, x_{\mathrm{obs}}\right) \approx \hat{\beta}_r .$$

در این حالت $f^{(1)} = \left(n^{(0)}_1 - n^{(0)}_{1r}\right)/n^{(0)}_1$ نرخ بی‌پاسخی در طبقه‌ی $s^{(1)}$ و $f^{(0)} = \left(n^{(0)}_0 - n^{(0)}_{0r}\right)/n^{(0)}_0$ نرخ بی‌پاسخی در طبقه‌ی $s^{(0)}$ است. لازم به ذکر است که $\hat{q}_{1r} = n^{(0)}_{1r}/n^{(0)}_1 = 1-f^{(1)}$ و $\hat{q}_{0r} = n^{(0)}_{0r}/n^{(0)}_0$ است. بنابراین مخرج رابطه‌ی (۵) به‌صورت زیر در می‌آید.

$$(18) \qquad E\left\{f^{(1)}\left(1-f^{(1)}\right)\left(\frac{1}{n_{11r}}+\frac{1}{n_{01r}}\right)+f^{(0)}\left(1-f^{(0)}\right)\left(\frac{1}{n_{10r}}+\frac{1}{n_{00r}}\right)\right\}$$

صورت کسر رابطه‌ی (۵) نیز همانند قبل معادل، $\mathrm{var}\left(\hat{\beta}_r\right) - \mathrm{var}\left(\hat{\beta}\right)$، است و می‌توان به طور تقریبی به‌صورت زیر بیان کرد

$$\text{var}\left(\hat{\beta}_r\right)-\text{var}\left(\hat{\beta}\right)=\frac{1}{n_1\pi_1\left(1-\pi_1\right)}\cdot\frac{1-p_{1r}}{p_{1r}}+\frac{1}{n_0\pi_0\left(1-\pi_0\right)}\cdot\frac{1-p_{0r}}{p_{0r}} \quad (19)$$

که در آن همانند قبل، $p_{1r}=P\left(R_i=1\middle|x_i=1\right)$ و $p_{0r}=P\left(R_i=1\middle|x_i=0\right)$ می‌باشد. به براوردهای دیگری از p_{1r} و p_{0r} نیز نیاز داریم. چون $p_{1r}=\pi_1q_{1r}+\left(1-\pi_1\right)q_{0r}$، آن‌گاه داریم $\hat{p}_{1r}=\hat{\pi}_1\left(1-f^1\right)+\left(1-\hat{\pi}_1\right)\left(1-f^0\right)$. به همین ترتیب، $\hat{p}_{0r}=\hat{\pi}_0\left(1-f^1\right)+\left(1-\hat{\pi}_0\right)\left(1-f^0\right)$.

همچنین می‌توانیم از براوردگرهای دیگری چون $n_1\hat{p}_{1r}\approx n_{1r}$ و $n_0\hat{p}_{0r}\approx n_{0r}$ نیز استفاده کنیم. با استفاده از روابط (۱۸) و (۱۹) می‌توان نتیجه گرفت که رابطه‌ی زیر برقرار است.

$$k=\frac{\left(\frac{1}{n_{11r}}+\frac{1}{n_{10r}}\right)\left(\hat{\pi}_{1r}f^1+\left(1-\hat{\pi}_{1r}\right)f^0\right)+\left(\frac{1}{n_{01r}}+\frac{1}{n_{00r}}\right)\left(\hat{\pi}_{0r}f^1+\left(1-\hat{\pi}_{0r}\right)f^0\right)}{f^1\left(1-f^1\right)\left(\frac{1}{n_{11r}}+\frac{1}{n_{01r}}\right)+f^0\left(1-f^0\right)\left(\frac{1}{n_{01r}}+\frac{1}{n_{00r}}\right)}$$

$$=\frac{f^1\left(\frac{1}{n_{10r}}+\frac{1}{n_{00r}}\right)+f^0\left(\frac{1}{n_{11r}}+\frac{1}{n_{01r}}\right)}{f^1\left(1-f^1\right)\left(\frac{1}{n_{11r}}+\frac{1}{n_{01r}}\right)+f^0\left(1-f^0\right)\left(\frac{1}{n_{01r}}+\frac{1}{n_{00r}}\right)}$$

لازم به ذکر است که اگر $f^1=f^0=f$ باشد، آن‌گاه $k=1/\left(1-f\right)$ است. در غیر این‌صورت، می‌توان $1/k$ را به‌عنوان یک ترکیب خطی از نرخ‌های پاسخ $\left(1-f^1,1-f^0\right)$ بیان کرد. فرض کنید $w_1=\frac{1}{n_{11r}}+\frac{1}{n_{01r}}$ و $w_0=\frac{1}{n_{10r}}+\frac{1}{n_{00r}}$ باشد. بنا بر این خواهیم داشت

$$\frac{1}{k}=a_1\left(1-f^1\right)+a_0\left(1-f^0\right)$$

که در آن $a_1 = f^{-1} w_1 / \left(f^{-1} w_0 + f^{-1} w_1\right)$ و $a_0 = f^{-1} w_0 / \left(f^{-1} w_0 + f^{-1} w_1\right)$ است. لازم به ذکر است که در حالت کلی $a_1 + a_0 \neq 1$.

۶- سؤال: آیا می‌توان از فرمول ترکیبی یکسان برای یک وضعیت مشخص و یک روش جانهی برای همه‌ی براوردهای علمی استفاده کرد؟

در این بخش سعی می‌شود یک رهیافت کلی برای این مسئله ارائه شود. فرض کنید s معرف کل‌نمونه و y داده‌های کل‌نمونه باشد. سه حالت ممکن است وجود داشته باشد:

۱- s یک نمونه از یک جامعه‌ی متناهی و $y = \left(y_i : i \in s\right)$ با مدل طرح باشد. بنا بر این متغیرهای تصادفی مشاهده‌شده، (s, s_r) هستند و y_{obs} معادل (s, s_r) است.

