

## فصل ۴

# مطالعات مقطعی: داده‌های دارای پاسخ اسمی

در این فصل به معرفی و بررسی مطالعات مقطعی با پاسخ یک متغیره‌ی اسمی می‌پردازیم. برای این منظور، در بخش اول، مفهومی از مطالعات مقطعی برای پاسخ‌های اسمی، همراه با نمادگذاری در این داده‌ها بیان می‌شود. در بخش دوم، توزیع‌های مورد استفاده برای این نوع داده‌ها معرفی می‌شوند. مدل‌های لوجیت چندجمله‌ای و لوجیت با استفاده از متغیر پنهان برای داده‌های دارای پاسخ اسمی در بخش سوم ارائه می‌شوند. در بخش چهارم، روش‌های محاسباتی برای براورد کردن پارامترهای مدل‌های معرفی شده ارائه می‌شوند. مثال‌های کاربردی با پاسخ اسمی مقطعی در بخش پنجم آمده‌اند. در بخش ششم به نیکوبی برازش مدل پرداخته شده است و در بخش هفتم به چگونگی استفاده از نرم‌افزار R برای برازش مدل دارای پاسخ اسمی.

## ۱.۴ مفاهیم و نمادگذاری

اغلب، زمانی که پاسخ مورد مطالعه یک متغیر اسمی است، رده‌های این متغیر نمی‌توانند ترتیبی باشند. پاسخ‌های اسمی در هر حوزه‌ای از علوم اجتماعی کاربرد دارند. اسمیت واستراس (۱۹۷۵) دست‌یابی به شغل را به عنوان یک کاربرد اولیه از مدل‌های لوجیت چندجمله‌ای برای تحلیل داده‌های دارای پاسخ اسمی بررسی کردند. منگ و میلر (۱۹۹۵) تفاوت‌های جنسیتی در اشتغال را در کشور چین مورد بررسی قرار دادند. آریوم و شاویت (۱۹۹۵) اثر تحصیلات در دبیرستان را روی دست‌یابی به شغل، مطالعه کردند. مثال‌های دیگری نیز در حوزه‌های دیگر وجود دارد. هافمان و دانکن (۱۹۸۸) مدل‌های لوجیت شرطی و مدل‌های لوجیت چندجمله‌ای را در یک مطالعه‌ی وضع رفاه و ازدواج مقایسه کردند. اسپیکتور و مازو (۱۹۸۰) اثر یک روش تدریس تجربی را روی کارایی آموزش، مورد بررسی قرار دادند. از جمله‌ی مثال‌های دیگر می‌توان به دلایل ترک منزل (گلدشايدر و داوانزو، ۱۹۸۹) و انتخاب زبان در یک جامعه‌ی چندزبانی (استون، ۱۹۹۲) اشاره کرد.

مدل مورد استفاده برای پاسخ‌های اسمی، اغلب زمانی هم که متغیرهای وابسته ترتیبی‌اند، مورد استفاده قرار می‌گیرد. گاهی این کار به منظور جلوگیری از فرض‌های رگرسیون ترتیبی انجام می‌شود. در موارد دیگر ممکن است به دلیل اطلاع نداشتن از این‌که آیا متغیر وابسته باید به عنوان متغیر ترتیبی باشد، از مدل‌های اسمی استفاده شود. همچنین پژوهشگر ممکن است فقط با مدل‌های لوجیت چندجمله‌ای آشنا باشد. به هر حال، اگر یک متغیر وابسته‌ی ترتیبی داشته باشیم و برای تحلیل آن از یک مدل مربوط به متغیرهای اسمی استفاده کنیم، با از دست دادن کارایی روبرو خواهیم بود؛ چرا که اطلاع ترتیبی بودن متغیر پاسخ را از دست می‌دهیم. به عبارت دیگر، زمانی که مدل برای متغیرهای پاسخ اسمی به عنوان مدلی برای تحلیل متغیر وابسته‌ی ترتیبی به کار برده شود، برآوردهای به دست آمده ممکن است اریب و ناکارا باشند.

## ۲.۴ توزیع مناسب برای داده‌های مقطوعی دارای پاسخ

### اسمی

در این فصل، با متغیر پاسخ اسمی دارای  $J = 1, \dots, J$  رده موافق هستیم که هیچ‌گونه ترتیبی روی رده‌های پاسخ قائل نمی‌شویم. توزیع مورد استفاده برای مطالعات مقطوعی با پاسخ‌های اسمی، مانند مطالعات مقطوعی با پاسخ ترتیبی، توزیع چندجمله‌ای است که در فصل قبل به آن پرداخته شده است. بنا بر این به تکرار آن نمی‌پردازیم.

## ۳.۴ مدل‌های مختلف برای تحلیل داده‌های مقطوعی دارای

### پاسخ اسمی

#### ۱.۳.۴ مدل‌های لوجیت چندجمله‌ای

ساده‌ترین روش برای مدل‌بندی پاسخ‌های اسمی با  $J$  رده، قرار دادن یکی از رده‌های پاسخ به عنوان مرجع، و محاسبه‌ی لگاریتم بخت همه‌ی رده‌های دیگر نسبت به رده‌ای است که به عنوان مرجع در نظر گرفته شده است. این مطلب نیز در نظر گرفته خواهد شد که لگاریتم بخت، تابعی خطی از پیشگوها یا متغیرهای کمکی است. عموماً آخرین رده به عنوان مرجع در نظر گرفته می‌شود و بخت این که یک عضو گروه  $\theta$  در رده‌ی  $J$  قرار گیرد در مقابل این که عضو  $\theta$  ام در رده‌ی مبنای مرجع قرار گیرد،  $\frac{\pi_{ij}}{\pi_{iJ}}$ ، برای  $j = 1, \dots, J - 1$  را محاسبه می‌شود. شایان ذکر است که رده‌ی مبنای مرجع می‌تواند رده‌ای باشد که فراوانی آن نسبت به رده‌های دیگر بیشتر است.

