

جانهای چندگانه‌ی غیر بیزی

جان جورن استد[†]

اداره‌ی آمار نروژ

مترجم: زهرا رضایی قهرودی

پژوهشکده‌ی آمار

چکیده: جانهای چندگانه روشی است که به طور مشخص برای برآورد واریانس در حضور داده‌های گم‌شده طراحی شده است. فرمول ترکیبی روبین نیازمند آن است که روش جانهای «مناسب» باشد، که اساساً به این معنی است که جانهایها، 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 جانهای بی‌پاسخی استفاده می‌شوند، مانند بعضی از انواع جانهای بی‌درنگ طبقه‌بندی‌شده، نوعاً غیر بیزی هستند. بنا بر این روش جانهای 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 جانهای 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 جانهای از نوع بی‌درنگ تعیین شده‌اند.

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

واژگان کلیدی: برآورد واریانس؛ آمارگیری نمونه‌ای؛ نمونه‌گیری طبقه‌بندی‌شده؛ رگرسیون لوژیستیک؛ بی‌پاسخی، جانهای بی‌درنگ.

دریافت: ۱۳۸۸/۲/۲، پذیرش: ۱۳۸۸/۵/۱۷

۱- مقدمه

جانهدی چندگانه روشی است که به طور مشخص برای برآورد 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 جانهدی شده برای هر مقدار گم‌شده و ترکیب m مجموعه‌ی داده‌های کامل شده به‌وسیله‌ی فرمول ترکیبی روبین برای برآورد واریانس است. برای این که برآوردگر معتبر باشد، جانهدی باید سطح مناسبی از تغییرپذیری را نمایش دهد. در اصطلاح روبین، یکی از الزامات روش جانهدی «مناسب بودن» است. در مؤسسات ملی آمار روش‌هایی که برای جانهدی بی‌پاسخی استفاده می‌شود، حتی اگر الزام «مناسب بودن» را داشته باشند، خیلی کم این نیاز را تأمین می‌کنند. اما ایده‌ی ایجاد جانهدی چندگانه برای اندازه‌گیری عدم حتمیت جانهدی و استفاده از آن برای برآورد واریانس و محاسبه‌ی بازه‌ی اطمینان هنوز مورد علاقه است. مشکل این است که فرمول ترکیبی روبین حالا دیگر برای جانهدی‌های نامناسب معمول که توسط مؤسسات ملی آمار استفاده می‌شود، معتبر نیست. دلیل آن این است که تغییرپذیری در جانهدی‌های نامناسب خیلی کم است و باید برای مؤلفه‌ی بین جانهدی، وزن بیش‌تری در برآورد واریانس در نظر گرفته شود. مشکل دیگر این است که مشخص شود این وزن چقدر باید باشد تا استنباط آماری معتبری ارائه دهد و همچنین برای چه نوع مکانیسم‌های بی‌پاسخی و مسایل برآورد، 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 جانهدی پیشنهاد می‌شود. بخش سوم سه کاربرد با بی‌پاسخی تصادفی را بیان می‌کند:

الف) برآورد میانگین جامعه از روی نمونه‌های تصادفی ساده با استفاده از روش جانهدی بی‌درنگ؛

ب) برآورد ضریب رگرسیونی در مدل نسبت، با استفاده از جانهدی رگرسیونی مانده‌ها؛

پ) برآورد ضریب رگرسیونی در رگرسیون خطی ساده با جانهدی رگرسیونی مانده‌ها.

بخش چهارم به مسئله‌ی کلی جانهی چندگانه برای نمونه‌های طبقه‌بندی شده 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) برآورد میانگین جامعه با استفاده از جانهی بی‌درنگ طبقه‌بندی شده؛

ب) برآورد 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 برآورد با یک روش جانهی مشخص و داده‌ها و مدل پاسخ معلوم را بیان می‌کند. یک نتیجه‌ی کلی برای جانهی بی‌درنگ و برآورد خطی نیز ارائه شده است.

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

فرض کنید $s = (1, \dots, n)$ معرف کل نمونه با داده‌های کل نمونه‌ی $y = (y_1, \dots, y_n)$ از مقادیر متغیرهای تصادفی Y_1, \dots, Y_n باشند. در نمونه‌گیری از جامعه‌ی متناهی تحت یک مدل طرح، شماره‌گذاری مجدد واحدهای انتخاب شده، انجام می‌شود و ماهیت تصادفی بودن y به وسیله‌ی طرح نمونه‌گیری تعیین می‌شود. هدف برآورد برخی از پارامترهای θ است. داده‌های مشاهده شده که به وسیله‌ی $\{y_i : i \in s_r, s_r\}$ نمایش داده می‌شوند، بخش مشاهده شده‌ی y و نمونه‌ی پاسخ s_r از اندازه‌ی n_r است. فرض کنید $\hat{\theta}$ برآوردگری براساس داده‌های کل نمونه‌ی y باشد که با $\text{var}(\hat{\theta})$ به وسیله‌ی $\hat{V}(y)$ برآورد شده است. برای $i \in s - s_r$ ، y_i^* را به وسیله‌ی 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 $\hat{V}^* = \hat{V}(y^*)$. جانهی چندگانه‌ی m جانهی تکرار شده، منجر به m مجموعه‌ی داده‌های کامل شده با m برآورد $i = 1, \dots, m$ و $\hat{\theta}_i^*$ ، و برآوردهای واریانس مربوط به آن‌ها، \hat{V}_i^* ، برای $i = 1, \dots, m$ می‌شود. برآورد ترکیب شده به وسیله‌ی $\bar{\theta}^* = \sum_{i=1}^m \hat{\theta}_i^* / m$ تعیین می‌شود.

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

$$(۱) \quad W = \bar{V}^* + (k + \frac{1}{m})B^* .$$

پیشنهاد می‌شود. همچنین لازم است k را به گونه‌ای تعیین کنیم که

$$(۲) \quad E(W) = \text{var}(\bar{\theta}^*) .$$

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.

به طور کلی، باید عبارت (۲) را تعیین کرد. یک راه آزمایش و انجام آن، استفاده از امید ریاضی دوگانه به شرط y_{obs} ، است به طوری که $E(W) = E\{E(W | Y_{\text{obs}})\}$ و

$$\text{var}(\bar{\theta}^*) = E\{\text{var}(\bar{\theta}^* | Y_{\text{obs}})\} + \text{var}\{E(\bar{\theta}^* | Y_{\text{obs}})\} \text{ نوعاً}$$

$$(۳) \quad E(\bar{V}^*) \approx \text{var}(\hat{\theta})$$

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

$$(۴) \quad E(W) = \text{var}(\hat{\theta}) + \left(E(k) + \frac{1}{m} \right) E \text{var}(\hat{\theta}^* | Y_{\text{obs}})$$

است. علاوه بر این، $\text{var}(\bar{\theta}^* | y_{\text{obs}}) = \frac{\text{var}(\hat{\theta}^* | y_{\text{obs}})}{m}$ و $E(\bar{\theta}^* | y_{\text{obs}}) = E(\hat{\theta}^* | y_{\text{obs}})$

این حاکی از آن است که $\text{var}(\bar{\theta}^*) = m^{-1} E\{\text{var}(\hat{\theta}^* | Y_{\text{obs}})\} + \text{var}\{E(\hat{\theta}^* | Y_{\text{obs}})\}$ با استفاده از روابط (۳) و (۴)، معادله‌ی (۲) به صورت

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

در می آید و عبارت کلی زیر را ارائه می دهد.

$$(5) \quad E(k) = \frac{\text{var} E(\hat{\theta}^* | Y_{\text{obs}}) - \text{var}(\hat{\theta})}{E \text{var}(\hat{\theta}^* | Y_{\text{obs}})}$$

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 این که چگونه رابطه‌ی (۵) می تواند استفاده شود، در بخش بعدی سه حالت خاص با بی پاسخی تصادفی را مورد بررسی قرار می دهیم.