۲- وضعیت مشابهی همانند حالت ۱، ولی با یک مدل جامعه به جای مدل طرح وجود داشته باشد. در این حالت، متغیرهای تصادفی مشاهده‌شده $y_{\text{obs}} = \left\{\left(y_i : i \in s_r\right), s_r, s\right\}$ هستند.

۳- یک مطالعه‌ی مشاهده‌ای که در آن $s = (1, \ldots, n)$ باشد و $y = \left(y_1, \ldots, y_n\right)$ مدل‌بندی شده باشد. در این حالت متغیرهای تصادفی مشاهده‌شده به‌صورت $y_{\text{obs}} = \left\{\left(y_i : i \in s_r\right), s_r\right\}$ هستند.

به‌عنوان یک مثال، حالتی را با وجود بی‌پاسخی با گم‌شدگی کاملاً تصادفی (که متغیرهای پاسخ R_i با احتمال $p_r = P\left(R_i = 1\right)$ مستقل هستند) و جانهی بی‌درنگ در نظر می‌گیریم. حالت معرفی‌شده در بخش ۱-۳ با نمونه‌ی تصادفی ساده، یک حالت خاص شماره‌ی ۱ است و دریافتیم که برای براورد میانگین جامعه با استفاده از میانگین نمونه،

$$k = \frac{1}{1-f} \qquad (۲۰)$$

که در آن $f = \left(n - n_r\right)/n = 1 - \hat{p}_r$ نرخ بی پاسخی است.

با محدود کردن توجه به براوردهای خطی که در آن براوردگر جانهی‌شده‌ی $\hat{\theta}^*$، پارامتر مشابهی مانند $\hat{\theta}$ را براورد می‌کند، نشان خواهیم داد که رابطه‌ی (۲۰) در حالت کلی برای

هر سه حالت معرفی‌شده‌ی بالا، زمانی که مکانیسم بی‌پاسخی، گمشدگی کاملاً تصادفی است و از روش جانهی بی‌درنگ استفاده شده باشد، برقرار است. اولین سئوالی که بررسی می‌کنیم این است که آیا روش جانهی بی‌درنگ به براوردگرهای معتبر جانهی- مبنا منجر می‌شود به‌گونه‌ای که این مقدار k بتواند استفاده شود. جواب در حالت کلی خیر است. نیازی آشکار برای یک روش جانهی، حد اقل به طور تقریبی، آن است که

$$E\left(\hat{\theta}^{*}\,|y,s\right)=\hat{\theta} \qquad (۲۱)$$

یعنی براوردگر جانهی‌شده باید پارامتر مشابهی مانند $\hat{\theta}$ را براورد کند. می‌توان گفت که مقدار مورد انتظار براوردگر جانهی‌شده به شرط کل داده‌های نمونه‌گیری‌شده، باید معادل براورد کل نمونه باشد. در حالت ۱، زمانی که s مشخص باشد، y اضافی است و رابطه‌ی (۲۱) بیان می‌کند که $E\left(\hat{\theta}^{*}\,|s\right)=\hat{\theta}$ است. در حالت ۳، s تصادفی نیست و بنا بر این غیر ضروری است، در حالی که در حالت ۲ هر دوی s و y مورد نیاز است.

در این مقاله براوردهایی را درنظر گرفته‌ایم که خطی در $\left(y_i : i\in s\right)$ هستند. نتایج زیر که در پیوست ۵ اثبات شده‌اند، براوردهای خطی را مشخص می‌کنند که با استفاده از روش جانهی بی‌درنگ در رابطه‌ی (۲۱) صدق می‌کند و نشان می‌دهند که برای چنین براوردگرهایی، $k=۱/\left(۱-f\right)$ است.

لم. *فرض کنید* $\hat{\theta}=\sum_{i\in s}a_i\left(s\right)y_i$ *است. بنا بر این* $E\left(\hat{\theta}^{*}\,|y,s\right)=\hat{\theta}$ *اگر و تنها اگر برای همه‌ی* $i\in s$، $a_i\left(s\right)=a(s)$. *یعنی* $\hat{\theta}=a(s)\sum_{i\in s}y_i=n\,a(s)\,\bar{y}_s$.

تذکر. در حالت ۳، s هیچ اطلاعی نمی‌دهد و $a_i\left(s\right)=a_i$.

قضیه. فرض کنید $\hat{\theta}=\sum_{i\in s}a_i\left(s\right)y_i$ و $E\left(\hat{\theta}^{*}\,|\,y,s\right)=\hat{\theta}$ باشد. همچنین فرض کنید که رابطه‌ی (۳) برقرار است. بنا بر این $E\left(k\right)=\dfrac{E\left(۱/\hat{p}_r\right)-۱}{۱-p_r-\frac{۱}{n}\left[E\left(۱/\hat{p}_r\right)-۱\right]}\approx\dfrac{۱}{p_r}$ و رابطه‌ی $k=۱/\left(۱-f\right)$ می‌تواند مورد استفاده قرار گیرد.

اکنون به برخی از حالت‌های خاص می‌پردازیم؛

۱- با درنظر گرفتن $a(s)=۱/n$، مشابه بخش ۱-۳، می‌بینیم که رابطه‌ی (۲۱) برقرار است.

۲- ضریب رگرسیون برای رگرسیونی که از مبدأ می‌گذرد، به‌صورت $\hat{\beta}=\sum_{i=۱}^{n} y_i / \sum_{i=۱}^{n} x_i$ است. در این حالت، با در نظر گرفتن $a=۱/\sum_{i=۱}^{n} x_i$ رابطه‌ی (۲۱) برقرار است و بنا بر این $k=۱/(۱-f)$ است.