در مدل لوجیت چندجمله‌ای، فرض می‌کنیم که لگاریتم بخت هر پاسخ به صورت تابع خطی زیر باشد:

$$\eta_{ij} = \log \frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + x'_i \beta_j, \quad j = 1, \dots, J - 1 \quad (1.4)$$

که  $\alpha$  مقدار ثابت و  $\beta$  برداری از ضرایب رگرسیونی برای  $j = 1, 2, \dots, J - 1$  است. این مدل شبیه مدل رگرسیون است با این تفاوت که توزیع احتمال پاسخ به جای دو جمله‌ای،

چندجمله‌ای است و ما  $1 - J$  معادله به جای یک معادله داریم.  $1 - J$  معادله‌ی لوجیت چندجمله‌ای، رده‌های دیگر را با رده‌ی  $J$  مقایسه می‌کنند، در حالی که رابطه‌ی رگرسیون لوژستیک فقط یک مقایسه بین موفقیت‌ها و شکست‌ها است. اگر  $2 = J$  باشد مدل لوجیت چندجمله‌ای به مدل رگرسیون لوژستیک معمولی تبدیل می‌شود. مدل لوجیت چندجمله‌ای ممکن است بر حسب احتمال‌های  $\pi_{ij}$  به جای نسبت بخت‌ها تفسیر شود. بنا بر این لازم است که احتمال‌های  $\pi_{ij}$  را با استفاده از رابطه‌ی (۱.۴) به دست آوریم. با شروع از رابطه‌ی (۱.۴) و پذیرفتن این قید که  $\eta_{iJ} = 0$  است، برای  $J = 1, \dots, J$  داشت:

$$\pi_{ij} = \frac{\exp(\eta_{ij})}{\sum_{k=1}^J \exp(\eta_{ik})}. \quad (2.4)$$

با استفاده از رابطه‌ی (۱.۴) و برقراری قید  $\eta_{iJ} = 0$  و انجام دادن محاسبات زیر می‌توان به رابطه‌ی (۲.۴) دست یافت:

$$\frac{\pi_{ij}}{\pi_{iJ}} = \exp(\eta_{ij}).$$

بنا بر این:

$$\pi_{ij} = \pi_{iJ} \exp(\eta_{ij}).$$

با جمع بستن در دو طرف تساوی فوق داریم:

$$\sum_{j=1}^J \pi_{ij} = \sum_{j=1}^J \pi_{iJ} \exp(\eta_{ij}).$$

بنا بر این:

$$1 = \pi_{iJ} \sum_{j=1}^J \exp(\eta_{ij})$$

و

$$\pi_{iJ} = \frac{1}{\sum_{j=1}^J \exp(\eta_{ij})}.$$

در نتیجه:

$$\pi_{ij} = \pi_{iJ} \exp(\eta_{ij}) = \frac{\exp(\eta_{ij})}{\sum_{j=1}^J \exp(\eta_{ij})}.$$

مدل‌های لوجیت چندجمله‌ای می‌توانند بر حسب بخت‌ها در هر دو سطح دلخواه نیز بیان شوند. به عنوان مثال، بخت پاسخ رده‌ی  $m$  در برابر پاسخ رده‌ی  $n$  برابر است با

$$\begin{aligned}\frac{\pi_{im}}{\pi_{in}} &= \frac{\frac{\exp(\eta_{im})}{\sum_{k=1}^J \exp(\eta_{ik})}}{\frac{\exp(\eta_{in})}{\sum_{k=1}^J \exp(\eta_{ik})}} = \frac{\exp(\eta_{im})}{\exp(\eta_{in})} = \frac{\exp(\alpha_m + x'_i \beta_m)}{\exp(\alpha_n + x'_i \beta_n)} \\ &= \exp[(\alpha_m - \alpha_n) + x'_i (\beta_m - \beta_n)].\end{aligned}$$

تفسیر رابطه‌ی فوق، وقتی پیشگوی خطی دارای یک متغیر کمکی است، این است که انتظار داریم به ازای یک واحد تغییر در متغیر  $x$ ، لگاریتم بخت رده‌ی  $m$  نسبت به رده‌ی  $n$  به اندازه‌ی  $\beta_m - \beta_n$  تغییر کند.

پس از معرفی مدل لوجیت چندجمله‌ای برای پاسخ‌های اسمی، به بررسی آزمون استقلال شرطی با استفاده از این مدل می‌پردازیم. فرض کنید  $Z$  یک متغیر کمکی اسمی باشد. مدل استقلال شرطی  $X$  و  $Y$  به شرط  $Z$  برای  $h = 1, 2, \dots, H$  به صورت زیر است:

$$\log \left[ \frac{\Pr(Y = j | X = h, Z = k)}{\Pr(Y = J | X = h, Z = k)} \right] = \alpha_{jk}, \quad j = 1, \dots, J-1 \quad k = 1, \dots, K$$

که در آن  $\alpha_{jk}$ ‌ها نقاط آستانه هستند که با تغییر رده‌ی پاسخ و رده‌ای که متغیر  $Z$  به آن تعلق دارد، تغییر می‌کنند. برای متغیر تصادفی ترتیبی یا پیوسته‌ی  $X$ ، مدل شرطی

$$\log \left[ \frac{\Pr(Y = j | X = x_i, Z = k)}{\Pr(Y = J | X = x_i, Z = k)} \right] = \alpha_{jk} + \beta_j x_i, \quad j = 1, \dots, J-1 \quad k = 1, \dots, K$$

می‌تواند برای آزمون استقلال شرطی  $X$  و  $Y$  به شرط  $Z$  به کار رود. اگر گواهی کافی بر رده  $H$  برای هر  $j$  به دست نیاید، می‌توان فرض استقلال شرطی  $X$  و  $Y$  (به شرط  $Z$ ) را پذیرفت.

یک مدل شرطی دیگر برای زمانی که متغیر تصادفی  $X$  اسمی باشد به صورت زیر معرفی می‌شود:

$$\begin{aligned}\log \left[ \frac{\Pr(Y = j | X = h, Z = k)}{\Pr(Y = J | X = h, Z = k)} \right] &= \alpha_{jk} + \beta_{jh}, \\ h &= 1, 2, \dots, H \quad j = 1, \dots, J-1 \quad k = 1, \dots, K\end{aligned}$$

که این مدل دارای محدودیت  $\beta_{Hj} = 0$  برای هر  $j$  است. این مدل می‌تواند برای آزمون استقلال شرطی به کار رود. اگر گواهی کافی بر رده  $H$  برای هر  $j$  به دست نیاید، می‌توان فرض استقلال شرطی  $X$  و  $Y$  (به شرط  $Z$ ) را پذیرفت.