۳- سه کاربرد برای بی پاسخی تصادفی

۳-۱- برآورد میانگین جامعه با جانہی بی درنگ

یک نمونه‌ی تصادفی ساده از جامعه‌ی متناهی به اندازه‌ی N را در نظر می گیریم، که هدف آن برآورد میانگین جامعه، μ از متغیر y است. همچنین فرض بی پاسخی کاملاً تصادفی نیز در نظر گرفته شده است. در اصطلاح‌شناسی روبین [۲] و لیتل و روبین [۱]، مکانیسم گم‌شدگی، گم شدن کاملاً تصادفی (Missing Completely at Random) MCAR نامیده می شود. گم شدن کاملاً تصادفی به این معنا است که نماگرهای پاسخ R_1, \dots, R_N با احتمال‌های پاسخ مشابه $\Pr = P(R_i = 1)$ مستقل هستند. روش جانہی، روش بی درنگ است که y_i^* به تصادف از y_{obs} با جایگذاری قرعه‌کشی شده است و برآورد، میانگین نمونه است. فرض کنید \bar{y}_r میانگین نمونه‌ی مشاهده شده و

$$\hat{\sigma}_r^2 = \frac{1}{n_r - 1} \sum_{i \in S_r} (y_i - \bar{y}_r)^2$$

واریانس نمونه‌ی مشاهده شده باشد. پس \bar{Y}^*

میانگین نمونه بر اساس جانہی برای نمونه‌ی کامل شده است، و برآوردگر ترکیب شده به وسیله‌ی $\bar{Y}^* = \sum_{i=1}^m \bar{Y}_i^* / m$ تعیین می شود. فرض کنید \bar{Y}_s معرف میانگین نمونه بر اساس کل نمونه باشد. بنا بر این $\text{var}(\bar{Y}_s) = \sigma^2 \left(\frac{1}{n} - \frac{1}{N} \right)$ است که در آن

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

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

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

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

۲-۳- برآورد ضریب رگرسیون در مدل نسبت با جانهای 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 $\text{var}(\varepsilon_i) = \sigma^2 x_i; i = 1, \dots, n$ را در نظر می‌گیریم. فرض شده است که همه‌ی x_i ها، همچنین در نمونه‌ی بی‌پاسخ

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

روش جانہی رگرسیون مانده‌ها در نظر گرفته شده است. فرض کنید $\hat{\beta}_r$ ، برآورد $\hat{\beta}$ 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 y_i جانہی شده به‌وسیله‌ی $y_i^* = \hat{\beta}_r x_i + e_i^* \sqrt{x_i}$ مشخص می‌شود.

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

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

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

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

$$(۷) \quad Evar(\hat{\beta}^* | y_{\text{obs}}) = \frac{X_{nr}}{X^2} \cdot \frac{\sigma^2}{n_r} (n_r + d_1 - 2)$$

که در آن $0 \leq d_1 \leq 1$ ، $d_2 \leq 1$. 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 در نظر 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- برآورد ضریب رگرسیون در رگرسیون خطی ساده با جانمایی مانده‌ها

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

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

جانمایی مانده‌های ساده به صورت زیر تعریف می‌شود؛ مانده‌های مشاهده‌شده برای $e_j = (y_j - \hat{\alpha}_r - \hat{\beta}_r x_j)$ ، $j \in S_r$ است. برای $i \in S - S_r$ ؛ e_i^* به تصادف و با

جایگذاری از $(e_j, j \in s_r)$ قرعه کشی شده است. مقدار y_i جانہی شده به وسیلهی

$$y_i^* = \hat{\alpha}_r + \hat{\beta}_r \bar{x}_i + e_i^*$$

برآوردهای بر اساس جانہی 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 $\hat{\alpha}^* = (n_r \bar{y}_r + (n - n_r) \bar{y}_{nr}^*) / n$

و $\bar{y}_{nr}^* = \sum_{s-s_r} y_i^* / (n - n_r)$ می توان نشان داد

که (برای خلاصه‌ی اثبات به پیوست ۲ نگاه کنید)

$$E(\hat{\sigma}_*^2) = \sigma^2 E \left(\frac{n_r - 2}{n_r} \cdot \frac{n - 2f}{n - 2} \right) \approx \sigma^2$$

شد $f = (n - n_r) / n$ است. چون $\text{var}(\hat{\beta}) = \sigma^2 / SS_x$ بنا بر این رابطه‌ی (۳) برقرار

است. به سهولت دیده می شود 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 $c_r = \sum_{s-s_r} x_i^2 / SS_x \in \langle 0, 1 \rangle$ می توان نشان داد که

$$E(c_r) = 1 - p_r \quad E(s_e^2 | s_r) = \frac{n_r - 2}{n_r} \sigma^2$$

و $\text{var}(\hat{\beta}_r | s_r) = \sigma^2 / SS_{x,r}$ از رابطه‌ی (۵) نتیجه می شود 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 به گونه‌ای که

نماگرهای پاسخ آن با $\text{cov}(R_i, R_j) = -f(1-f)/(n-1)$ همبسته‌اند، درمی یابیم که

$$E(SS_{x,r}) = \left(p_r - \frac{1 - p_r}{n - 1} \right) SS_x$$

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

۴- جانهای چندگانه برای نمونه‌های طبقه‌بندی شده

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 k می‌جزا برای هر طبقه تعیین شود. پس ساختار کلی به صورت زیر است؛ نمونه‌ی s به H طبقه‌ی نمونه به صورت s_1, \dots, s_H تقسیم می‌شود. فرض کنید y_h کل داده‌های طرح ریزی شده از زیرنمونه‌ی s_h به اندازه‌ی n_h باشد. فرض شده است که y_1, \dots, y_H مستقل هستند. بخش مشاهده‌شده‌ی y_h با نماد $y_{h,obs}$ نمایش داده می‌شود که s_{hr} نمونه‌ی پاسخ از s_h به اندازه‌ی n_{hr} است. برآوردگر 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. $\text{var}(\hat{\theta}) = \sum_{h=1}^H \text{var}(\hat{\theta}_h)$ از طریق $\hat{V}(\hat{\theta}) = \sum_{h=1}^H \hat{V}_h(y_h)$ برآورد می‌شود که در آن $\hat{V}_h(y_h)$ برآورد واریانس $\hat{\theta}_h$ بر اساس y_h است. برای $i \in s_h - s_{hr}$ ، y_i^* با استفاده از برخی روش‌ها بر اساس $y_{h,obs}$ جانهای می‌شود و y_h^* را به عنوان داده‌های کامل در نظر می‌گیریم ($y_{h,obs}, y_i^*, i \in s_h - s_{hr}$). بر اساس y_h^* ، داریم $\hat{\theta}_h^* = \hat{\theta}_h(y_h^*)$ و $\hat{V}_h^* = \hat{V}_h(y_h^*)$. سپس برآوردگر مبتنی بر جانهای به وسیله‌ی $\hat{\theta}^* = \sum_{h=1}^H \hat{\theta}_h^*$ و $\hat{V}^* = \sum_{h=1}^H \hat{V}_h^*$ مشخص می‌شوند. جانهای چندگانه‌ی m جانهای تکرار شده به m مجموعه‌ی داده‌های کامل شده با m برآورد برای هر طبقه‌ی h ، به صورت $\hat{\theta}_{h,i}$ و برآورد واریانس مربوط $\hat{V}_{h,i}^*$ برای $i = 1, \dots, m$ منجر می‌شود. برآوردهای کل و واریانس‌های مربوط به صورت $\hat{\theta}_i^* = \sum_{h=1}^H \hat{\theta}_{h,i}^*$ برای $i = 1, \dots, m$ و $\hat{V}_i^* = \sum_{h=1}^H \hat{V}_{h,i}^*$ برای $i = 1, \dots, m$ است. برآورد ترکیب‌شده برای طبقه‌ی h ام به وسیله‌ی $\bar{\theta}_h^* = \sum_{i=1}^m \hat{\theta}_{h,i}^* / m$ مشخص می‌شود. واریانس داخل جانهای برای طبقه‌ی h ام عبارت است از $\bar{V}_h^* = \sum_{i=1}^m \hat{V}_{h,i}^* / m$ و مؤلفه‌ی بین جانهای به وسیله‌ی رابطه‌ی $B_h^* = \sum_{i=1}^m (\hat{\theta}_{h,i}^* - \bar{\theta}_h^*)^2 / (m-1)$ مشخص می‌شود. با دنبال کردن همان ایده‌ی بخش ۲،

رابطه‌ی (۱)، واریانس برآوردشده‌ی کل $\bar{\theta}_h^*$ ، به صورت $W_h = \bar{V}_h^* + (k_h + \frac{1}{m})B_h^*$ پیشنهاد می‌شود. برآورد کل ترکیب‌شده نیز به وسیله‌ی $\bar{\theta}^* = \sum_{i=1}^m \hat{\theta}_i^* / m = \sum_{h=1}^H \bar{\theta}_h^*$ مشخص می‌شود. نتیجه می‌گیریم که واریانس برآوردشده‌ی کل $\bar{\theta}^*$ نیز می‌تواند به صورت رابطه‌ی زیر بیان شود.

$$(۸) \quad W_{\text{sep}} = \sum_{h=1}^H W_h = \bar{V}^* + \sum_{h=1}^H (k_h + \frac{1}{m})B_h^*$$

که در آن $\bar{V}^* = \sum_{i=1}^m \hat{V}_i^* / m = \sum_{h=1}^H \bar{V}_h^*$ است. اگر رابطه‌ی (۳) برای هر طبقه‌ی h برقرار باشد

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

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

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

فرمول ترکیبی (۸) یک شق دیگر برای فرمول ترکیبی معمول (۱) است که به خصوص 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

$$(۱۱) \quad E \left\{ \sum_{h=1}^H \left(k_h + \frac{1}{m} \right) B_h^* \right\} = E \left(k + \frac{1}{m} \right) B^*$$