۳- براورد ضریب رگرسیون در رگرسیون خطی معمول، حالتی است که در رابطه‌ی (۲۱) صدق نمی‌کند و در آن $\hat{\beta}=\sum(x_i-\bar{x})y_i/\sum(x_i-\bar{x})^۲$ است. در این حالت، $a_i=(x_i-\bar{x})/\sum_{j=۱}^{n}(x_j-\bar{x})^۲$ است که مستقل از i نیست. می‌توان نشان داد که به‌طور تقریبی $E(\hat{\beta}^*|y)\approx p_r\hat{\beta}$ و به طور دقیق به‌صورت $\frac{np_r-۱}{n-۱}\hat{\beta}$ است. بنا بر این برای مسائل رگرسیون عادی، روش جانهی بی‌درنگ نمی‌تواند مؤثر واقع شود. لازم به ذکر است که با استفاده از مطالب بخش ۳-۳ می‌توان با استفاده از جانهی مانده‌ها از $k=۱/(۱-f)$ استفاده کرد.

بدیهی است که زمانی که y با مقدار معلوم x در یک گروه بی‌پاسخی همبسته باشد، باید صرف نظر از مشکلات براورد تحت بررسی، این همبستگی را در جانهی‌ها مورد استفاده قرار داد.

۷- بحث و نتیجه‌گیری

در این مقاله نشان داده شد که امکان بسط یک قضیه‌ی کلی برای جانهی چندگانه، بدون نیاز به قرعه‌کشی تصادفی جانهی‌ها از یک توزیع پسین بیزی وجود دارد. برای نمونه‌های طبقه‌بندی‌شده با براوردهای طبقه‌بندی‌شده، نیاز به مطالعات بیش‌تر برای این‌که چه

براورد واریانسی به‌کار گرفته شود، وجود دارد. یعنی آیا رابطه‌ی (۸) با استفاده از ترکیب طبقات مجزا که در رابطه‌ی (۱۰) مشخص شده است، استفاده شود و یا ترکیب کلی رابطه‌ی (۱) با مقدار k مشخص شده در رابطه‌ی (۱۲) به کار گرفته شود.

از مواردی که در این مقاله ارائه شده است می‌توان نتیجه گرفت که فرمول روش جانهی غیر بیزی نوعاً به معیاری از سهم اطلاع گم‌شده در نمونه‌ی پاسخ در مقایسه با کل نمونه بستگی دارد. در ساده‌ترین حالت در بخش ۱-۳، اطلاعات گم‌شده به‌وسیله‌ی $۱/(۱-f)$ که عکس نرخ پاسخ است، اندازه‌گیری شده است. هر چه این عامل بیش‌تر باشد، وزن مؤلفه‌ی بین جانهی نیز بیش‌تر است. در مدل نسبت بخش ۲-۳ با روش جانهی رگرسیونی بی‌درنگ مانده‌ها، میزان اطلاعات گم‌شده عکس نسبت $x-$ کل در نمونه‌ی پاسخ در مقایسه با کل نمونه است. نشان داده شده است که در رگرسیون خطی ساده با عبارت واریانسی که مستقل از متغیر تبیینی است، مجدداً اطلاع گم‌شده به‌وسیله‌ی $۱/(۱-f)$ اندازه‌گیری می‌شود. یک پیشنهاد برای مطالعات بعدی، بررسی امکان تعمیم این نتیجه برای تعیین k، از طریق تعریف معیارهای مرتبط با اطلاعات گم‌شده، با استفاده از فرمول تعریف پایه‌ای (۵) می‌باشد.

مطالعه‌ی عملکرد بازه‌های اطمینان مربوط نیز باقی می‌ماند. برخی مطالعات شبیه‌سازی مقدماتی که در این مقاله گنجانده نشده‌اند، نشان می‌دهند که برای رگرسیون خطی ساده با استفاده از روش جانهی مانده‌ها و $k = ۱/(۱-f)$، بازه‌ی اطمینان به‌صورت $\bar{\beta}^* \pm z_{\alpha/۲}\sqrt{W}$ (که در آن $z_{\alpha/۲}$، $\frac{\alpha}{۲}$ نقطه‌ی بالایی توزیع نرمال استاندارد می‌باشد) است که به طور تقریبی به سطح اسمی $(۱-\alpha)$ دست می‌یابد.

مرجع‌ها

[1] Little, R.J.A.; Rubin, D.B. (2002). *Statistical Analysis with Missing* Data. Wiley, New York.

[2] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

[3] Rubin, D.B. (1996). Multiple imputation after 18 + years (with discussion). Journal of *the American Statistical Association,* ***91***, 473–489.

[4] Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.

پیوست

۱- جانهی چندگانه برای مدل نسبت در بخش ۲-۳

ابتدا شرط (۳) که معادل $E\left(\hat{\sigma}_*^2\right) \approx \sigma^2$ می‌باشد را درنظر بگیرید. فرض کنید $\hat{\beta}_{nr} = \sum_{s-s_r} y_i^* / X_{nr}$ و $\hat{\sigma}_{nr}^2 = \sum_{s-s_r} \frac{1}{x_i}\left(y_i^* - \hat{\beta}_{nr} x_i\right)^2 / \left(n_{nr}-1\right)$. در این حالت، $n_{nr} = n - n_r$. سپس عبارت $\hat{\sigma}_*^2$ می‌تواند به‌صورت زیر بیان شود:

$$\hat{\sigma}_*^2 = \frac{1}{n-1}\left(\left(n_r-1\right)\hat{\sigma}_r^2 + \left(n_{nr}-1\right)\hat{\sigma}_{nr}^2 + \frac{X_r X_{nr}}{X}\left(\hat{\beta}_r - \hat{\beta}_{nr}\right)^2\right)$$

در این حالت، $E\left(Y_i^* \middle| y_{\text{obs}}\right) = \hat{\beta}_r x_i + \bar{e}\sqrt{x_i}$، که در آن $\bar{e} = \sum_{s_r} e_i / n_r$ و $\text{var}\left(Y_i^* \middle| y_{\text{obs}}\right) = x_i s_e^2$ است که در آن $s_e^2 = \frac{1}{n_r}\sum_{s_r}\left(e_i - \bar{e}\right)^2$ است. با استفاده از این اطلاعات، می‌توان نشان داد که

$$E\left(\hat{\sigma}_*^2\right) = \sigma^2\left(1 - \frac{c_1}{n-1} - \frac{4c_2}{\left(n-1\right)n_r} - c_3 f\,\frac{n-1}{n.n_r}\right)$$

که در آن c_1، c_2، c_3 در فاصله‌ی (۱و ۰) قرار می‌گیرند. بنا بر این، $E\left(\hat{\sigma}_*^2\right) \approx \sigma^2$ و رابطه‌ی (۳) برای مقادیر زیاد و متوسط n_r برقرار است.