## ۲.۳.۴ مدل لوجیت با استفاده از متغیر پنهان

فرض کنید  $Y_i$  معرف یک انتخاب گستته از میان  $J$  انتخاب باشد و  $U_{ij}$  معرف ارزش یا مطلوبیت زامین انتخاب برای نامین فرد باشد.  $U_{ij}$  را به عنوان متغیرهای تصادفی مستقل با یک مؤلفه‌ی سیستماتیک  $\eta_{ij}$  و یک مؤلفه‌ی  $\varepsilon_{ij}$  به‌گونه‌ای در نظر می‌گیریم که

$$U_{ij} = \eta_{ij} + \varepsilon_{ij}, \quad (3.4)$$

که در آن  $\eta_{ij}$  مقدار ثابتی است که با زامین رده‌ی پاسخ همبسته است و  $\varepsilon_{ij}, \dots, \varepsilon_{iJ}$  متغیرهای تصادفی مستقل با تابع توزیع پیوسته‌ی  $F$  هستند. به عبارت دیگر می‌توان گفت مدل‌هایی که برای پاسخ‌های اسمی در نظر گرفته می‌شوند، ممکن است از متغیرهای پنهان به وجود آمده باشند. در نظریه‌ی انتخاب تصادفی، اغلب فرض می‌شود که یک متغیر غیر قابل مشاهده‌ی  $U_{ij}$  با رده‌ی زام پاسخ، همبسته است. بنا بر این برای انتخاب نوع وسیله‌ی حمل و نقل، متغیر پنهان ممکن است به عنوان مطلوبیت مصرف‌کننده تعبیر و تفسیر شود. فرض می‌کنیم که افراد به یک روش منطقی عمل کنند و آن این باشد که مطلوبیت را ماکسیمم کنند. به این ترتیب، فرد نام رده‌ی زام را زمانی انتخاب می‌کند که  $U_{ij}$  بزرگ‌ترین مقدار  $U_{i1}, \dots, U_{iJ}$  را داشته باشد. بر اساس اصل ماکسیمم کردن مطلوبیت تصادفی، شرط لازم و کافی برای این که پاسخ مشاهده‌شده‌ی  $Y_i$  مقدار ز را اخذ کند، این است که

$$U_{ij} = \max_{j=1, \dots, J} U_{ij},$$

که این به آن معنا است که پاسخ، رده‌ی ز را می‌گیرد اگر متغیر پنهان  $U_{ij}$  بر اساس این رده ماکسیمم باشد. بنا بر این احتمال این که نامین فرد، انتخاب ز را داشته باشد برابر است با

$$\pi_{ij} = \Pr(Y_i = j) = \Pr(\max(U_{i1}, \dots, U_{iJ}) = U_{ij}) \quad (4.4)$$

واز رابطه‌ی (۴.۴) داریم:

$$\begin{aligned} \Pr(Y_i = j) &= \Pr(U_{ij} - U_{i1} \geq 0, \dots, U_{ij} - U_{iJ} \geq 0) \\ &= \Pr(\varepsilon_{i1} \leq \eta_{ij} - \eta_{i1} + \varepsilon_{ij}, \dots, \varepsilon_{iJ} \leq \eta_{ij} - \eta_{iJ} + \varepsilon_{ij}) \\ &= \int_{-\infty}^{\infty} \prod_{s \neq j} F(\eta_{ij} - \eta_{is} + \varepsilon) f(\varepsilon) d\varepsilon, \end{aligned} \quad (5.4)$$

که در آن  $f' = F$  تابع چگالی  $\varepsilon$  است و  $F$  تابع توزیع  $\varepsilon$  است. بسته به فرضی که برای توزیع متغیر  $\varepsilon$  در رابطه‌ی (۳.۴) در نظر می‌گیریم، به مدل‌های مختلفی دست می‌یابیم. اگر  $\varepsilon$ ‌ها به طور مستقل دارای توزیع نرمال باشند، به یک مدل پربویت مستقل دست می‌یابیم. تعمیم دیگر، مدل پربویت چندمتغیره است که با در نظر گرفتن متغیرهای خطای همبسته به دست می‌آید.

ارتباط بین توزیع مقدار فرین و مدل لوجیت توسط یلوت (۱۹۷۷) و مک‌فادن (۱۹۷۳) بیان شده است. مدل‌های کلی دیگر بر اساس ماکسیمم کردن مطلوبیت تصادفی در نوشتگان آمده است. مک‌فادن (۱۹۸۱) توزیع مقادیر فرین تعمیم‌یافته را در نظر گرفته است. هاسمان و وایز (۱۹۷۸)، داگانزو (۱۹۸۰)، لرمن و منسکی (۱۹۸۱) و مک‌فادن (۱۹۸۴) مدل‌های پربویت را در نظر گرفته‌اند که در آن‌ها فرض استقلال مطلوبیت‌ها را در نظر نمی‌گیرند (همچنین می‌توان به اسمال، ۱۹۸۷، و برش‌سوپان، ۱۹۹۰، مراجعه کرد).

می‌توان نشان داد که اگر جمله‌های خطای  $\varepsilon$  دارای توزیع مقادیر فرین نوع I با تابع چگالی

$$f(\varepsilon) = \exp(-\varepsilon - \exp(-\varepsilon))$$

باشند، رابطه‌ی

$$\pi_{ij} = \frac{\exp(\eta_{ij})}{\sum_{k=1}^J \exp(\eta_{ik})} \quad (۳.۴)$$

یک معادله‌ی ساده است که در مدل لوجیت چندجمله‌ای تعریف کردیم (به عنوان مثال به مادلا، ۱۹۸۳، صص. ۶۰-۶۱ مراجعه شود). در حالت ساده که  $J = 2$  باشد، فرم آن انتخاب اول را خواهد داشت اگر  $U_2 > U_1$  باشد. اگر مطلوبیت‌های تصادفی  $\varepsilon$  دارای توزیع‌های مستقل مقدار فرین باشند، می‌توان نشان داد که تفاضل‌های آن‌ها دارای توزیع لوژستیک است و می‌توانیم مدل رگرسیون لوژستیک استاندارد را به دست آوریم. با تغییر توزیع عبارت خطای در رابطه‌ی (۳.۴)، مدل‌های دیگری به وجود می‌آیند. یک انتخاب مورد توجه به جای مدل‌های لوجیت این است که  $\varepsilon$ ‌ها دارای توزیع‌های نرمال استاندارد مستقل برای همه‌ی  $(z_i, \eta)$  باشند. مدل به دست آمده، مدل پربویت

شرطی-چندجمله‌ای نامیده می‌شود که به نتایجی بسیار مشابه با مدل لوجیت شرطی-چندجمله‌ای منجر می‌شود.