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 $E(B^* | y_{obs}) = \text{var}(\hat{\theta}^* | y_{obs}) = \sum_{h=1}^H \text{var}(\hat{\theta}_h^* | y_{obs})$ و همچنین $E(B_h^* | y_{obs}) = \text{var}(\hat{\theta}_h^* | y_{obs})$ بنا بر این

$$E(B^* | y_{obs}) = E\left(\sum_{h=1}^H B_h^* | y_{obs}\right).$$

بنا بر این اتحاد (۱۱)، به صورت

$$E\left\{\sum_{h=1}^H k_h E(B_h^* | Y_{obs})\right\} = E\{k E(B^* | Y_{obs})\}$$

می‌شود. در صورتی که بخواهیم از فرمول ترکیبی معمول (۱) استفاده کنیم، به حل

$$k = \sum_{h=1}^H k_h E(B_h^* | y_{obs}) / E(B^* | y_{obs})$$

دست می‌یابیم و لذا

$$(۱۲) \quad k = \frac{\sum_{h=1}^H k_h \text{var}(\hat{\theta}_h^* | y_{obs})}{\text{var}(\hat{\theta}^* | y_{obs})} = \sum_{h=1}^H k_h \cdot \frac{\text{var}(\hat{\theta}_h^* | y_{obs})}{\text{var}(\hat{\theta}^* | y_{obs})}$$

که این رابطه یک میانگین موزون از k_h است. زمانی که همه k_h ها برابر باشند، مثلاً $k_h = k$ ، به عبارت ساده‌ای برای k یعنی $k = k_0$ دست می‌یابیم.

۵- چهار کاربرد برای نمونه‌های طبقه‌بندی شده و بی‌پاسخی تصادفی درون طبقات

۵-۱- برآورد میانگین جامعه از نمونه‌ی طبقه‌بندی شده با جانهای بی‌درنگ طبقه‌بندی شده

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

داشته باشد. این مطلب به این معنا است که نماگرهای پاسخ در طبقه‌ی h ام، $R_{h,1}, \dots, R_{h,N_h}$ با احتمال‌های $p_{hr} = P(R_{h,i} = 1)$ از هم مستقل هستند. روش جانهی، روش بی‌درنگ طبقه‌بندی شده است. فرض کنید $y_{h,obs}$ بخش مشاهده‌شده از نمونه‌ی پاسخ s_{hr} به‌اندازه‌ی n_{hr} از طبقه‌ی h باشد، $y_{h,obs} = (y_i : i \in s_{hr})$. سپس مقدار جانهی‌شده‌ی y_i^* در طبقه‌ی h ام به‌تصادف از $y_{h,obs}$ قرعه‌کشی می‌شود. برآوردگر مبتنی بر داده‌های کل نمونه، میانگین موزون طبقه‌بندی‌شده‌ی معمول $\bar{Y}_{strat} = \sum_{h=1}^H N_h \bar{y}_h / N = \sum_{h=1}^H v_h \bar{y}_h$ 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 h ام و $n_h = |s_h|$ است. پس $\text{var}(\bar{Y}_{strat}) = \sum_{h=1}^H v_h^2 \sigma_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right)$ با $\sigma_h^2 = \sum_{i \in U_h} (y_i - \mu_h)^2 / (N_h - 1)$ واریانس جامعه در طبقه‌ی h ام می‌باشد. در این‌جا U_h ، جمعیت طبقه‌ی h ام و μ_h میانگین در U_h است.

فرض کنید \bar{y}_{hr} میانگین نمونه‌ی مشاهده‌شده از طبقه‌ی h ام و $\hat{\sigma}_{hr}^2 = \frac{1}{n_{hr} - 1} \sum_{i \in s_{hr}} (y_i - \bar{y}_{hr})^2$ واریانس نمونه‌ی مشاهده‌شده باشد. برآوردگر مبتنی بر جانهی به‌وسیله‌ی $\bar{Y}_{strat}^* = \sum_{h=1}^H N_h \bar{y}_h^* / N$ 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}_{strat}^* به‌وسیله‌ی $\bar{Y}_{strat,i}^*$ برای $i = 1, \dots, m$ نمایش داده شود. برآوردگر ترکیب‌شده به‌وسیله‌ی $\bar{\bar{Y}}_{strat}^* = \sum_{i=1}^m \bar{Y}_{strat,i}^* / m$ مشخص می‌شود.

۱-۱-۵- ترکیبات طبقات مجزا

از بخش ۱-۳ نتیجه شد که $k_h = 1/(1-f_h)$ ، که در آن $f_h = (n_h - n_{hr})/n_h$ نرخ بی‌پاسخی در طبقه‌ی h ام است. با استفاده از رابطه‌ی (۸)، فرمول ترکیبی برآورد واریانس $\bar{\bar{Y}}_{strat}^*$ به‌صورت

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

است. در این جا، $\bar{V}^* = \sum_{h=1}^H \bar{V}_h^*$ و \bar{V}_h^* میانگین مقدار از برآورد واریانس مبتنی بر
 جانمایی $\hat{V}_h^* = v_h^* \hat{\sigma}_h^{*2} \left(\frac{1}{n_h} - \frac{1}{N_h} \right)$ 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 $\text{var}(v_h \bar{Y}_h^* | y_{\text{obs}})$ و $\text{var}(\bar{Y}_{\text{strat}}^* | y_{\text{obs}}) = \sum_{h=1}^H \text{var}(v_h \bar{Y}_h^* | y_{\text{obs}})$ تعیین شوند. پس از آن

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

اکنون،

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

و

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

بنابراین می‌توانیم مقدار k را به صورت زیر تعیین کنیم.

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

اگر اندازه‌های طبقه‌ی N_h بزرگ باشد، $\hat{V}(v_h \bar{Y}_h) = v_h \hat{\sigma}_{hr}^2 / n_h$ در نظر گرفته می‌شود. همچنین اگر فرض کنیم $b_h = f_h \hat{V}(v_h \bar{Y}_h) / \sum_{k=1}^H f_k \hat{V}(v_k \bar{Y}_k)$ باشد آن‌گاه

$$(۱۳) \quad k = \frac{\sum_{h=1}^H \hat{V}(v_h \bar{Y}_h) f_h \frac{1}{1-f_h}}{\sum_{h=1}^H \hat{V}(v_h \bar{Y}_h) f_h} = \sum_{h=1}^H b_h \cdot \frac{1}{1-f_h}$$

چون $\sum_{h=1}^H b_h = 1$ 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 به صورت $k = 1/(1-f)$ در نظر گرفته می‌شود. در غیر این صورت، اگر یا نرخ بی‌پاسخی زیاد باشد و یا واریانس برآوردشده‌ی $v_h \bar{Y}_h$ زیاد باشد، نرخ پاسخ طبقه $1-f_h$ ، 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 جانهی شده‌ی \bar{Y}_h^* مستقل هستند، که این مطلب دلالت بر این دارد که $E(\bar{Y}_{strat}^* | y_{obs}) = \sum_{h=1}^H N_h \bar{y}_{hr} / N = \bar{y}_{strat,r}$ و $\text{var}(\bar{Y}_{strat}^* | y_{obs}) \approx \sum_{h=1}^H v_h f_h \hat{\sigma}_{hr}^2 / n_h$. درست مانند بخش ۱-۳، رابطه‌ی (۳) برقرار است و با استفاده از رابطه‌ی (۵)، رابطه‌ی زیر به‌دست می‌آید.

$$E(k) \approx \frac{\text{var}(\bar{Y}_{strat,r}) - \text{var}(\bar{Y}_{strat})}{E\left(\sum_h v_h f_h \hat{\sigma}_{hr}^2 / n_h\right)}$$

$$= \frac{\sum_{h=1}^H v_h \hat{\sigma}_h^2 \left(E\left(\frac{1}{n_{hr}}\right) - \frac{1}{N_h} \right) - \sum_{h=1}^H v_h \hat{\sigma}_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right)}{\sum_{h=1}^H v_h \hat{\sigma}_h^2 \cdot E\left\{ \frac{f_h}{n_h} E\left(\hat{\sigma}_{hr}^2 | n_{hr}\right) \right\}}$$

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

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

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

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 بخش ۲-۱-۵ 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 برآورد شده‌ی $v_h \bar{Y}_{hr}$ زیاد باشد، وزن زیادی اخذ می‌کند. برآورد مجموع براساس نمونه‌ی پاسخ به‌وسیله‌ی $\bar{Y}_{strat,r} = \sum_h v_h \bar{Y}_{hr}$ مشخص می‌شود. با در نظر 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 k ای که در رابطه‌ی (۱۳) مشخص شده است، فرمول $E(k)$ به‌طور تقریبی برقرار است.