در مرحله‌ی بعد، به $\mathrm{var}\left(\hat{\beta}^{*}\,|\,y_{\mathrm{obs}}\right)$ و $E\left(\hat{\beta}^{*}\,|\,y_{\mathrm{obs}}\right)$ نگاه می‌کنیم. می‌بینیم که $\hat{\beta}^{*}=\left(\hat{\beta}_{r}X_{r}+\hat{\beta}_{nr}X_{nr}\right)/X$ و $E\left(\hat{\beta}_{nr}\,|\,y_{\mathrm{obs}}\right)=\hat{\beta}_{r}+\left(\bar{e}\,/\,X_{nr}\right)\sum_{s-s_r}\sqrt{x_i}$ و $\mathrm{var}\left(\hat{\beta}_{nr}\,|\,y_{\mathrm{obs}}\right)=s_e^{2}/X_{nr}$ می‌باشد. این اطلاعات بیان می‌کند که $E\left(\hat{\beta}^{*}\,|\,y_{\mathrm{obs}}\right)=\hat{\beta}_{r}+\left(\bar{e}\,/\,X\right)\sum_{s-s_r}\sqrt{x_i}$ و $\mathrm{var}\left(\hat{\beta}^{*}\,|\,y_{\mathrm{obs}}\right)=\left(X_{nr}/X^{2}\right)s_e^{2}$ است. و این نتیجه می‌شود که

$$\mathrm{var}\,E\left(\hat{\beta}^{*}\,|\,y_{\mathrm{obs}}\right)=\mathrm{var}\left(\hat{\beta}_{r}\right)+\frac{\left(\sum_{s-s_r}\sqrt{x_i}\right)^{2}}{X^{2}}\mathrm{var}\left(\bar{e}\right)+2\frac{\sum_{s-s_r}\sqrt{x_i}}{X}\mathrm{cov}\left(\hat{\beta}_{r},\bar{e}\right)$$

اکنون، $\mathrm{cov}\left(\hat{\beta}_{r},\bar{e}\right)=0$ است. با استفاده از نابرابری کوشی شوارتس، $\left(\sum a_i b_i\right)^{2}\le\sum a_i^{2}\sum b_i^{2}$ با $a_i=\sqrt{x_i}$ و $b_i=1$، درمی‌یابیم که $\left(\sum_{i=1}^{n}\sqrt{x_i}\right)^{2}\le nX$ است. نتیجه این می‌شود که

$$\mathrm{var}\left(\bar{e}\right)=\left(\sigma^{2}/n_r\right)\left(1-\left(\Sigma_{s_r}\sqrt{x_i}\right)^{2}/n_r X_r\right)=\left(1-d_1\right)\sigma^{2}/n_r,\quad 0\le d_1\le 1$$

$$\left(\sum_{s-s_r}\sqrt{x_i}\right)^{2}/X^{2}=d_2 n_{nr}X_{nr}/X^{2},\qquad 0\le d_2\le 1$$

بنا بر این،

$$\mathrm{var}\,E\left(\hat{\beta}^{*}\,|\,y_{\mathrm{obs}}\right)=\frac{\sigma^{2}}{X_r}+\frac{\left(1-d_1\right)d_2 n_{nr}X_{nr}}{X^{2}}\cdot\frac{\sigma^{2}}{n_r}$$

سپس درمی‌یابیم که $E\left(s_e^{2}\right)=\sigma^{2}\left(1-\frac{1}{n_r}\right)-\mathrm{var}\left(\bar{e}\right)=\sigma^{2}\left(n_r+d_1-2\right)/n_r$ است که از این اطلاع خواهیم داشت

$$E\ \mathrm{var}\left(\hat{\beta}^{*}\,|y_{\mathrm{obs}}\right)=\frac{X_{nr}}{X^{۲}}\cdot\frac{\sigma^{۲}}{n_r}\left(n_r+d_{۱}-۲\right)$$

۲- جانهی چندگانه در رگرسیون خطی ساده در بخش۳-۳. خلاصه‌ای از اثبات:

در این‌جا $f=\left(n-n_r\right)/n$ ، $E\left(\hat{\sigma}_{*}^{۲}\right)=\sigma^{۲}E\left(\frac{n_r-۲}{n_r}\cdot\frac{n-۲f}{n-۲}\right)$

$\hat{\sigma}_{*}^{۲}=\frac{۱}{n-۲}\left(SS_e^{r}+SS_e^{nr}\right)$، که در آن $SS_e^{r}=\sum_{s_r}\left(y_i-\hat{\alpha}^{*}-\hat{\beta}^{*}x_i\right)^{۲}$ و $SS_e^{nr}=\sum_{s-s_r}\left(y_i^{*}-\hat{\alpha}^{*}-\hat{\beta}^{*}x_i\right)^{۲}$ است.