## ۴.۴ روش‌های براورد

### ۱.۴.۴ روش شبه درست‌نمایی

همان‌طور که در بخش قبل نیز به آن اشاره کردیم، در مطالعات مقطعي، معادلات براوردگر تعمیم‌یافته برای پاسخ‌های مختلف گستته دارای صورت یکسان‌اند و تنها تفاوت آن‌ها در توزیع مورد استفاده در آن‌ها است، که این اختلاف منجر به تغییر میانگین و واریانس در معادلات براوردگر تعمیم‌یافته می‌شود. از آنجا که پاسخ‌های اسمی نیز از توزیع چندجمله‌ای استفاده می‌کنند، با در نظر گرفتن توزیع چندجمله‌ای برای پاسخ‌های اسمی در معادلات براوردگر، پارامترهای مدل براورد می‌شوند.

### ۲.۴.۴ معادلات درست‌نمایی و براوردگرهای ماکسیمم درست‌نمایی

اگر مشاهدات مستقل باشند، معادله‌ی درست‌نمایی به صورت زیر بیان می‌شود:

$$\begin{aligned} L(\beta, \alpha | y, X) &= \prod_{i=1}^n \Pr(Y_i = y_i | x_i, \beta, \alpha) \\ &= \prod_{i=1}^n \left[ \frac{\exp(\eta_{iy_i})}{\sum_{k=1}^J \exp(\eta_{ik})} \right], \end{aligned} \quad (7.4)$$

که در آن  $\eta_{iy_i}$  در معادله‌ی (۱.۴) تعریف شده است. این معادله می‌تواند از طریق نرم‌افزارهای موجود، مانند R و SPSS و SAS، به براورد پارامترهای مدل منجر شود. در صورتی که مشاهدات دارای وزن باشند (در آمارگیری‌ها ممکن است این وزن‌ها از چند مرحله‌ی وزن‌دهی به وجود آمده باشند)، باید وزن‌ها نیز در معادلات درست‌نمایی در نظر گرفته شوند. با در نظر گرفتن وزن‌ها معادلات درست‌نمایی، که بهتر است آن‌ها را معادلات

شبه درست‌نمایی بنامیم، به صورت زیر بیان می‌شوند:

$$\begin{aligned} L(\beta, \alpha | y, X) &= \prod_{i=1}^n [\Pr(Y_i = y_i | x_i, \beta, \alpha)]^{W_i} \\ &= \prod_{i=1}^n \left[ \frac{\exp(\eta_i y_i)}{\sum_{k=1}^J \exp(\eta_i k)} \right]^{W_i}. \end{aligned} \quad (8.4)$$

برآورد پارامترها در این معادلات نیز در نرم‌افزارهای موجود همچون SAS و R قابل محاسبه است.

## ۵.۴ مثال‌های کاربردی

### ۱.۵.۴ وضع فعالیت اقتصادی سرپرست خانوار در طرح آمارگیری نیروی کار سال ۱۳۸۵

در این بخش، از داده‌های بیکاران طرح آمارگیری نیروی کار مرکز آمار ایران در بهار ۱۳۸۵ استفاده شده است. وضع فعالیت اقتصادی این بیکاران در تابستان ۱۳۸۵ به عنوان پاسخ اسمی در نظر گرفته می‌شود و اثر برخی متغیرهای کمکی روی پاسخ، مورد مطالعه قرار می‌گیرد. برخی از آمارهای توصیفی و معرفی داده‌ها در فصل اول زیربخش ۳.۲.۱ بیان شده‌اند. مدل مورد استفاده برای این داده‌ها مدل لوجیت تعمیم‌یافته است که به صورت زیر بیان می‌شود:

$$\begin{aligned} \log \left[ \frac{\Pr(Y_i = l | \alpha, \beta)}{\Pr(Y_i = J | \alpha, \beta)} \right] &= \alpha_l + \beta_{l,1} AGE_i + \beta_{l,2} RHH_i + \beta_{l,3} MIS_i \\ &\quad + \beta_{l,4} AS_i + \beta_{l,5} LA_i + \beta_{l,6} GENDER_i \\ &\quad + \sum_{j=1}^2 \beta_{l,7,j} AL_{ij} + \sum_{j=1}^2 \beta_{l,8,j} MS_{ij} + \sum_{j=1}^2 \beta_{l,9,j} NEIH_{ij} \\ &\quad + \sum_{j=1}^2 \beta_{l,10,j} NIH_{ij}, \end{aligned} \quad (9.4)$$

که در آن  $J = 3$  و  $l = 1, 2$  است. متغیرهای کمکی و کدهای مربوط به آن‌ها در جدول ۲.۳ آمده‌اند. همان‌طور که قبلاً اشاره شد، متغیرهای کمکی‌ای که بیش از ۲ سطح داشته

## فصل ۴. مطالعات مقطعی: داده‌های دارای پاسخ اسمی

باشند (مانند سطح تحصیلات و وضع زناشویی)، به عنوان عامل در نظر گرفته می‌شوند و با تعریف معمول به متغیرهای دودویی تبدیل می‌شوند. به عنوان مثال، برای وضع زناشویی با ۳ سطح،  $MS_{i1}$ ، یک متغیر دودویی برای افراد ازدواج کرده ( $1 = MS_{i1}$  برای افراد ازدواج کرده و  $0 = MS_{i1}$  در غیر این صورت) و  $MS_{i2}$  یک متغیر نشانگر دودویی برای افراد بیوه و طلاق گرفته است.