۲-۵- رگرسیون لوژیستیک با متغیرهای تبیینی دودویی. برآورد لگاریتم (نسبت بخت‌ها)

متغیرهای Y_1, \dots, Y_n متغیرهای مستقلی هستند که مقادیر صفر و ۱ را اخذ می‌کنند و متغیر تبیینی x با مقادیر ثابت معلوم x_1, \dots, x_n که مقادیر صفر و ۱ را اخذ می‌کند، در نظر گرفته شده است. احتمال رده‌ها با $\pi_1 = P(Y_i = 1 | x_i = 1)$ و $\pi_0 = P(Y_i = 1 | x_i = 0)$ مشخص شده است. یک مدل گم‌شدن تصادفی برای متغیرهای پاسخ R_1, \dots, 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 به صورت $\log \{P(Y = 1 | x) / P(Y = 0 | x)\} = \alpha + \beta x$ بازپارامتری کرد که در آن

$$\beta = \log \frac{\pi_1 / (1 - \pi_1)}{\pi_0 / (1 - \pi_0)} \text{ و } \alpha = \log \{ \pi_0 / (1 - \pi_0) \}$$

برآورد β است. فرض کنید $s = (1, \dots, n)$ 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_1 و s_0 با n_1 و n_0 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 معرف تعداد مشاهدات با $x = i$ و $y = j$ است، به صورت زیر نمایش داد (جدول ۱ را ببینید).

پس برآوردهای ماکسیمم درست‌نمایی برای π_{1r} و π_{0r} به ترتیب برابر $\hat{\pi}_{1r} = n_{11r} / n_{1r}$ و $\hat{\pi}_{0r} = n_{01r} / n_{0r}$ است و در نتیجه برآورد درست‌نمایی β به صورت $\hat{\beta}_r = \log \frac{\hat{\pi}_{1r} / (1 - \hat{\pi}_{1r})}{\hat{\pi}_{0r} / (1 - \hat{\pi}_{0r})} = \log (n_{11r} n_{00r} / n_{10r} n_{01r})$ است. به همین ترتیب، برآوردگر

مبتنی بر کل نمونه به وسیله‌ی $\hat{\beta} = \log \frac{\hat{\pi}_1 / (1 - \hat{\pi}_1)}{\hat{\pi}_0 / (1 - \hat{\pi}_0)} = \log (n_{11} n_{00} / n_{10} n_{01})$ با

نمادهایی آشکارا متناظر نمایش داده می‌شود. می‌توان این برآورد را به صورت

جدول ۱- داده‌های مشاهده‌شده و مجموع بی‌پاسخی برای دو رده

بی‌پاسخی	مجموع	$y = 1$	$y = 0$	y / x
$n_0 - n_{0r}$	n_{0r}	n_{01r}	n_{00r}	$x = 0$
$n_1 - n_{1r}$	n_{1r}	n_{11r}	n_{10r}	$x = 1$

بیان کرد. همچنین $\hat{\beta}_0$ و $\hat{\beta}_1$ که از طبقاتی با نمونه‌ی مجزای s_0 و s_1 به دست آمده‌اند، مستقل هستند. برای اندازه‌های نمونه‌ی n_0 و n_1 بزرگ، $\hat{\beta}$ به طور تقریبی دارای توزیع $N(\beta, \sigma_{\hat{\beta}}^2)$ است که در آن $\sigma_{\hat{\beta}}^2 = \{n_1 \pi_1 (1 - \pi_1)\}^{-1} + \{n_0 \pi_0 (1 - \pi_0)\}^{-1}$ است. همچنین به طریقی تقریبی $\text{var}(\hat{\beta}_1) = 1 / \{n_1 \pi_1 (1 - \pi_1)\}$ و $\text{var}(\hat{\beta}_0) = 1 / \{n_0 \pi_0 (1 - \pi_0)\}$ و برآورد $\text{var}(\hat{\beta})$ از طریق رابطه‌ی زیر مشخص می‌شود:

$$\hat{V}(\hat{\beta}) = \frac{1}{n_1 \hat{\pi}_1 (1 - \hat{\pi}_1)} + \frac{1}{n_0 \hat{\pi}_0 (1 - \hat{\pi}_0)} = \left(\frac{1}{n_{11}} + \frac{1}{n_{10}} \right) + \left(\frac{1}{n_{01}} + \frac{1}{n_{00}} \right)$$

به گونه‌ای که $\hat{V}(\hat{\beta}) = \hat{V}_1 + \hat{V}_0$ ، که در آن $\hat{V}_1 = \left(\frac{1}{n_{11}} + \frac{1}{n_{10}} \right)$ و $\hat{V}_0 = \left(\frac{1}{n_{01}} + \frac{1}{n_{00}} \right)$ به ترتیب برآورد واریانس $\hat{\beta}_1$ و $\hat{\beta}_0$ هستند.

اکنون روش جانپی زیر د ر نظر گرفته خواهد شد؛ برای هر مقدار گم‌شده در $s_{1r} - s_{0r}$ ، مقدار جانپی شده‌ی y^* به تصادف از توزیع برآوردشده‌ی Y به شرط $x = 1$ به صورت زیر استخراج شده است:

$$\text{با احتمال } \hat{\pi}_{1r} = n_{11r} / n_{1r}, y^* = 1 \text{ و با احتمال } 1 - \hat{\pi}_{1r}, y^* = 0.$$

همین روش جانپی برای $s_{0r} - s_{0r}$ استفاده می‌شود به گونه‌ای که y^* به تصادف از توزیع برآوردشده‌ی Y به شرط $x = 0$ استخراج می‌شود. این روش همان روش جانپی بی‌درنگ طبقه‌بندی شده است به گونه‌ای که مقادیر جانپی شده به تصادف و با جایگذاری از $y_{1, \text{obs}} = (y_i : i \in s_{1r})$ و $y_{0, \text{obs}} = (y_i : i \in s_{0r})$ استخراج می‌شوند.

مقادیر برآورد شده در $s - s_r$ را می‌توان به همان شکل داده‌های اصلی نشان داد که در آن n_{ij}^* معرف تعداد مقادیر جانہی شده با $x = i$ و $y = j$ است (جدول ۲ را ببینید). برآورد π_1 براساس جانہی به‌وسیله‌ی $n_{1r}^* = (n_{1r} + n_{11}^*) / n_1$ مشخص می‌شود به‌گونه‌ای که برآورد مبتنی بر جانہی

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

است. به همین ترتیب، برآوردهای β_0 و β براساس جانہی به‌صورت $\hat{\beta}_0^* = \log \left\{ (n_{0r} + n_{01}^*) / (n_0 - n_{0r} - n_{01}^*) \right\}$ و $\hat{\beta}_i^* = \hat{\beta}_1^* - \hat{\beta}_0^*$ تعیین می‌شوند. m جانہی تکرار شده به m برآورد $\hat{\beta}_{1,i}^*$ ، $\hat{\beta}_{0,i}^*$ ، $\hat{\beta}_{\cdot,i}^*$ برای $i = 1, \dots, m$ منجر می‌شود. برآورد ترکیب‌شده به وسیله‌ی

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

بیان می‌شود. برآورد واریانس جانہی‌شده‌ی \hat{V}^* برای $\hat{\beta}$ به‌صورت زیر مشخص می‌شود؛

$$(۱۶) \quad \hat{V}^* = \frac{1}{n_{1r} + n_{11}^*} + \frac{1}{n_{0r} + n_{01}^*} + \frac{1}{n_{01r} + n_{01}^*} + \frac{1}{n_{00r} + n_{00}^*}$$

می‌بینیم که $E(\hat{V}^* | y_{\text{obs}}) \approx \frac{1}{n_1 \hat{\pi}_{1r} (1 - \hat{\pi}_{1r})} + \frac{1}{n_0 \hat{\pi}_{0r} (1 - \hat{\pi}_{0r})}$ و رابطه‌ی (۳) برقرار است. همچنین توجه می‌کنیم که رابطه‌ی (۹) نیز به‌طور 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 $E(\hat{\beta}_1^* | y_{1,\text{obs}}) \approx \hat{\beta}_{1r}$ و $E(\hat{\beta}_0^* | y_{0,\text{obs}}) \approx \hat{\beta}_{0r}$ با $\text{var}(\hat{\beta}_1^* | y_{1,\text{obs}}) \approx f_1(1 - f_1) \hat{V}(\hat{\beta}_{1r})$ استفاده از رابطه‌ی (۱۰)، به‌طور تقریبی داریم

جدول ۲- مجموع‌های جانمایی شده برای دو رده

مجموع	$y = ۱$	$y = ۰$	y / x
$n_۰ - n_{۰r}$	$n_{۰۱}^*$	$n_{۰۰}^*$	$x = ۰$
$n_۱ - n_{۱r}$	$n_{۱۱}^*$	$n_{۱۰}^*$	$x = ۱$

$$E(k_۱) = \frac{\text{var}(\hat{\beta}_{1r}) - \text{var}(\hat{\beta}_1)}{E\{f_1(1-f_1)\hat{V}(\hat{\beta}_{1r})\}} = \frac{E \text{var}(\hat{\beta}_{1r}|n_{1r}) - \text{var}(\hat{\beta}_1)}{E\{f_1(1-f_1)E[\hat{V}(\hat{\beta}_{1r})|n_{1r}]\}}$$

$$\approx \frac{\frac{1}{\pi_1(1-\pi_1)} \left(E\left(\frac{1}{n_{1r}}\right) - \frac{1}{n_1} \right)}{E f_1(1-f_1) \frac{1}{n_{1r}\pi_1(1-\pi_1)}} \approx \frac{(1-p_{1r})/p_{1r}}{1-p_{1r}} = \frac{1}{p_{1r}}$$

