

# روش معادلات برآوردگر تعمیم یافته استوار و کاربرد آن در مدل‌هایی با برآمدهای دودویی

همبسته

مسعود یارمحمدی<sup>1\*</sup>؛ سعید معدنی<sup>2</sup>

دریافت: 1394/02/13

پذیرش: 1394/04/20

## چکیده

روش معادلات برآوردگر تعمیم یافته توسط لیانگ و زیگر (1986) به عنوان روشی برای تسهیل در تحلیل داده‌های جمع‌آوری شده به صورت طولی، آشیانه‌ای و اندازه‌های مکرر معرفی شد. این روش از مدل خطی تعمیم یافته برای برآوردهایی کارا و نااریب پارامترهای رگرسیونی نسبت به برآورد ضرایب رگرسیونی در مدل‌های خطی تعمیم یافته، هنگامی که همبستگی نامشخصی در میان مشاهدات موجود باشد، استفاده می‌کند. این روش در رابطه با داده‌های دورافتاده متاثر شده و کارایی خود را از دست می‌دهد. لذا به منظور کاهش اثرات این دسته از مشاهدات، روش‌های استوارسازی معادلات برآوردگر تعمیم یافته را برای دو کلاس شوئیپ و مالوس معرفی کرده و سپس آنها را با استفاده از روش‌های شبیه‌سازی برای مدل‌هایی با برآمدهای دودویی همبسته مورد بحث و بررسی قرار می‌دهیم. **واژگان کلیدی:** معادلات برآوردگر تعمیم یافته، معادلات برآوردگر تعمیم یافته استوار، کلاس شوئیپ، کلاس مالوس.

---

1. دانشیار، گروه آمار دانشگاه پیام نور (\*نویسنده مسئول) [masyar@pnu.ac.ir](mailto:masyar@pnu.ac.ir)

2. دانشجوی دکترا، آمار دانشگاه پیام نور [s\\_madani2000@yahoo.com](mailto:s_madani2000@yahoo.com)

## 1. مقدمه

انجام یافته در زمینه استوارسازی روش معادلات برآوردگر تعمیم یافته می توان به تحقیقات پان<sup>7</sup> (2001) و وانگ و لونگ<sup>8</sup> (2010) اشاره کرد. به طور کلی در روش **GEE** مدل های حاشیه ای را برای پاسخ ها یا خوشه های همبسته برازش داده و از یک برآوردگر فشرده برای ماتریس واریانس کواریانس ضرایب رگرسیونی استفاده می شود. اگرچه این برآوردگر نسبت به ساختار همبستگی پاسخ ها نسبتا استوار است، پان و سپس وانگ و لونگ، برآوردگرهای تصحیح شده استوار دیگری که برای نمونه های متناهی از اربیی کمتر و کارایی بالاتری برخوردار می باشند، معرفی کردند.

در این مقاله با استفاده از روش کارول و پدرسن<sup>9</sup> (1993) در استوار سازی رگرسیون لوژستیک، روش های استوارسازی معادلات برآوردگر تعمیم یافته را مطرح و آنها را برای مدلی شبیه سازی شده با پاسخ های دودویی همبسته به کار خواهیم برد. در این روش ها یک ماتریس وزن دهی قطری که به هر مشاهده وزنی بین صفر و یک می دهد، به کار می رود. این عمل تاثیر مشاهدات دورافتاده را روی برآورد پارامترها کم می کند. وزن دهی به مشاهدات در دو کلاس مالوس و شوئیپ انجام می شود. در کلاس مالوس وزن دهی بر مبنای خاصیت اهرمی مشاهدات و در کلاس شوئیپ بر مبنای مانده ها انجام می شود، برای توضیحات بیشتر به کاکیش و پرایسر<sup>10</sup> (1999) مراجعه شود. مطالب ارایه شده در بخش های بعدی به شرح زیر است: نخست معادلات برآوردگر تعمیم یافته را معرفی می شود. سپس روش کمترین مربعات دوباره وزنی شده تکراری برای حل معادلات برآوردگر تعمیم یافته بیان می گردد. در ادامه معادلات

روش معادلات برآوردگر تعمیم یافته (**GEE**)<sup>1</sup> امروزه به طور وسیعی در بیولوژی و پزشکی و صنایعی که در تحلیل داده های آنها همبستگی بین مشاهدات جمع آوری شده دیده می شود، به کار می رود. بیشترین کاربرد این روش در اندازه گیری های مکرر یا مطالعات طولی است، که در آنها یک فرد یا یک موضوع در موقعیت های زمانی یا مکانی به صورت مکرر مورد بررسی قرار می گیرند. این روش برای نخستین بار توسط لیانگ و زیگر<sup>2</sup> (1986) معرفی شد و سپس توسط پرنیتیس<sup>3</sup> (1988) و ژائو و پرنیتیس<sup>4</sup> (1990) تعمیم یافت. در این روش ها که به ترتیب به **GEE1** و **GEE2** مشهورند، اندازه های تکراری به صورت دوحالتی (0/1) در نظر گرفته می شوند، با این تفاوت که در روش **GEE1** همبستگی بین مشاهدات به عنوان یک پارامتر مزاحم در نظر گرفته می شود، اما در **GEE2** پارامتر ارتباط یا همبستگی به اندازه پارامترهای رگرسیونی مهم تلقی می شود و برآوردی از آن به دست می آید. مطالعات مروری مختلفی در مورد این دو روش صورت گرفته است. به عنوان نمونه خواننده می تواند به مقالات لیانگ و همکاران<sup>5</sup> (1992)، زیگر و لیانگ<sup>6</sup> (1992) مراجعه کند.