اکنون با استفاده از مدل یک متغیره ای اسمی به تحلیل داده‌های نیروی کار می‌پردازیم. نتایج مدل‌بندی وضع فعالیت اقتصادی به عنوان پاسخ اسمی در جدول ۱.۴ ارائه شده است. در این مطالعه، نخست همه‌ی متغیرهایی که در رابطه‌ی (۹.۴) آمده‌اند وارد مدل شده‌اند. سپس با استفاده از روش انتخاب پس‌رو، متغیرهایی که معنادار نبوده‌اند از مدل حذف شده‌اند. نتایج این جدول که بر اساس مدل لوجیت تعییم‌یافته (وضع غیر فعال به عنوان مبنای) به دست آمده است، اثرهای متغیرهایی کمکی مختلف را روی وضع فعالیت اقتصادی نشان می‌دهد. این نتایج همچنین برای مردان و زنان به‌طور مجزا بیان شده است. برای مردان، اثر سن معنادار است، اما این اثر برای کل جامعه و زنان، معنادار نیست. این اثر نشان می‌دهد که هرچه سن مردان افزایش می‌یابد، بخت بیکار ماندن آن‌ها نسبت به غیر فعال شدن شان کاهش می‌یابد. اثر «سرپرست خانوار بودن» نیز معنادار نیست. نتایج همچنین نشان می‌دهد که اثر وضع تحصیل برای همه‌ی مردم و مردان معنادار است و این اثر نشان می‌دهد که بخت برآورد شده‌ی بیکار ماندن نسبت به غیر فعال شدن برای افرادی که در حال تحصیل هستند، نسبت به افرادی که تحصیل نمی‌کنند، کمتر است. همچنین این بخت در مناطق شهری نسبت به مناطق روستایی برای مردان و کل جمعیت، نسبتاً بیشتر است. نتایج همچنین نشان می‌دهد که بخت برآورد شده‌ی بیکار ماندن نسبت به غیر فعال شدن برای مردان بیشتر است. این به آن معنا است که احتمال باقی ماندن در نیروی کار (به عنوان بیکار) در طول مدت مطالعه، برای مردان بیکار نسبت به زنان بیکار، بیشتر است. اثر وضع زناشویی نیز نشان می‌دهد که بخت برآورد شده‌ی بیکار ماندن نسبت به غیر فعال شدن برای زنانی که ازدواج کرده‌اند، نسبت به زنانی که ازدواج نکرده‌اند، کمتر است. این به آن معنا است که زنانی که ازدواج کرده‌اند با احتمال زیادی غیر فعال می‌شوند. اثر وضع زناشویی برای مردان، معنادار نشده است.

به منظور مقایسه‌ی اثر متغیرهایی کمکی بر روی بخت برآورد شده‌ی شاغل شدن نسبت

به غیرفعال شدن، نتایج بخش دوم جدول ۱.۴ تحت عنوان «وضع فعالیت (شاغل)» به صورت زیر بیان شده است. اثر سن نشان می‌دهد که بخت برآورد شده‌ی شاغل شدن نسبت به غیرفعال شدن برای افراد مسن کمتر است. نتیجه‌ی مشابه برای مردان نیز وجود دارد. اثر «سرپرست خانوار بودن» نشان می‌دهد که بخت برآورد شده‌ی شاغل شدن نسبت به غیرفعال شدن برای سرپرست خانوار نسبت به بقیه‌ی اعضای خانوار بیشتر است. نتایج همچنین نشان می‌دهد که بخت برآورد شده‌ی شاغل شدن نسبت به غیرفعال شدن برای افرادی که در حال تحصیل هستند، نسبت به افرادی که تحصیل نمی‌کنند، کمتر است. برای اثر جنس، نتایج نشان می‌دهد که این بخت برای مردان نسبت به زنان بیشتر است. اثر وضع زناشویی نشان می‌دهد که این بخت برای مردانی که ازدواج کرده‌اند نسبت به زنانی که ازدواج نکرده‌اند، کمتر است.

جدول ۱.۴: برآورد پارامترها و انحراف مدل لوجیت تعیین‌یافته برای وضع فعالیت اقتصادی در نیروی کار

پارامتر		کل جمعیت		مردان		زنان	
		برآورد	انحراف معیار	برآورد	انحراف معیار	برآورد	انحراف معیار
وضع فعالیت (بیکار)							
۰/۴۷۵	-۰/۱۸۱	۰/۳۶۵	۱/۴۱۷	۰/۳۱۵	۰/۱۷۳	۰۲,۱	
۰/۵۱۸	۰/۰۰۹	۰/۰۱۴	-۰/۰۳۵	۰/۰۱۱	-۰/۰۱۸	سن	
-	۰/۲۰۹	۰/۲۲۷	۰/۲۲۸	-۰/۲۶۴	۰/۲۶۳	۰/۴۹۰	بسنگی با سرپرست خانوار مبنای: بقیه‌ی اعضای خانوار سرپرست خانوار
-	۰/۴۶۹	-۰/۷۰۰	۰/۴۳۰	-۲/۹۴۶	۰/۳۹۶	-۲/۰۴۳	وضع تحصیل مبنای: در حال تحصیل بودن در حال تحصیل بودن
-	۰/۲۵۷	۰/۱۱۸	۰/۲۱۲	۰/۷۱۰	۰/۱۶۹	۰/۴۶۷	محل زندگی مبنای: روستا شهر
-	-	-	-	-	-	۱/۰۲۷	جنس مبنای: زن
-	-	-	-	۰/۱۶۰	-	-	مرد
-	۰/۲۴۸	-۱/۰۲۲	۰/۳۱۸	۰/۵۲۵	۰/۱۲۹	-۰/۶۳۷	وضع زناشویی مبنای: هرگز ازدواج نکرده ازدواج کرده
-	۰/۸۵۰	-۱/۰۷۴	۰/۸۰۳	-۰/۵۲۱	۰/۰۶۵	-۰/۵۹۲	طلاق گرفته یا بیوه شده اطلاق گرفته یا بیوه شده
وضع فعالیت (شاغل)							
۰/۶۶۳	-۱/۴۹۰	۰/۲۴۶	۲/۴۹۷	۰/۲۳۷	-۰/۰۸۲	۰۲,۲	
۰/۰۲۴	۰/۰۲۳	۰/۰۱۳	-۰/۰۶۷	۰/۰۱۰	-۰/۰۴۸	سن	
-	۱/۱۱۱	-۰/۸۳۱	۰/۳۱۸	۰/۲۴۲	۰/۲۴۸	۰/۸۹۳	بسنگی با سرپرست خانوار مبنای: بقیه‌ی اعضای خانوار سرپرست خانوار
-	۰/۱۸۶	-۲/۰۴۳	۰/۲۸۵	-۲/۴۸۰	۰/۳۷۸	-۲/۲۰۹	وضع تحصیل مبنای: در حال تحصیل بودن در حال تحصیل بودن
-	۰/۲۸۱	-۰/۲۰۳	۰/۱۹۷	-۰/۰۲۷	۰/۱۶۷	-۰/۲۰۱	محل زندگی مبنای: روستا شهر
-	-	-	-	۰/۲۰۴	-	۲/۳۱۰	جنس مبنای: زن
-	-	-	-	۰/۲۰۴	-	-	مرد
-	۰/۳۸۰	-۰/۸۷۳	۰/۳۱۴	۱/۴۳۴	-	-	وضع زناشویی مبنای: هرگز ازدواج نکرده ازدواج کرده
-	۰/۷۷۲	۰/۸۶۳	۰/۶۷۶	۱/۰۴۷	۰/۱۹۴	۰/۲۹۶	طلاق گرفته یا بیوه شده اطلاق گرفته یا بیوه شده
۵۹۸۸۴۹/۰		۱۶۰۰۹۷۵۰/۷		۲۲۵۹۳۳۳/۴		- $\log L_{Null}$	
۵۷۵۹۲۰/۲		۱۴۷۲۲۷۴/۷		۲۰۸۱۶۴۳/۱		- $\log L_{Full}$	