که با قرار دادن $k_1 = 1/(1-f_1)$ به طور تقریبی در رابطه صدق می‌کند. درست به همین ترتیب پی می‌بریم که $k_0 = 1/(1-f_0)$ است که در آن $f_0 = (n_0 - n_{0r})/n_0$ نسبت بی‌پاسخی در طبقه s_0 است. مؤلفه‌ی بین جانمایی برای $\hat{\beta}_1^*$ به وسیله‌ی $B_1^* = \frac{1}{m-1} \sum_{i=1}^m (\hat{\beta}_{1,i}^* - \bar{\beta}_1^*)^2$ مشخص می‌شود و همچنان B_0^* ، مؤلفه‌ی بین جانمایی برای $\hat{\beta}_0^*$ است. بنا بر این یک واریانس برآوردشده‌ی برآورد ترکیبی $\hat{\beta}$ برای β بر اساس جانمایی، با استفاده از رابطه‌ی (۸) به صورت زیر مشخص می‌شود

$$W_{\text{sep}} = \bar{V}^* + \sum_{x=0}^1 \left(\frac{1}{1-f_x} + \frac{1}{m} \right) B_x^*$$

که در آن \bar{V}^* میانگین m تکرار برآورد واریانس جانمایی شده‌ی \hat{V}^* است که از طریق رابطه‌ی (۱۶) مشخص شده است.

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

چون $\text{var}(\hat{\beta}_1^* | y_{1,\text{obs}}) = f_1(1-f_1)\hat{V}(\hat{\beta}_1)$ و $\text{var}(\hat{\beta}_0^* | y_{0,\text{obs}}) = f_0(1-f_0)\hat{V}(\hat{\beta}_0)$ است، از رابطه‌ی (۱۲) داریم؛

$$(۱۷) \quad k = \frac{1}{1-f_1} \cdot \frac{f_1(1-f_1)\hat{V}(\hat{\beta}_1)}{\sum_{x=0}^1 f_x(1-f_x)\hat{V}(\hat{\beta}_{xr})} + \frac{1}{1-f_0} \cdot \frac{f_0(1-f_0)\hat{V}(\hat{\beta}_0)}{\sum_{x=0}^1 f_x(1-f_x)\hat{V}(\hat{\beta}_{xr})}$$

نمونه‌ی $\hat{\beta}_1$ و $\hat{\beta}_0$ را به ترتیب از طریق $\hat{V}(\hat{\beta}_1) = (1-f_1)\hat{V}(\hat{\beta}_1)$ و $\hat{V}(\hat{\beta}_0) = (1-f_0)\hat{V}(\hat{\beta}_0)$ برآورد کرد. بنا بر این

$$k = \frac{1}{1-f_1} \cdot \frac{f_1\hat{V}(\hat{\beta}_1)}{\sum_{x=0}^1 f_x\hat{V}(\hat{\beta}_x)} + \frac{1}{1-f_0} \cdot \frac{f_0\hat{V}(\hat{\beta}_0)}{\sum_{x=0}^1 f_x\hat{V}(\hat{\beta}_x)} = \frac{1}{1-f_1} b_1 + \frac{1}{1-f_0} (1-b_1)$$

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 $k = 1/(1-f)$ در غیر این صورت، نرخ پاسخ طبقه‌ی $1-f_x$ زمانی که یا نرخ بی‌پاسخی زیاد باشد و یا واریانس برآورد شده‌ی $\hat{\beta}_x$ زیاد باشد، وزن زیادی خواهد داشت. یا این که از رابطه‌ی (۱۷)

$$\frac{1}{k} = \frac{\sum_{x=0}^1 (1-f_x) f_x \hat{V}(\hat{\beta}_{xr})}{\sum_{x=0}^1 f_x \hat{V}(\hat{\beta}_{xr})} = \sum_{x=0}^1 a_x (1-f_x)$$

که در آن وزن‌ها به صورت $a_x = f_x \hat{V}(\hat{\beta}_{xr}) / \{f_1 \hat{V}(\hat{\beta}_1) + f_0 \hat{V}(\hat{\beta}_0)\}$ می‌باشد. بنا بر این می‌توان $1/k$ را به‌عنوان میانگین موزون از نرخ‌های پاسخ بیان کرد.

اگر هدف برآورد π_1 و π_0 باشد، مقدار $k = 1/(1-f_1)$ و $k = 1/(1-f_0)$ به ترتیب برای π_1 و π_0 به دست می‌آید.

۳-۵- رگرسیون لوژیستیک با متغیرهای تبیینی رسته‌ای گسسته. برآورد لگاریتم (نسبت بخت‌ها)

اگر متغیر تبیینی x به صورت رسته‌ای است که مثلاً H رده را تعریف می‌کند، می‌توان نتایج را به صورت زیر تعمیم داد.

فرض کنید برای $h = 0, \dots, H-1$ ، $\pi_h = P(Y = 1 | x = h)$ است. رگرسیون لوژیستیک که برای این داده‌های رده‌ای تعریف می‌شود، از طریق معرفی $H-1$ متغیر تبیینی دودویی x_1, \dots, x_{H-1} صورت می‌گیرد که در آن اگر مشاهدات به رده‌ی h تعلق داشته باشد، برای $h = 0, \dots, H-1$ ، $x_h = 1$ و در غیر این صورت صفر خواهد بود. پس اگر $x_1 = x_2 = \dots = x_{H-1} = 0$ باشد، مشاهده به رده‌ی صفر تعلق دارد. مدل لوجیت با $x = (x_1, x_2, \dots, x_{H-1})$ به صورت

$$\log \left\{ \frac{P(Y = 1 | x)}{P(Y = 0 | x)} \right\} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + x_{H-1} \beta_{H-1}$$

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. برآورد β_h از طریق جانه‌ی چندگانه دقیقاً مشابه متغیر دودویی x با جایگزینی رده‌ی h با رده‌ی ۱ انجام می‌شود.

۴-۵- رگرسیون لوژیستیک 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، $y = (y_1, \dots, y_n)$ ، و مقادیر گم‌شده برای متغیر تبیینی x وجود دارد. Y_1, \dots, 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

$y = 1$	$y = 0$	y / x
n_{01r}	n_{00r}	$x = 0$
n_{11r}	n_{10r}	$x = 1$
n_{1r}^0	n_{0r}^0	مجموع
$n_1^0 - n_{1r}^0$	$n_0^0 - n_{0r}^0$	بی‌پاسخی

مقادیر ثابت x_1, x_2, \dots, x_n که برخی از آن‌ها گم‌شده است نیز وجود دارد. متغیرهای پاسخ نشان‌دهنده‌ی گم‌شدگی x_i ها است اما اکنون با مدل گم‌شدن تصادفی، $P(R_i = 1 | y_i = 1) = q_{1r}$ و $P(R_i = 1 | y_i = 0) = q_{0r}$ روبه‌رو هستیم.

در غیر این صورت، مدل مشابه بخش ۲-۵ با احتمال‌های رده‌ای $\pi_1 = P(Y_i = 1 | x_i = 1)$ و $\pi_0 = P(Y_i = 1 | x_i = 0)$ و مدل لوجیت $\log\{P(Y = 1 | x) / P(Y = 0 | x)\} = \alpha + \beta x$ با $\beta = \log \frac{\pi_1 / (1 - \pi_1)}{\pi_0 / (1 - \pi_0)}$ است. هنوز هدف برآورد β است.

اکنون فرض کنید $s^1 = \{i \in S : y_i = 1\}$ و $s^0 = \{i \in S : y_i = 0\}$ با اندازه‌های n_1^0 و n_0^0 باشد. نمونه‌های پاسخ در طبقات، $s_r^1 = \{i \in s^1 : R_i = 1\}$ و $s_r^0 = \{i \in s^0 : R_i = 1\}$ با نمونه‌ی پاسخ کل $s_r = \{i \in S : R_i = 1\} = s_r^1 \cup s_r^0$ است. حال، داده‌ها می‌توانند همانند قبل معرفی شوند با این تفاوت که مجموع‌های بی‌پاسخی به هر طبقه‌ی y تعلق دارد (جدول ۳ را ببینید).