می‌توان دو مجموع مربعات مانده‌ها را به‌صورت زیر بیان کرد

$$SS_e^{r}=\sum_{s_r}\left(y_i-\hat{\alpha}_r-\hat{\beta}_r x_i\right)^{۲}+\sum_{s_r}\left[\left(\hat{\alpha}_r-\hat{\alpha}^{*}\right)^{۲}+(\hat{\beta}_r-\hat{\beta}^{*})x_i\right]^{۲}$$

$$SS_e^{nr}=\sum_{s-s_r}\left(y_i-\hat{\alpha}_{nr}^{*}-\hat{\beta}_{nr}^{*}x_i\right)^{۲}+\sum_{s-s_r}\left[\left(\hat{\alpha}_{nr}^{*}-\hat{\alpha}^{*}\right)+\left(\hat{\beta}_{nr}^{*}-\hat{\beta}^{*}\right)x_i\right]^{۲}$$

که در آن $\hat{\alpha}_{nr}^{*}$، $\hat{\beta}_{nr}^{*}$ براوردهایی هستند که تنها بر اساس y_i^{*} جانهی‌شده برای $i\in s-s_r$ به‌وجود آمده‌اند. از این مطلب می‌توان نتیجه گرفت که

$$E\left(SS_e^{r}\,|y_{\mathrm{obs}}\right)=n_r s_e^{۲}+n_r\ \mathrm{var}\left(\hat{\alpha}^{*}\,|y_{\mathrm{obs}}\right)+\left(\sum_{s-s_r}x_i^{۲}\right)\mathrm{var}\left(\hat{\beta}^{*}\,|y_{\mathrm{obs}}\right)$$
$$+۲n_r\bar{x}_r\ \mathrm{cov}\left(\hat{\alpha}^{*},\hat{\beta}^{*}\,|y_{\mathrm{obs}}\right)$$

$$= n_r s_e^2 + f(1-f)s_e^2 + (1-c_r)SS_x c_r s_e^2 / SS_x + 2n_r \bar{x}_r f \bar{x}_{nr} s_e^2 / SS_x$$

با $\left\{ \bar{x}_{nr} = \sum_{s-s_r} x_i / (n-n_r) \right\}$

$$\Rightarrow E\left(SS_e^r \mid y_{obs}\right) = s_e^2 \left\{ n_r + f(1-f) + c_r(1-c_r) + 2n_r \bar{x}_r \bar{x}_{nr} f(1/SS_x) \right\} \quad (22)$$

پس از بعضی محاسبات جبری، درمی‌یابیم که

$$\begin{aligned} E\left(SS_e^{nr} \mid y_{obs}\right) &= \sum_{s-s_r} \mathrm{var}\left(Y_i^* - \bar{Y}_{nr}^* \mid y_{obs}\right) + \sum_{s-s_r} (x_i - \bar{x}_{nr})^2 \mathrm{var}\left(\hat{\beta}_{nr}^* \mid y_{obs}\right) \\ &\quad - 2\sum_{s-s_r} (x_i - \bar{x}_{nr}) \mathrm{cov}\left(Y_i^* - \bar{Y}_{nr}^*, \hat{\beta}_{nr}^* \mid y_{obs}\right) \\ &\quad + (n-n_r)\mathrm{var}\left(\hat{\alpha}_{nr}^* - \hat{\alpha}^* \mid y_{obs}\right) + c_r SS_x \mathrm{var}\left(\hat{\beta}_{nr}^* - \hat{\beta}^* \mid y_{obs}\right) \\ &\quad + 2(n-n_r)\bar{x}_{nr} \mathrm{cov}\left(\hat{\alpha}_{nr}^* - \hat{\alpha}^*, \hat{\beta}_{nr}^* - \hat{\beta}^* \mid y_{obs}\right) \\ &= (n-n_r-1)s_e^2 + s_e^2 - 2s_e^2 + (n-n_r)\left\{ (1-f)^2 s_e^2 \frac{1}{n-n_r} + \bar{x}_{nr}^2 s_e^2 \frac{1}{SS_{x,nr}} \right\} \\ &\quad + c_r SS_x s_e^2 \left(\frac{1}{SS_{x,nr}} + \frac{c_r}{SS_x} - 2\frac{1}{SS_x} \right) \\ &\quad + 2(n-n_r)\bar{x}_{nr} \left(f\bar{x}_{nr} s_e^2 \frac{1}{SS_x} - \bar{x}_{nr} s_e^2 \frac{1}{SS_{x,nr}} \right) \end{aligned}$$

که در آن $SS_{x,nr} = \sum_{s-s_r} (x_i - \bar{x}_{nr})^2$. می‌بینیم که $SS_{x,nr} = \sum_{s-s_r} x_i^2 - (n-n_r)\bar{x}_{nr}^2 = c_r SS_x - (n-n_r)\bar{x}_{nr}^2$ و بنابراین $c_r SS_x / SS_{x,nr} = 1 + (n-n_r)\bar{x}_{nr}^2 / SS_{x,nr}$. از مطالب بالا نتیجه می‌شود که

$$E(SS_e^{nr} \mid y_{obs}) = s_e^2((n-n_r-1) + (1-f)^2 + c_r^2 - 2c_r + 2f\,\bar{x}_{nr}^2 (n-n_r)/SS_x) \quad (23)$$

از روابط (۲۲) و (۲۳) درمی‌یابیم که

$$(n-۲)E\left(\hat{\sigma}_*^۲\middle|y_{\text{obs}}\right)=s_e^۲\left(n-f-c_r+۲f\frac{۱}{SS_x}\bar{x}_{nr}n\bar{x}\right)=s_e^۲\left(n-f-c_r\right)$$

چون $E\left(c_r\middle|n_r\right)=\frac{۱}{SS_x}\sum_{i=۱}^{n}E\left(۱-R_i\middle|n_r\right)x_i^۲=\left(۱-n_r/n\right)=f$، داریم

$$\begin{aligned}(n-۲)E\left(\hat{\sigma}_*^۲\right)Es_e^۲\left(n-f-c_r\right)&=E\left(n-f-c_r\right)E\left(s_e^۲\middle|s_r\right)\\&=E\left(n-f-c_r\right)\frac{n_r-۲}{n_r}\sigma^۲\\&=\sigma^۲E\left\{\frac{n_r-۲}{n_r}\left(n-f-E\left(c_r\middle|n_r\right)\right)\right\}\\&=\sigma^۲E\left\{\frac{n_r-۲}{n_r}\left(n-۲f\right)\right\}\end{aligned}$$

۳- رگرسیون لوژستیک با متغیرهای تبیینی دودویی. ترکیب رده‌های مجزا

هدف تعیین $E\left(\hat{\beta}_۱^*\middle|y_{۱,\text{obs}}\right)$ و $\text{var}\left(\hat{\beta}_۱^*\middle|y_{۱,\text{obs}}\right)$ است.