در جدول ۱.۴،  $2 \log L_{\text{Null}}$  - منفی دو برابر لگاریتم درست‌نمایی بدون در نظر گرفتن هیچ متغیر کمکی است و  $2 \log L_{\text{Full}}$  - منفی دو برابر لگاریتم درست‌نمایی با در نظر گرفتن همهٔ متغیرهای کمکی در مدل است. اختلاف فاحش این دو مقدار نشان می‌دهد که باید متغیرهای تبیینی ذکر شده در مدل در نظر گرفته شوند.

## ۲.۵.۴ بررسی نحوهٔ مشارکت در نیروی کار بر اساس اطلاعات آمارگیری اجتماعی جمعیت کانادا در سال ۱۹۷۷ میلادی

بخشی از داده‌هایی که از یک آمارگیری اجتماعی جمعیت کانادا در سال ۱۹۷۷ میلادی به دست آمده، در جدول ۲.۴ داده شده است. این داده‌ها در مورد اطلاعات مربوط به ۲۶۳ زن ازدواج کرده‌ی بین سنین ۲۱ تا ۳۰ سال می‌باشد. در این جدول، متغیر «نحوهٔ حضور یا مشارکت در نیروی کار» (Participation) که سه رده‌ی بیکار (not.work)، حضور نیمه‌وقت (parttime) و شاغل تمام‌وقت (fulltime) را اخذ می‌کند، به عنوان شاغل نیمه‌وقت (parttime) و شاغل تمام‌وقت (fulltime) را اخذ می‌کند، به عنوان متغیر پاسخ در نظر گرفته شده است. همچنین متغیر «درآمد شوهر»، بر حسب ۱۰۰۰ دلار (Hincome) بیان شده است. متغیر «حضور فرزند در خانوار» (Children) که دو رده‌ی غایب (absent) و حاضر (present) را اخذ می‌کند و متغیر «ناحیه» که ۵ ناحیه در کشور کانادا را نشان می‌دهد نیز در جدول آمده‌اند. این نواحی عبارت‌اند از آتلانتیک (Atlantic)، اوونتاریو (Ontario)، بریتیش کلمبیا (BC)، کیبک (Quebec) و پرایری (Prairie).

جدول ۲.۴: بخشی از داده‌های مربوط به آمارگیری اجتماعی جمعیت کانادا

ناحیه	حضور فرزند در خانوار	درآمد شوهر	نحوهٔ حضور و مشارکت در نیروی کار
Atlantic	۱۵	present	not.work
Ontario	۲۸	absent	parttime
BC	۲۲	absent	fulltime
Ontario	۱۵	absent	fulltime
Atlantic	۲۳	present	not.work
Ontario	۱۹	present	not.work
Ontario	۷	present	not.work
Ontario	۱۸	absent	fulltime
Ontario	۱۳	absent	not.work
Quebec	۱۳	present	not.work

## ۶.۴ نیکویی برازش مدل اسمی

۱۲۱

مدل مورد استفاده برای تحلیل این داده‌ها، مدل لوجیت چندجمله‌ای است که به صورت زیر بیان می‌شود:

$$\log \left[ \frac{\Pr(Y_i=l|\alpha, \beta)}{\Pr(Y_i=J|\alpha, \beta)} \right] = \alpha_l + \beta_{l,1} \text{Hincome}_i + \beta_{l,2} \text{Children}_i$$

که در آن  $J = 1, 2 = l$  است. مبنای متغیر پاسخ، سطح حضور و مشارکت تمام وقت در نظر گرفته شده است. نتایج این مدل که به مدل‌بندی نحوه‌ی حضور و مشارکت در نیروی کار می‌پردازد، تحت عنوان مدل ۱ در زیربخش ۷.۴ ارائه شده است. می‌توان اثر متغیر ناحیه بر این متغیر پاسخ را نیز مورد بررسی قرار داد که تحت عنوان مدل ۲ در زیربخش ذکر شده آمده است.

## ۶.۴ نیکویی برازش مدل اسمی

قبل از وارد شدن به مبحث مربوط به محاسبه‌ی نتایج مربوط به مطالعات مقطعی با پاسخ اسمی با استفاده از نرم‌افزار  $R$ ، به معرفی معیار نیکویی برازش مدل در این نوع پاسخ‌ها می‌پردازیم؛ چرا که در نظر گرفتن نیکویی برازش مدل نیز یک مسئله‌ی مهم است که باید به آن پرداخته شود. برای پاسخ‌های اسمی، ضریب پیش‌اینده  $R_c^2$  (آلدریچ و نلسون، ۱۹۸۴) به عنوان یک آماره‌ی نیکویی برازش به کار می‌رود، که به صورت زیر تعریف می‌شود:

$$R_c^2 = \frac{G_M}{(G_M + n)},$$

که در آن داریم  $G_M = -2[\log L_{\text{Null}} - \log L_{\text{Full}}]$ ، و  $n$  تعداد مشاهدات است. این آماره در بازه‌ی  $(0, 1)$  مقدار اختیار می‌کند. هرچه  $R_c^2$  بزرگ‌تر باشد، برازش مدل بهتر بوده است.