برآورد ماکسیمم درست‌نمایی $\hat{\pi}_{1r}, \hat{\pi}_{0r}, \hat{\beta}_r$ براساس s_r مشابه قبل است که برای برآورد کل نمونه‌ی $\hat{\beta}$ بوده است. روش جانہی، روش بی‌درنگ طبقه‌بندی شده برای طبقه‌های y است. برای هر مقدار گم‌شده‌ی x در $s_r^1 - s_r^0$ ، مقدار جانہی شده‌ی x^* به تصادف از $(x_i : i \in s_r^1)$ قرعه‌کشی می‌شود. به همین ترتیب، مقادیر جانہی شده در $s^0 - s_r^0$ نیز به تصادف از $(x_i : i \in s_r^0)$ قرعه‌کشی می‌شود. مقادیر

جدول ۴- مجموع‌های جانپ‌های شده برای طبقه‌های y

$y = ۱$	$y = ۰$	y / x
$n_{۰۱}^*$	$n_{۰۰}^*$	$x = ۰$
$n_{۱۱}^*$	$n_{۱۰}^*$	$x = ۱$
$n_{۱}^{\circ} - n_{۱r}^{\circ}$	$n_{۰}^{\circ} - n_{or}^{\circ}$	مجموع

جانپ‌های شده در $s - s_r$ می‌تواند به همان صورت داده‌های اصلی بیان شود به طوری که اکنون n_{ij}^* معرف تعداد مقادیر جانپ‌های شده 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

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

۹

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

در این حالت $f^{-1} = (n_{11}^{\circ} - n_{1r}^{\circ}) / n_{11}^{\circ}$ نرخ بی‌پاسخی در طبقه‌ی s^{-1} و $f^{\circ} = (n_{00}^{\circ} - n_{or}^{\circ}) / n_{00}^{\circ}$ نرخ بی‌پاسخی در طبقه‌ی s° است. لازم به ذکر است که $\hat{q}_{or} = n_{or}^{\circ} / n_{00}^{\circ}$ و $\hat{q}_{1r} = n_{1r}^{\circ} / n_{11}^{\circ} = 1 - f^{-1}$ زیر در می‌آید.

$$(۱۸) \quad E\left\{f^{-1}(1-f^{-1})\left(\frac{1}{n_{11r}} + \frac{1}{n_{01r}}\right) + f^{\circ}(1-f^{\circ})\left(\frac{1}{n_{10r}} + \frac{1}{n_{00r}}\right)\right\}$$

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

$$(۱۹) \quad \text{var}(\hat{\beta}_r) - \text{var}(\hat{\beta}) = \frac{1}{n_1 \pi_1 (1 - \pi_1)} \cdot \frac{1 - p_{1r}}{p_{1r}} + \frac{1}{n_0 \pi_0 (1 - \pi_0)} \cdot \frac{1 - p_{0r}}{p_{0r}}$$

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

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

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

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

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

که در آن $a_1 = f'w_1 / (f'w_0 + f'w_1)$ و $a_0 = f'w_0 / (f'w_0 + f'w_1)$ 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 جانمایی برای همه‌ی برآوردهای علمی استفاده کرد؟

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

۱- S یک نمونه از یک جامعه‌ی متناهی و $y = (y_i : i \in S)$ 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

$$y_{\text{obs}} = \{(y_i : i \in s_r), s_r, s\}$$

۳- یک مطالعه‌ی مشاهده‌ای که در آن $s = (1, \dots, n)$ باشد و $y = (y_1, \dots, y_n)$ مدل‌بندی شده باشد. در این حالت متغیرهای تصادفی مشاهده‌شده به صورت

$$y_{\text{obs}} = \{(y_i : i \in s_r), s_r\}$$

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

$$(20) \quad k = \frac{1}{1-f}$$

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

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

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

$$E(\hat{\theta}^* | y, s) = \hat{\theta} \quad (21)$$

یعنی برآوردهای جانہی شده باید پارامتر مشابهی مانند $\hat{\theta}$ را برآورد کند. می‌توان گفت که مقدار مورد انتظار برآوردهای جانہی شده به شرط کل داده‌های نمونه‌گیری شده، باید معادل برآورد کل نمونه باشد. در حالت ۱، زمانی که s مشخص باشد، y اضافی است و رابطه‌ی (۲۱) بیان می‌کند که $E(\hat{\theta}^* | s) = \hat{\theta}$ 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 برآوردهایی را در نظر گرفته‌ایم که خطی در $(y_i : i \in s)$ هستند. نتایج زیر که در پیوست ۵ اثبات شده‌اند، برآوردهای خطی را مشخص می‌کنند که با استفاده از روش جانہی بی‌درنگ در رابطه‌ی (۲۱) صدق می‌کند و نشان می‌دهند که برای چنین برآوردهایی، $k = 1/(1-f)$ است.

لم. فرض کنید $\hat{\theta} = \sum_{i \in s} a_i(s) y_i$ است. بنا بر این $E(\hat{\theta}^* | y, s) = \hat{\theta}$ اگر و تنها اگر برای همه‌ی $i \in s$ ، $a_i(s) = a(s)$ ، یعنی $\hat{\theta} = a(s) \sum_{i \in s} y_i = n a(s) \bar{y}_s$ ، $a_i(s) = a_i$ و $a_i(s) = a_i$ تذکر. در حالت ۳، s هیچ اطلاعی نمی‌دهد و

قضیه. فرض کنید $\hat{\theta} = \sum_{i \in s} a_i(s) y_i$ و $E(\hat{\theta}^* | y, s) = \hat{\theta}$ باشد. همچنین فرض کنید که رابطه‌ی (۳) برقرار است. بنا بر این $E(k) = \frac{E(1/\hat{p}_r) - 1}{1 - p_r - \frac{1}{n} [E(1/\hat{p}_r) - 1]} \approx \frac{1}{p_r}$ و رابطه‌ی $k = 1/(1-f)$ می‌تواند مورد استفاده قرار گیرد.

اکنون به برخی از حالت‌های خاص می‌پردازیم؛

۱- با در نظر گرفتن $a(s) = 1/n$ ، مشابه بخش ۱-۳، می‌بینیم که رابطه‌ی (۲۱) برقرار است.

۲- ضریب رگرسیون برای رگرسیون‌ی که از مبدأ می‌گذرد، به صورت $\hat{\beta} = \sum_{i=1}^n y_i / \sum_{i=1}^n x_i$ است. در این حالت، با در نظر گرفتن $a = 1 / \sum_{i=1}^n x_i$ رابطه‌ی (۲۱) برقرار است و بنا بر این $k = 1 / (1-f)$ است.

۳- برآورد ضریب رگرسیون در رگرسیون خطی معمول، حالتی است که در رابطه‌ی (۲۱) صدق نمی‌کند و در آن $\hat{\beta} = \sum (x_i - \bar{x}) y_i / \sum (x_i - \bar{x})^2$ است. در این حالت، $a_i = (x_i - \bar{x}) / \sum_{j=1}^n (x_j - \bar{x})^2$ است که مستقل از i نیست. می‌توان نشان داد که به طور تقریبی $E(\hat{\beta}^* | y) \approx p_r \hat{\beta}$ و به طور دقیق به صورت $\frac{np_r - 1}{n - 1} \hat{\beta}$ است. بنا بر این برای مسائل رگرسیون عادی، روش جانهای بی‌درنگ نمی‌تواند مؤثر واقع شود. لازم به ذکر است که با استفاده از مطالب بخش ۳-۳ می‌توان با استفاده از جانهای مانده‌ها از $k = 1 / (1-f)$ استفاده کرد.

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

۷- بحث و نتیجه‌گیری

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

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

از مواردی که در این مقاله ارائه شده است می‌توان نتیجه گرفت که فرمول روش جانہی غیر بیزی نوعاً به معیاری از سهم اطلاع گم‌شده در نمونہی پاسخ در مقایسه با کل نمونہ بستگی دارد. در ساده‌ترین حالت در بخش ۱-۳، اطلاعات گم‌شده به وسیلہی $(1-f)/1$ که عکس نرخ پاسخ است، اندازه‌گیری شده است. هر چه این عامل بیش‌تر باشد، وزن مؤلفہی بین جانہی نیز بیش‌تر است. در مدل نسبت بخش ۲-۳ با روش جانہی رگرسیون بی‌درنگ مانده‌ها، میزان اطلاعات گم‌شده عکس نسبت x -کل در نمونہی پاسخ در مقایسه با کل نمونہ است. نشان داده شده است که در رگرسیون خطی ساده با عبارت واریانسی که مستقل از متغیر تبیینی است، مجدداً اطلاع گم‌شده به وسیلہی $(1-f)/1$ 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 پایہ‌ای (۵) می‌باشد.