به شرط $y_{۱,\text{obs}}$، $n_{۱۱}^*$ دوجمله‌ای با $\left(n_۱-n_{۱r},\hat{\pi}_{۱r}\right)$ است. بنا بر این، $E\left(n_{۱۱}^*\middle|y_{۱,\text{obs}}\right)=\left(n_۱-n_{۱r}\right)\hat{\pi}_{۱r}$ و $\text{var}\left(n_{۱۱}^*\middle|y_{۱,\text{obs}}\right)=\left(n_۱-n_{۱r}\right)\hat{\pi}_{۱r}\left(۱-\hat{\pi}_{۱r}\right)$ است. به شرط $y_{۱,\text{obs}}$، $\hat{\beta}_۱^*$ به‌صورت $T=\log\left\{\left(a+Z\right)/\left(b-Z\right)\right\}$ است که در آن Z دارای توزیع دوجمله‌ای با پارامترهای (n,p) و a و b مقادیر ثابت هستند. براساس خطی‌سازی تیلور در نزدیکی $E(Z)=np$ به رابطه‌ی $T\approx\log\left\{\left(a+np\right)/\left(b-np\right)\right\}+\left(Z-np\right)\left(a+b\right)/\left\{\left(a+z\right)\left(b-z\right)\right\}$ دست می‌یابیم و خواهیم داشت.

$$E(T) \approx \log \frac{a+np}{b-np}$$

(۲۴)

و

$$\operatorname{var}(T) \approx \left(\frac{a+b}{(a+np)(b-np)} \right)^{2} np(1-p)$$

با درنظر گرفتن $a = n_{11r}$ و $b = n_{1} - n_{11r}$ به نتیجه‌ی $E\left(\hat{\beta}_{1}^{*} \middle| y_{1,\text{obs}}\right) \approx \hat{\beta}_{1r}$ و $\operatorname{var}\left(\hat{\beta}_{1}^{*} \middle| y_{1,\text{obs}}\right) \approx \left(n_{1} / \left\{n_{1}\hat{\pi}_{1r} n_{1}(1-\hat{\pi}_{1r})\right\}\right)^{2} (n_{1} - n_{1r})\hat{\pi}_{1r}(1-\hat{\pi}_{1r})$ دست می‌یابیم. فرض کنید $f_{1} = (n_{1} - n_{1r}) / n_{1}$ نرخ بی‌پاسخی در طبقه‌ی s_1 باشد. بنا بر این

$$\begin{aligned}\operatorname{var}\left(\hat{\beta}_{1}^{*} \middle| y_{1,\text{obs}}\right) &\approx \frac{f_{1} n_{1}}{n_{1}^{2}} \cdot \frac{1}{\hat{\pi}_{1r}(1-\hat{\pi}_{1r})} = f_{1}(1-f_{1}) \cdot \frac{1}{n_{1r}\hat{\pi}_{1r}(1-\hat{\pi}_{1r})} \\ &= f_{1}(1-f_{1}) \hat{\mathrm{V}}\left(\hat{\beta}_{1r}\right)\end{aligned}$$

۴- رگرسیون لوژستیک با مقادیر گم‌شده در یک متغیر تبیینی دودویی

به‌منظور تعیین $\operatorname{var}\left(\hat{\beta}_{*} \middle| y, x_{\text{obs}}\right)$ و $E\left(\hat{\beta}_{*} \middle| y, x_{\text{obs}}\right)$، نیاز است که $\hat{\beta}_{*}$ به‌روشی متفاوت با روش بخش ۲-۵ بیان شود تا بتواند مجموع دو عبارت مستقل به شرط داده‌های مشاهده‌شده‌ی (y, x_{obs}) باشد:

$$\hat{\beta}_{*} = \log \frac{(n_{11r} + n_{11}^{*})(n_{00r} + n_{00}^{*})}{(n_{10r} + n_{10}^{*})(n_{01r} + n_{01}^{*})} = \log \frac{(n_{11r} + n_{11}^{*})}{(n_{01r} + n_{01}^{*})} - \log \frac{(n_{10r} + n_{10}^{*})}{(n_{00r} + n_{00}^{*})} = \hat{\beta}_{*}^{1} - \hat{\beta}_{*}^{0}$$

و $\operatorname{var}\left(\hat{\beta}_{*} \middle| y, x_{\text{obs}}\right) = \operatorname{var}\left(\hat{\beta}_{*}^{1} \middle| y, x_{1,\text{obs}}\right) + \operatorname{var}\left(\hat{\beta}_{*}^{0} \middle| y, x_{0,\text{obs}}\right)$. به شرط (y, x_{obs})، n_{11}^{*} دارای توزیع دوجمله‌ای با $\left(n_{1}^{0} - n_{1r}^{0}, p^{1}\right)$ است که در آن $p^{1} = n_{11r} / n_{1r}^{0}$ است و n_{10}^{*} دارای توزیع دوجمله‌ای با $\left(n_{0}^{0} - n_{0r}^{0}, p^{0}\right)$ است که در آن $p^{0} = n_{10r} / n_{0r}^{0}$ می‌باشد. پس با استفاده از رابطه‌ی (۲۴)، می‌توان به این نتیجه رسید که به طور تقریبی $E\left(\hat{\beta}_{*}^{1} \middle| y, x_{1,\text{obs}}\right) = \log\left\{p^{1} / (1-p^{1})\right\}$ و

$\operatorname{var}\left(\hat{\beta}_{*}^{(1)} \middle| y, x_{1,\mathrm{obs}}\right)=\left(n_{1}^{(0)} / \left\{n_{1}^{(0)} p^{(1)} n_{1}^{(0)}\left(1-p^{(1)}\right)\right\}\right)^{2}\left(n_{1}^{(0)}-n_{1r}^{(0)}\right) p^{(1)}\left(1-p^{(1)}\right)$ است.