## ۷.۴ دستورهای R برای برازش مدل در داده‌های مقطعی با

### پاسخ اسمی

پس از فراخوانی بخشی از داده‌های مربوط به اطلاعات آمارگیری اجتماعی جمعیت کانادا در سال ۱۹۷۷ میلادی (پیوست ۳.۲ را ببینید)، با استفاده از دستورهای زیر به برازش مدل لوجیت چندجمله‌ای می‌پردازیم.

```
data<-read.table("womenlf.txt",header=TRUE)
attach(data)
```

برای برازش مدل لوجیت چندجمله‌ای به داده‌های فوق با استفاده از تابع `multinom` در کتابخانه `nnet` و همچنین با استفاده از کتابخانه `MASS` می‌توان به برازش مدلی مناسب به این داده‌ها با پاسخ اسمی پرداخت.

```
library(MASS)
library(nnet)
mod1.multinom <- multinom(Participation ~ hincome+children)
summary(mod1.multinom, cor=F, Wald=T)
```

در مدل فوق که به مدل ۱ (`mod1.multinom`) معروف است، فقط اثر متغیر کمکی درآمد شوهر (`hincome`) و حضور فرزند (`children`) بر نحوهٔ حضور و مشارکت در نیروی کار (`participation`) را مورد بررسی قرار داده‌ایم، که نتایج آن در شکل ۱.۴ آمده است. مشخص کردن گزینه‌ی `cor=T` در دستور `summary`، ماتریس همبستگی ضرایب برآورد شده را به دست می‌دهد و قرار دادن `Wald=T` در دستور `summary` نیز آماره‌ی والد برای هر یک از ضرایب را ارائه می‌کند.

اگر بخواهیم اثر متغیر ناحیه (`region`) بر نحوهٔ حضور و مشارکت در نیروی کار را نیز ببینیم، از مدل ۲ (`mod2.multinom`) به صورت زیر استفاده می‌کنیم. نتایج این مدل نیز در شکل ۲.۴ آمده است.

```

Call:
multinom(formula = partic ~ hincome + children)

Coefficients:
            (Intercept)      hincome   childrenpresent
not.work     99.49448    -7.116437    74.137310
parttime   -244.35286    8.886173   -2.002361

Std. Errors:
            (Intercept)      hincome   childrenpresent
not.work     346.8797    24.95399   2.712987e+02
parttime   362.8280    13.19552   1.734951e-16

Value/SE (Wald statistics):
            (Intercept)      hincome   childrenpresent
not.work     0.2868270   -0.2851824   2.732682e-01
parttime   -0.6734676    0.6734233   -1.154131e+16

Residual Deviance: 0.05014073
AIC: 12.05014

```

شکل ۱.۴: نتایج برازش مدل ۱ برای داده‌های آمارگیری اجتماعی جمعیت کانادا در سال ۱۹۷۷ میلادی

```

Coefficients:
            (Intercept)      hincome   childrenpresent
not.work     120.7543   -10.829391   187.91068
parttime   -173.6612    9.073531   -16.77298
            regionB   region@tario   regionQuebec
not.work     -36.72773   31.92860   20.0135694
parttime   -136.97019   -25.84745   -0.1494796

Std. Errors:
            (Intercept)      hincome   childrenpresent
not.work     1.501756e+03   2.029033e+02   1.613586e-08
parttime   1.613587e-08   3.711249e-07   1.613586e-08
            regionB   region@tario   regionQuebec
not.work     5.814868e-81   1.501756e+03   1.289138e-72
parttime   1.019301e-27   4.239289e-15   3.702430e-104

Residual Deviance: 0.0001292281
AIC: 24.00013

```

شکل ۲.۴: نتایج برازش مدل ۲ برای داده‌های آمارگیری اجتماعی جمعیت کانادا در سال ۱۹۷۷ میلادی

```
mod2.multinom <- multinom(Participation ~ hincome+children+region )
summary(mod2.multinom)
```

به منظور مقایسهٔ دو مدل می‌توان از آماره‌ی اختلاف انحرافات به صورت زیر استفاده

کرد:

```
deviance(mod2.multinom)-deviance(mod1.multinom)
```

که با مقایسهٔ این کمیت با مقدار خی دو با درجهٔ آزادی‌ای مساوی با اختلاف تعداد پارامترهای دو مدل، در سطح ۵ درصد می‌توان به مقایسهٔ دو مدل پرداخت. در مورد این دو مدل می‌توان احتمال‌های برازنده شده را برای هر سه ردیهٔ پاسخ با استفاده از یکی از دستورهای زیر به دست آورد:

```
p.fit <- predict(mod2.multinom,type="probs")
p.fit <- fitted.values(mod2.multinom,type="probs") .
```

## ۸.۴ تمرین‌ها

۱ - در جدول ۳.۴، فرض کنید  $Y$  اعتقاد به زندگی پس از مرگ (با سه ردیهٔ بلهٔ ۱، نامعلوم = ۲ و نه = ۳)،  $X_1$  جنسیت ( $G$ ، ۱ = زن، ۰ = مرد)، و  $X_2$  نژاد ( $R$ ، ۱ = سفید، ۰ = سیاه) باشد. جدول ۴.۴ نتایج برآش مدل زیر را با مقادیر انحراف معیار در داخل پرانتز نشان می‌دهد.

$$\log\left(\frac{\pi_j}{\pi_3}\right) = \alpha_j + \beta_j^G X_1 + \beta_j^R X_2, \quad j = 1, 2$$

(۱) معادله‌ی پیشگو برای  $\log\left(\frac{\pi_1}{\pi_3}\right)$  را بیابید.

(۲) با استفاده از ردیه‌های پاسخ «بلی و نه»، اثر شرطی جنسیت را با استفاده از فاصله‌ی

اطمینان ۹۵ درصدی برای نسبت بخت‌ها تفسیر کنید.

پ) نشان دهید که برای زنان سفید، داریم  $0/76 = 1, G = 1) = R|_{\text{بلى}} = \hat{\pi}_1 = \hat{P}_{\text{r}}(Y =$

ت) بدون محاسبه‌ی احتمال‌های براورد شده، توضیح دهید که چرا براوردهای عرض از مبدأ نشان می‌دهند که برای مردان سیاهپوست،  $\hat{\pi}_2 > \hat{\pi}_1$  است. از براوردهای عرض از مبدأ و جنس برای نشان دادن این که چنین ترتیبی برای زنان سیاهپوست نیز برقرار است، استفاده کنید.