مطالعه‌ی عملکرد بازہ‌های 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 جانہی مانده‌ها و $k = 1/(1-f)$ ، بازہ‌ی اطمینان به صورت $\bar{\beta}^* \pm z_{\alpha/2} \sqrt{W}$ (که در آن $z_{\alpha/2}$ ، نقطہ‌ی بالایی توزیع نرمال استاندارد می‌باشد) است که به طور تقریبی به سطح اسمی $(1-\alpha)$ 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- جانمایی چندگانه برای مدل نسبت در بخش ۲-۳

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

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

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

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

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

در مرحله ی بعد، به $\text{var}(\hat{\beta}^* | y_{\text{obs}})$ و $E(\hat{\beta}^* | y_{\text{obs}})$ نگاه می کنیم. می بینیم که $E(\hat{\beta}_{nr} | y_{\text{obs}}) = \hat{\beta}_r + (\bar{e} / X_{nr}) \sum_{s-s_r} \sqrt{x_i}$ و $\hat{\beta}^* = (\hat{\beta}_r X_r + \hat{\beta}_{nr} X_{nr}) / X$ و $\text{var}(\hat{\beta}_{nr} | y_{\text{obs}}) = s_e^2 / X_{nr}$ می باشد. این اطلاعات بیان می کند که $\text{var}(\hat{\beta}^* | y_{\text{obs}}) = (X_{nr} / X^2) s_e^2$ و $E(\hat{\beta}^* | y_{\text{obs}}) = \hat{\beta}_r + (\bar{e} / X) \sum_{s-s_r} \sqrt{x_i}$ است. و این نتیجه می شود که

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

اکنون، $\text{cov}(\hat{\beta}_r, \bar{e}) = 0$ است. با استفاده از نابرابری کوشی شوارتس، $\left(\sum a_i b_i\right)^2 \leq \sum a_i^2 \sum b_i^2$ با $a_i = \sqrt{x_i}$ و $b_i = 1$ ، درمی یابیم که $\left(\sum_{i=1}^n \sqrt{x_i}\right)^2 \leq nX$ است. نتیجه این می شود که

$$\text{var}(\bar{e}) = \left(\sigma^2 / n_r\right) \left(1 - \left(\sum_{s-s_r} \sqrt{x_i}\right)^2 / n_r X_r\right) = (1 - d_1) \sigma^2 / n_r, \quad 0 \leq d_1 \leq 1$$

$$\left(\sum_{s-s_r} \sqrt{x_i}\right)^2 / X^2 = d_1 n_{nr} X_{nr} / X^2, \quad 0 \leq d_1 \leq 1$$

بنا بر این،

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

سپس درمی یابیم که $E(s_e^2) = \sigma^2 \left(1 - \frac{1}{n_r}\right) - \text{var}(\bar{e}) = \sigma^2 (n_r + d_1 - 2) / n_r$ است که از

این اطلاع خواهیم داشت

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

۲- جانهای چندگانه در رگرسیون خطی ساده در بخش ۳-۳. خلاصه‌ای از

اثبات:

$$E(\hat{\sigma}_*^{\top}) = \sigma^{\top} E\left(\frac{n_r - 2}{n_r} \cdot \frac{n - 2f}{n - 2}\right), f = (n - n_r)/n$$

در این جا $f = (n - n_r)/n$

$$SS_e^r = \sum_{s_r} (y_i - \hat{\alpha}^* - \hat{\beta}^* x_i)^{\top} \quad \text{و} \quad \hat{\sigma}_*^{\top} = \frac{1}{n - 2} (SS_e^r + SS_e^{nr})$$

$$SS_e^{nr} = \sum_{s-s_r} (y_i^* - \hat{\alpha}^* - \hat{\beta}^* x_i)^{\top}$$

می‌توان دو مجموع مربعات مانده‌ها را به صورت زیر بیان کرد

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

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

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

$$E(SS_e^r | y_{\text{obs}}) = n_r s_e^{\top} + n_r \text{ var}(\hat{\alpha}^* | y_{\text{obs}}) + \left(\sum_{s-s_r} x_i^{\top} \right) \text{ var}(\hat{\beta}^* | y_{\text{obs}})$$

$$+ 2n_r \bar{x}_r \text{ cov}(\hat{\alpha}^*, \hat{\beta}^* | y_{\text{obs}})$$

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

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

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

پس از بعضی محاسبات جبری، درمی یابیم که

$$E(SS_e^{nr} | y_{\text{obs}}) = \sum_{s-s_r} \text{var}(Y_i^* - \bar{Y}_{nr}^* | y_{\text{obs}}) + \sum_{s-s_r} (x_i - \bar{x}_{nr})^{\vee} \text{var}(\hat{\beta}_{nr}^* | y_{\text{obs}})$$

$$- 2 \sum_{s-s_r} (x_i - \bar{x}_{nr}) \text{cov}(Y_i^* - \bar{Y}_{nr}^*, \hat{\beta}_{nr}^* | y_{\text{obs}})$$

$$+ (n-n_r) \text{var}(\hat{\alpha}_{nr}^* - \hat{\alpha}^* | y_{\text{obs}}) + c_r SS_x \text{var}(\hat{\beta}_{nr}^* - \hat{\beta}^* | y_{\text{obs}})$$

$$+ 2(n-n_r) \bar{x}_{nr} \text{cov}(\hat{\alpha}_{nr}^* - \hat{\alpha}^*, \hat{\beta}_{nr}^* - \hat{\beta}^* | y_{\text{obs}})$$

$$= (n-n_r-1) s_e^{\vee} + s_e^{\vee} - 2s_e^{\vee} + (n-n_r) \left\{ (1-f)^{\vee} s_e^{\vee} \frac{1}{n-n_r} + \bar{x}_{nr}^{\vee} s_e^{\vee} \frac{1}{SS_{x,nr}} \right\}$$

$$+ c_r SS_x s_e^{\vee} \left(\frac{1}{SS_{x,nr}} + \frac{c_r}{SS_x} - 2 \frac{1}{SS_x} \right)$$

$$+ 2(n-n_r) \bar{x}_{nr} \left(f \bar{x}_{nr} s_e^{\vee} \frac{1}{SS_x} - \bar{x}_{nr} s_e^{\vee} \frac{1}{SS_{x,nr}} \right)$$

که در آن $SS_{x,nr} = \sum_{s-s_r} (x_i - \bar{x}_{nr})^{\vee}$ می بینیم که

$$SS_{x,nr} = \sum_{s-s_r} x_i^{\vee} - (n-n_r) \bar{x}_{nr}^{\vee} = c_r SS_x - (n-n_r) \bar{x}_{nr}^{\vee}$$

از مطالب بالا نتیجه می شود که $c_r SS_x / SS_{x,nr} = 1 + (n-n_r) \bar{x}_{nr}^{\vee} / SS_{x,nr}$

$$E(SS_e^{nr} | y_{\text{obs}}) = s_e^{\vee} ((n-n_r-1) + (1-f)^{\vee} + c_r^{\vee} - 2c_r^{\vee})$$

$$(23) \quad + 2f \bar{x}_{nr}^{\vee} (n-n_r) / SS_x$$

از روابط (22) و (23) درمی یابیم که

$$(n-2)E(\hat{\sigma}_*^2 | y_{\text{obs}}) = s_e^2 \left(n-f - c_r + 2f \frac{1}{SS_x} \bar{x}_{nr} n \bar{x} \right) = s_e^2 (n-f - c_r)$$

$$\text{چون } E(c_r | n_r) = \frac{1}{SS_x} \sum_{i=1}^n E(1-R_i | n_r) x_i^2 = (1-n_r/n) = f \text{ داریم}$$

$$\begin{aligned} (n-2)E(\hat{\sigma}_*^2) E s_e^2 (n-f - c_r) &= E(n-f - c_r) E(s_e^2 | s_r) \\ &= E(n-f - c_r) \frac{n_r - 2}{n_r} \sigma^2 \\ &= \sigma^2 E \left\{ \frac{n_r - 2}{n_r} (n-f - E(c_r | n_r)) \right\} \\ &= \sigma^2 E \left\{ \frac{n_r - 2}{n_r} (n - 2f) \right\} \end{aligned}$$

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

هدف تعیین $E(\hat{\beta}_1^* | y_{1,\text{obs}})$ و $\text{var}(\hat{\beta}_1^* | y_{1,\text{obs}})$ است.