بنابراین $\operatorname{var}\left(\hat{\beta}_{*}^{(1)} \middle| y, x_{1,\mathrm{obs}}\right) \approx f^{(1)} /\left\{n_{1}^{(0)} p^{(1)}\left(1-p^{(1)}\right)\right\}=f^{(1)}\left(1-f^{(1)}\right) /\left\{n_{1r}^{(0)} p^{(1)}\left(1-p^{(1)}\right)\right\}$.

به همین ترتیب $E\left(\hat{\beta}_{*}^{(0)} \middle| y, x_{1,\mathrm{obs}}\right) \approx \log\left\{p^{(0)} /\left(1-p^{(0)}\right)\right\}$ است به گونه‌ای که $E\left(\hat{\beta}_{*} \middle| y, x_{\mathrm{obs}}\right) \approx \hat{\beta}_{r}$. همچنین

$$\operatorname{var}\left(\hat{\beta}_{*}^{(0)} \middle| y, x_{0,\mathrm{obs}}\right) \approx f^{(0)} /\left\{n_{0}^{(0)} p^{(0)}\left(1-p^{(0)}\right)\right\}=f^{(0)}\left(1-f^{(0)}\right) /\left\{n_{0r}^{(0)} p^{(0)}\left(1-p^{(0)}\right)\right\}$$

می‌دانیم $\frac{1}{n_{1r}^{(0)} p^{(0)}\left(1-p^{(0)}\right)}=\frac{1}{n_{10r}}+\frac{1}{n_{00r}}$ و $\frac{1}{n_{1r}^{(0)} p^{(1)}\left(1-p^{(1)}\right)}=\frac{n_{1r}^{(0)}}{n_{11r} n_{01r}}=\frac{1}{n_{11r}}+\frac{1}{n_{01r}}$

و از این‌ها نتیجه می‌شود که

$$\operatorname{var}\left(\hat{\beta}_{*} \middle| y, x_{\mathrm{obs}}\right) \approx f^{(1)}\left(1-f^{(1)}\right)\left(\frac{1}{n_{11r}}+\frac{1}{n_{01r}}\right)+f^{(0)}\left(1-f^{(0)}\right)\left(\frac{1}{n_{10r}}+\frac{1}{n_{00r}}\right).$$

۵- اثبات لم و قضیه‌ی مربوط به بخش ۶

به منظور اثبات لم و قضیه در بخش ۶، به دانستن بعضی واقعیت‌ها نیاز داریم. در همه‌ی سه حالت مطرح‌شده در بخش ۶؛

۱- n_r دارای توزیع دوجمله‌ای با (n, p_r) و مستقل از s می‌باشد.

۲- s_r به شرط n_r, s نمونه‌ای تصادفی ساده از s به‌اندازه‌ی n_r است.

۳- $P\left(R_i=1 \middle| n_r\right)=n_r / n$ و $P\left(R_i=1, R_j=1 \middle| n_r\right)=\frac{n_r}{n} \cdot \frac{n_r-1}{n-1}$ (از رابطه‌ی ۲ به‌دست می‌آید)

۴- $E\left(Y_i^* \middle| y_{\mathrm{obs}}\right)=\bar{y}_r \left(\Rightarrow E\left(Y_i^* \middle| y, s, n_r\right)=\bar{y}_s \Rightarrow E\left(Y_i^* \middle| y, s\right)=\bar{y}_s\right)$

۵- $\operatorname{var}\left(Y_i^* \middle| y_{\mathrm{obs}}\right)=\frac{n_r-1}{n_r} \hat{\sigma}_r^2$ که در آن $\hat{\sigma}_r^2=\frac{1}{n_r-1} \sum_{i \in s_r}\left(y_i-\bar{y}_r\right)^2$

۶- $E\left(\hat{\sigma}_r^{۲}\middle|y,s,n_r\right)=\hat{\sigma}^{۲}$ که در آن $\hat{\sigma}^{۲}=\frac{۱}{n-۱}\sum_{i\in s}\left(y_i-\bar{y}_s\right)^{۲}$

$$\left(\Rightarrow \operatorname{var}\left(Y_i^{*}\middle|y,s,n_r\right)=\frac{n-۱}{n}\hat{\sigma}^{۲}\approx\hat{\sigma}^{۲}\right)$$

۷- $$\operatorname{var}\left(\bar{Y}_r\middle|y,s,n_r\right)=f\,\hat{\sigma}^{۲}/n_r=\hat{\sigma}^{۲}\left(\frac{۱}{n_r}-\frac{۱}{n}\right)$$

اثبات لم:

$$E\left(\hat{\theta}^{*}\middle|y,s\right)=E\left\{E\left(\sum_{i\in s_r}a_i(s)y_i+\sum_{i\in s-s_r}a_i(s)Y_i^{*}\middle|Y_{\text{obs}}\right)\middle|y,s\right\}$$

$$=^{(۴)}E\left(\sum_{i\in s_r}a_i(s)y_i\middle|y,s\right)+E\left(\sum_{i\in s-s_r}a_i(s)\bar{Y}_r\middle|y,s\right)$$

عبارت اول:

$$E\left(\sum_{i\in s_r}a_i(s)y_i\middle|y,s\right)=E\left\{E\left(\sum_{i\in s}a_i(s)y_iR_i\middle|y,s,n_r\right)\middle|y,s\right\}$$

$$=^{(۳)}E\left(\sum_{i\in s}a_i(s)y_i\frac{n_r}{n}\middle|y,s\right)=^{(۱)}p_r\hat{\theta}$$

عبارت دوم:

$$E\left(\sum_{i\in s-s_r}a_i(s)\bar{Y}_r\middle|y,s\right)=E\left\{E\left(\frac{۱}{n_r}\sum_{i\in s}\sum_{j\in s}a_i(s)y_j(۱-R_i)R_j\middle|y,s,n_r\right)\middle|y,s\right\}$$

$$=^{(۳)}E\left(\frac{۱}{n_r}\sum_{i\in s}\sum_{j\in s,j\neq i}a_i(s)y_j\left(\frac{n_r}{n}-\frac{n_r}{n}.\frac{n_r-۱}{n-۱}\right)\middle|y,s\right)$$

$$=^{(۱)}\frac{۱-p_r}{n-۱}\sum_{i\in s}\sum_{j\in s,j\neq i}a_i(s)y_j=\frac{۱-p_r}{n-۱}\left(n\bar{a}(s)n\bar{y}_s-\hat{\theta}\right)$$

که در آن $\bar{a}(s)=\sum_{i\in s}a_i(s)/n$.