ث) بدون محاسبه‌ی احتمال‌های براورد شده، توضیح دهید که چرا براوردهای سطري جنسیت و نژاد نشان می‌دهند که  $\hat{\pi}_3$  بیشترین مقدار خود را برای مردان سیاهپوست دارد.

ج) برای این برازش،  $G_M = 9$  است. توضیح دهید که چرا درجه‌ی آزادی مانده‌ها برابر با ۲ است.

چ) با حذف اثر جنسیت،  $G_M = 8$  است. آزمون کنید که آیا به شرط نژاد، اعتقاد از جنسیت مستقل است یا نه. توضیح دهید.

جدول ۳.۴: داده‌های مربوط به اعتقاد به زندگی پس از مرگ

عقیده به زندگی پس از مرگ		نژاد	جنسیت	بلی	نامعلوم	نه
سفید	زن					
۷۴	۴۹	۳۷۱				
۷۱	۴۵	۲۵۰	مرد			
۱۵	۹	۶۴	زن			
۱۳	۵	۲۵	مرد			

جدول ۴.۴: نتایج مربوط به داده‌های مربوط به اعتقاد به زندگی پس از مرگ

ردۀای عقیده برای لوجیت		پارامتر	بلی / نه	نامعلوم / نه
-۰/۷۵۸ (۰/۳۶۱)	۰/۸۸۳ (۰/۲۴۳)	عرض از مبدأ		
۰/۱۰۵ (۰/۲۴۶)	۰/۴۱۹ (۰/۱۲۱)	جنس		
۰/۲۷۱ (۰/۳۵۴)	۰/۳۴۲ (۰/۲۳۷)	نژاد		

۲- یک مدل که برتری حزب رئیس جمهور آمریکا [دموکرات (Democrat)، جمهوری خواه (Republican) و مستقل (Independent)] را با استفاده از متغیر کمکی  $X$  درامد سالانه (بر حسب ۱۰۰۰۰ دلار) برازش می‌دهد، به صورت  $X - ۰/۲X^0 = ۳/۳ - \log(\frac{\hat{\pi}_D}{\hat{\pi}_I})$  است.

$$1 + \frac{1}{3}X = \log(\frac{\hat{\pi}_D}{\hat{\pi}_I})$$

## فصل ۴. مطالعات مقطعی: داده‌های دارای پاسخ اسمعی

- (آ) معادله‌ی پیش‌بینی برای  $\log(\frac{\hat{\pi}_R}{\hat{\pi}_D})$  را بیابید و شیب را تفسیر کنید. نشان دهید که برای چه دامنه‌ای داریم  $\hat{\pi}_R > \hat{\pi}_D$ .
- (ب) معادله‌ی پیش‌بینی برای  $\hat{\pi}_I$  را بیابید.
- (پ)  $\hat{\pi}_D$  و  $\hat{\pi}_I$  را برای مقادیر  $X$  بین  $0^\circ$  و  $10^\circ$  رسم کنید و نمودار حاصل را تفسیر کنید.

۳- جدول ۵.۴، اشاره به اثرهای جنسیت و نژاد در تشخیص حزب سیاسی افراد دارد. یک مدل لوجیت بر مبنای رده‌ی مرجع را که برازش خوبی به داده‌ها دهد پیدا کنید و اثرهای براورد شده را بر روی بخت این‌که حزب دموکرات به جای حزب جمهوری خواه انتخاب شود، تفسیر کنید.

جدول ۵.۴: داده‌های مربوط به تشخیص حزب سیاسی افراد

		حزب سیاسی		جنسيت			
		جمهوری خواه	دموکرات	نژاد	مرد	زن	
		مستقل	۱۲۲	۱۲۲	سفید	سفید	
۱۲		۶	۴۲	۴۲	سیاه	سیاه	
۱۳۰		۱۲۹	۱۷۲				
۱۵		۴	۵۶				

۴- جدول ۶.۴، نتایج مدل لوجیت پیشرفت شغلی در آمریکا را با استفاده از متغیرهای کمکی موردنظر S: سال‌های تحصیل، E: تجربه‌ی بازار کار (محاسبه شده از طریق، سن - سال‌های تحصیل - ۵)، R: نژاد ( $1 =$  سفید،  $0 =$  سیاه) و G: جنسیت ( $1 =$  مرد،  $0 =$  زن) بیان می‌کند. رده‌های پیشرفت شغلی، تخصصی (P)، کارمند دفتری (W)، کارگری (B)، پیشه‌ور (C) و خادم (M) است.

- (آ) براورد پارامترها را برای مدل‌بندی  $\log(\frac{\pi_W}{\pi_B})$  به دست آورید و آن‌ها را تفسیر کنید.
- (ب) توضیح دهید که چرا براورد ستون نژاد نشان می‌دهد که گروه‌های شغلی (W، C، P، M) بر حسب تعداد نسبی کارگران سفیدپوست، به شرط عامل‌های دیگر، مرتب شده‌اند.

جدول ۶.۴: نتایج مربوط به مدل لوجیت پیشرفت شغلی در آمریکا

تابع لوجیت	نقطه‌ی برش	تحصیل	تجربه‌ی بازار کار	نژاد	جنسيت
$\log(\frac{\pi_B}{\pi_M})$	$1/056$	$-0/124$	$-0/015$	$0/700$	$1/252$
$\log(\frac{\pi_C}{\pi_M})$	$-3/769$	$-0/001$	$-0/008$	$1/458$	$3/112$
$\log(\frac{\pi_W}{\pi_M})$	$-3/305$	$0/225$	$0/003$	$1/762$	$-0/522$
$\log(\frac{\pi_P}{\pi_M})$	$-5/959$	$0/429$	$0/008$	$0/976$	$0/656$

#### ۱.۴. تمرین‌ها

۱۲۷

۵- سه متغیر رسته‌ای  $X$  و  $Y$  و  $Z$  را در نظر بگیرید.

- (آ) نشان دهید که اگر  $Z$  مشترکاً مستقل از  $X$  و  $Y$  باشد،  $X$  و  $Y$  به شرط  $Z$  مستقل‌اند.
- (ب) اگر  $X$  از  $Y$ ، و  $Y$  از  $Z$  مستقل باشد، آیا این نتیجه می‌دهد که  $X$  و  $Z$  مستقل‌اند؟  
توضیح دهید.
- (پ) اگر هر دو متغیر به شرط متغیر دیگر مستقل باشند، آیا اثر متقابل سه‌تایی وجود دارد؟  
توضیح دهید.