به شرط $n_{11}^*, y_{1,\text{obs}}$ دوجمله‌ای با $(n_1 - n_{1r}, \hat{\pi}_{1r})$ است. بنا بر این،
 $E(n_{11}^* | y_{1,\text{obs}}) = (n_1 - n_{1r}) \hat{\pi}_{1r}$ و $\text{var}(n_{11}^* | y_{1,\text{obs}}) = (n_1 - n_{1r}) \hat{\pi}_{1r} (1 - \hat{\pi}_{1r})$ است.
 به شرط $\hat{\beta}_1^*, y_{1,\text{obs}}$ به صورت $T = \log \left\{ \frac{(a+Z)}{(b-Z)} \right\}$ 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
 $T \approx \log \left\{ \frac{(a+np)}{(b-np)} \right\} + (Z - np) \frac{(a+b)}{(a+np)(b-np)}$
 می‌یابیم و خواهیم داشت.

$$E(T) \approx \log \frac{a + np}{b - np} \tag{۲۴}$$

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

با در نظر گرفتن $a = n_{1r}$ و $b = n_1 - n_{1r}$ به نتیجه‌ی $E(\hat{\beta}_1^* | y_{1, \text{obs}}) \approx \hat{\beta}_{1r}$ و $\text{var}(\hat{\beta}_1^* | y_{1, \text{obs}}) \approx (n_1 / \{n_1 \hat{\pi}_{1r} n_1 (1 - \hat{\pi}_{1r})\})^2 (n_1 - n_{1r}) \hat{\pi}_{1r} (1 - \hat{\pi}_{1r})$ دست می‌یابیم. فرض کنید $f_1 = (n_1 - n_{1r}) / n_1$ نرخ بی‌پاسخی در طبقه‌ی s_1 باشد. بنا بر این

$$\begin{aligned} \text{var}(\hat{\beta}_1^* | y_{1, \text{obs}}) &\approx \frac{f_1 n_1}{n_1^2} \cdot \frac{1}{\hat{\pi}_{1r} (1 - \hat{\pi}_{1r})} = f_1 (1 - f_1) \cdot \frac{1}{n_{1r} \hat{\pi}_{1r} (1 - \hat{\pi}_{1r})} \\ &= f_1 (1 - f_1) \hat{V}(\hat{\beta}_{1r}) \end{aligned}$$

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

به منظور تعیین $\text{var}(\hat{\beta}_* | y, x_{\text{obs}})$ و $E(\hat{\beta}_* | y, x_{\text{obs}})$ ، نیاز است که $\hat{\beta}_*$ به روشی 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 $\text{var}(\hat{\beta}_* | y, x_{\text{obs}}) = \text{var}(\hat{\beta}_1^* | y, x_{1, \text{obs}}) + \text{var}(\hat{\beta}_0^* | y, x_{0, \text{obs}})$ به شرط (y, x_{obs}) ، n_{11}^* دارای توزیع دوجمله‌ای با $(n_{11}^* - n_{1r}^*, p^1)$ است که در آن $p^1 = n_{11r} / n_{1r}^*$ است و n_{10}^* دارای توزیع دوجمله‌ای با $(n_{10}^* - n_{1r}^*, p^1)$ است که در آن $p^1 = n_{10r} / n_{1r}^*$ می‌باشد. پس با استفاده از رابطه‌ی (۲۴)، می‌توان به این نتیجه رسید که $\text{var}(\hat{\beta}_* | y, x_{1, \text{obs}}) = \log \{p^1 / (1 - p^1)\}$ و

است. $\text{var}(\hat{\beta}_*^1 | y, x_{1, \text{obs}}) = (n_1^\circ / \{n_1^\circ p^\circ (1-p^\circ)\})^\top (n_1^\circ - n_{1r}^\circ) p^\circ (1-p^\circ)$
 بنابراین $\text{var}(\hat{\beta}_*^1 | y, x_{1, \text{obs}}) \approx f^1 / \{n_1^\circ p^\circ (1-p^\circ)\} = f^1 (1-f^1) / \{n_{1r}^\circ p^\circ (1-p^\circ)\}$
 به همین ترتیب $E(\hat{\beta}_*^1 | y, x_{1, \text{obs}}) \approx \log\{p^\circ / (1-p^\circ)\}$ است به گونه‌ای که
 $E(\hat{\beta}_*^1 | y, x_{\text{obs}}) \approx \hat{\beta}_r^1$ همچنین

$$\text{var}(\hat{\beta}_*^0 | y, x_{0, \text{obs}}) \approx f^0 / \{n_0^\circ p^\circ (1-p^\circ)\} = f^0 (1-f^0) / \{n_{0r}^\circ p^\circ (1-p^\circ)\}$$

$$\frac{1}{n_{1r}^\circ p^\circ (1-p^\circ)} = \frac{1}{n_{1or}} + \frac{1}{n_{0or}} \quad \text{و} \quad \frac{1}{n_{0r}^\circ p^\circ (1-p^\circ)} = \frac{n_{1r}^\circ}{n_{11r} n_{01r}} = \frac{1}{n_{11r}} + \frac{1}{n_{01r}}$$

می‌دانیم و از این‌ها نتیجه می‌شود که

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

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 به اندازه‌ی n_r است.

۳- $P(R_i = 1 | n_r) = n_r / n$ و $P(R_i = 1, R_j = 1 | n_r) = \frac{n_r}{n} \cdot \frac{n_r - 1}{n - 1}$ (از

رابطه‌ی ۲ به دست می‌آید)

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

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

$$\hat{\sigma}^r = \frac{1}{n-1} \sum_{i \in S} (y_i - \bar{y}_s)^r \quad \text{که در آن } E(\hat{\sigma}_r^r | y, s, n_r) = \hat{\sigma}^r - \epsilon$$

$$\left(\Rightarrow \text{var}(Y_i^* | y, s, n_r) = \frac{n-1}{n} \hat{\sigma}^r \approx \hat{\sigma}^r \right)$$

$$\text{var}(\bar{Y}_r | y, s, n_r) = f \hat{\sigma}^r / n_r = \hat{\sigma}^r \left(\frac{1}{n_r} - \frac{1}{n} \right) - \gamma$$

اثبات لم:

$$\begin{aligned} E(\hat{\theta}^* | y, s) &= E \left\{ E \left(\sum_{i \in S_r} a_i(s) y_i + \sum_{i \in S - S_r} a_i(s) Y_i^* | Y_{\text{obs}} \right) | y, s \right\} \\ &= {}^{(*)} E \left(\sum_{i \in S_r} a_i(s) y_i | y, s \right) + E \left(\sum_{i \in S - S_r} a_i(s) \bar{Y}_r | y, s \right) \end{aligned}$$

عبارت اول:

$$\begin{aligned} E \left(\sum_{i \in S_r} a_i(s) y_i | y, s \right) &= E \left\{ E \left(\sum_{i \in S} a_i(s) y_i R_i | y, s, n_r \right) | y, s \right\} \\ &= {}^{(*)} E \left(\sum_{i \in S} a_i(s) y_i \frac{n_r}{n} | y, s \right) = {}^{(*)} p_r \hat{\theta} \end{aligned}$$

عبارت دوم:

$$\begin{aligned} E \left(\sum_{i \in S - S_r} a_i(s) \bar{Y}_r | y, s \right) &= E \left\{ E \left(\frac{1}{n_r} \sum_{i \in S} \sum_{j \in S} a_i(s) y_j (1 - R_i) R_j | y, s, n_r \right) | y, s \right\} \\ &= {}^{(*)} E \left(\frac{1}{n_r} \sum_{i \in S} \sum_{j \in S, j \neq i} a_i(s) y_j \left(\frac{n_r}{n} - \frac{n_r}{n} \cdot \frac{n_r - 1}{n - 1} \right) | y, s \right) \\ &= {}^{(*)} \frac{1 - p_r}{n - 1} \sum_{i \in S} \sum_{j \in S, j \neq i} a_i(s) y_j = \frac{1 - p_r}{n - 1} (n \bar{a}(s) n \bar{y}_s - \hat{\theta}) \end{aligned}$$

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

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

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

اثبات قضیه:

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

است.

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

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

بنا بر این،

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

سپس،

$$\begin{aligned}
 E \text{ var}(\hat{\theta}^* | Y_{\text{obs}}) &= E \left(\left[a(s) \right]^T (n - n_r) \frac{n_r - 1}{n_r} \hat{\sigma}_r^* \right) \\
 &= E \left\{ (n - n_r) (1 - 1/n_r) \left[a(s) \right]^T E(\hat{\sigma}_r^* | Y, s, n_r) \right\} \\
 &= E \left\{ (n - n_r) (1/n_r - 1) \left[a(s) \right]^T \hat{\sigma}^* \right\} \\
 &=^{(1)} \left[n(1 - p_r) - \{E(1/\hat{p}_r) - 1\} \right] E \left(\left[a(s) \right]^T \hat{\sigma}^* \right)
 \end{aligned}$$

از رابطه‌ی (۵) دریافتیم که

$$\begin{aligned}
 E(k) &= \frac{\{E(1/\hat{p}_r) - 1\} E \left\{ \left[a(s) \right]^T \hat{\sigma}^* \right\}}{\left[(1 - p_r) - \frac{1}{n} \{E(1/\hat{p}_r) - 1\} \right] E \left(\left[a(s) \right]^T \hat{\sigma}^* \right)} \\
 &= \frac{\{E(1/\hat{p}_r) - 1\}}{1 - p_r - \frac{1}{n} \{E(1/\hat{p}_r) - 1\}} \\
 &\approx \frac{(1/p_r) - 1}{1 - p_r} = \frac{1}{p_r}
 \end{aligned}$$

زهرا رضایی قهرودی

دکتری آمار

تهران، خیابان دکتر فاطمی، خیابان باباطاهر، خیابان شهید فکوری، شماره‌ی ۱۴۵، پژوهشکده‌ی آمار.

رایانشانی: z_rezaei@srta.ac.ir