این به آن معنی است که $E\left(\hat{\theta}^*|y,s\right)=p_r\hat{\theta}+\frac{1-p_r}{n-1}\left(n^2\bar{a}(s)\bar{y}_s-\hat{\theta}\right)$ و رابطه‌ی (۲۱)

$$\hat{\theta}=n\bar{a}(s)\bar{y}_s=\bar{a}(s)\sum_{i\in s}y_i \Leftrightarrow$$

اثبات قضیه:

با استفاده از لم، $\hat{\theta}=a(s)\sum_{i\in s}y_i=na(s)\bar{y}_s$ و $\hat{\theta}^*=a(s)\left(\sum_{i\in s_r}y_i+\sum_{i\in s-s_r}y_i^*\right)$

است.

$$E\left(\hat{\theta}^*|y_{\text{obs}}\right)=^{(۴)}a(s)\left(n_r\bar{y}_r+\left(n-n_r\right)\bar{y}_r\right)=na(s)\bar{y}_r$$

$$\text{var}\left(\hat{\theta}^*|y_{\text{obs}}\right)=^{(۵)}\left\{a(s)\right\}^2\left(n-n_r\right)\frac{n_r-1}{n_r}\hat{\sigma}_r^2$$

بنا بر این،

$$\begin{aligned}\text{var}E\left(\hat{\theta}^*|Y_{\text{obs}}\right)&=\text{var}\left[na(s)\bar{\text{Y}}_r\right]=E\left[n^2\left\{a(s)\right\}^2\text{var}\left(\bar{\text{Y}}_r|Y,s\right)\right]\\&\quad+\text{var}\left\{na(s)E\left(\bar{\text{Y}}_r|Y,s\right)\right\}\\&=n^2E\left(\left[a(s)\right]^2\left\{E\left\{\text{var}\left(\bar{\text{Y}}_r|Y,s,n_r\right)|Y,s\right\}+\text{var}\left\{E\left(\bar{\text{Y}}_r|Y,s,n_r\right)|Y,s\right\}\right\}\right)\\&\quad+\text{var}\left\{na(s)E\left\{E\left(\bar{\text{Y}}_r|Y,s,n_r\right)|Y,s\right\}\right\}\\&=^{(۷)}n^2E\left(\left[a(s)\right]^2\left\{\hat{\sigma}^2E\left\{(1/n_r-1/n)|s\right\}\right\}+\text{var}\left(\bar{\text{Y}}_s|Y,s\right)\right)+\text{var}\left\{na(s)\bar{\text{Y}}_s\right\}\\&=nE\left(\left[a(s)\right]^2\hat{\sigma}^2\left[E\left(1/\hat{p}_r|s\right)-1\right]+0\right)\\&\quad+\text{var}\,\hat{\theta}=^{(۱)}n\left\{E\left(1/\hat{p}_r\right)-1\right\}E\left(\left[a(s)\right]^2\hat{\sigma}^2\right)+\text{var}\,\hat{\theta}\end{aligned}$$

سپس،

$$
\begin{aligned}
E\ \mathrm{var}\left(\hat{\theta}^{*}\,|\,Y_{\mathrm{obs}}\right) &= E\left(\left[a(s)\right]^{۲}\left(n-n_r\right)\frac{n_r-۱}{n_r}\hat{\sigma}_r^{۲}\right)\\
&= E\left\{\left(n-n_r\right)\left(۱-۱/n_r\right)\left[a(s)\right]^{۲}E\left(\hat{\sigma}_r^{۲}\,|\,Y,s,n_r\right)\right\}\\
&= E\left\{\left(n-n_r\right)\left(۱/n_r-۱\right)\left[a(s)\right]^{۲}\hat{\sigma}^{۲}\right\}\\
&=^{(۱)}\left[n\left(۱-p_r\right)-\left\{E\left(۱/\hat{p}_r\right)-۱\right\}\right]E\left(\left[a(s)\right]^{۲}\hat{\sigma}^{۲}\right)
\end{aligned}
$$

از رابطه‌ی (۵) دریافتیم که

$$
\begin{aligned}
E\left(k\right) &= \frac{\left\{E\left(۱/\hat{p}_r\right)-۱\right\}E\left\{\left[a(s)\right]^{۲}\hat{\sigma}^{۲}\right\}}{\left[\left(۱-p_r\right)-\frac{۱}{n}\left\{E\left(۱/\hat{p}_r\right)-۱\right\}\right]E\left(\left[a(s)\right]^{۲}\hat{\sigma}^{۲}\right)}\\
&= \frac{\left\{E\left(۱/\hat{p}_r\right)-۱\right\}}{۱-p_r-\frac{۱}{n}\left\{E\left(۱/\hat{p}_r\right)-۱\right\}}\\
&\approx \frac{\left(۱/p_r\right)-۱}{۱-p_r}=\frac{۱}{p_r}
\end{aligned}
$$

زهرا رضایی قهرودی
دکتری آمار
تهران، خیابان دکتر فاطمی، خیابان باباطاهر، خیابان شهید فکوری، شماره‌ی ۱۴۵، پژوهشکده‌ی آمار.
رایانشانی: z_rezaei@srtc.ac.ir