یکی از مشکلات مهم در استفاده از این روش وقوع داده های دورافتاده است که معمولا باعث تغییرات زیادی در برآورد ضرایب رگرسیونی می شود. جهت کاهش اثر این داده ها و مقاوم سازی این روش در برخورد با داده های دورافتاده روش های استوار سازی استفاده می شود. در رابطه با مطالعات

- 
1. Generalized Estimating Equations
  2. Liang, K.Y., Zeger, S.L.
  3. Prentice, R. L.
  4. Zhao, I., Prentice, R.L.
  5. Liang, K.Y., Zeger S.L., Qaqish, B.
  6. Zeger, S.L., Liang, K.

7. Pan, W.

8. Wang, M., Long, Q.

9. Carroll, R.J. and Pederson S.

10. Qaqish, B.F., Preisser, J.S.

فرض کنید در یک مطالعه طولی برای هر موضوع یا فرد مورد مطالعه  $n_i$  زمان اندازه گیری وجود داشته باشد بنابراین فرد  $i$ ام ( $i = 1, 2, \dots, k$ ) در موقعیت های  $t = 1, 2, \dots, n_i$  مشاهده می شود. متغیر پاسخ مربوط به فرد  $i$ ام در زمان  $t$  را می توان به صورت  $y_{it}$  نمایش داد که می تواند پیوسته، یا گسسته باشد.

$$Y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{pmatrix} = (y_{i1}, y_{i2}, \dots, y_{in_i})'$$

هر پاسخ  $y_{it}$  یک بردار  $1 \times p$  از متغیرهای کمکی دارد که به صورت زیر نمادگذاری می شود.

$$X_{it} = \begin{pmatrix} \chi_{it1} \\ \chi_{it2} \\ \vdots \\ \chi_{itp} \end{pmatrix} \quad \begin{matrix} i = 1, 2, \dots, K \\ t = 1, 2, \dots, n_i \end{matrix}$$

به عبارت دیگر برای فرد  $i$ ام ماتریسی  $n_i \times p$  از متغیرهای کمکی به صورت

$$X_i = \begin{pmatrix} x'_{i1} \\ x'_{i2} \\ \vdots \\ x'_{in_i} \end{pmatrix} = \begin{pmatrix} \chi_{i11} & \chi_{i12} & \dots & \chi_{i1p} \\ \chi_{i21} & \chi_{i22} & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \chi_{in_i1} & \chi_{in_i2} & \dots & \chi_{in_ip} \end{pmatrix}_{n_i \times p}$$

مشاهده می شود. مجموعه ای از مشاهدات همبسته مربوط به فرد  $i$ ام را خوشه نامیده و فرض می شود که مشاهدات درون خوشه ها همبسته اند، ولی بین یک خوشه و خوشه دیگر همبستگی وجود ندارد. مدل های بررسی شده به روش (GEE) مدل هایی با میانگین حاشیه ای<sup>2</sup> نامیده می شوند. در یک مدل حاشیه ای برای مطالعات طولی امید حاشیه ای متغیر پاسخ یعنی  $E(y_{it}) = \mu_{it}$  از طریق یک تابع پیوند به نام  $g$  به  $x_{it}$  به صورت زیر مربوط می شود:

برآوردگر تعمیم یافته استوار را مطرح کرده و سپس با استفاده از شبیه سازی روش معادلات برآوردگر تعمیم یافته استوار (REGEE)<sup>1</sup> با روش معمول آن (GEE) مقایسه می شود.

## 2. معادلات برآوردگر تعمیم یافته

در دو دهه اخیر توجه بسیاری از محققان به تجزیه و تحلیل داده های چندمتغیره و همبسته جلب شده است. در این گونه مطالعات متغیر پاسخ برای هر فرد در چندین نوبت متوالی مشاهده می شوند. به مطالعه ای که اندازه گیری مربوط به یک صفت در طول زمان مورد بررسی قرار می گیرد، مطالعه طولی می گویند. این گونه از اندازه گیری ها معمولاً روی فرد و یا واحد اندازه گیری در زمان های مختلف در طول مطالعه انجام می شوند. به فرد و یا واحد نمونه گیری که برای وی در طول زمان چند اندازه گیری انجام شده است، یک خوشه گفته می شود. بدیهی است مشاهدات حاصل از هر فرد یا خوشه که در واقع اندازه های تکراری را تشکیل می دهند، با یکدیگر همبسته بوده، روش های معمول برای تحلیل داده های مستقل در مورد آن ها کارایی لازم را ندارند.

از آنجا که روش معادلات برآوردگر تعمیم یافته روشی برای برآورد ضرایب رگرسیونی در مدل های خطی تعمیم یافته می باشد، برای متغیر پاسخ  $y_i, i = 1, 2, \dots, n$  مدل خطی تعمیم یافته به شرح زیر است:

$$g(\mu_i) = g[E(y_i)] = x'_i \beta$$

که در آن بردار متغیرهای کمکی برای مشاهده  $i$ ام و  $\beta$  بردار پارامترها (ضرایب رگرسیون) و  $g$  تابع پیوند (ربط) می باشند. بردار پاسخ فرد  $i$ ام به صورت زیر بیان می شود.

در این رابطه  $\eta$  بردار میانگین پاسخ با عناصر  $\eta$  است و داریم:

$$\mu_{it} = \mu_{it}(\beta) = g^{-1}(X'_{it}\beta)$$

برای برآورد پارامترهای  $\beta$  روش مینیمم کردن تابع حداقل مربعات تعمیم یافته منجر به معادلات برآوردگر تعمیم یافته به صورت زیر شده که با حل آن می توان  $\beta$  را برآورد کرد:

$$\sum_{i=1}^N \left( \frac{\partial \mu_i}{\partial \beta} \right)' V_i^{-1} (y_i - \mu_i(\beta)) = 0$$

به طور معادل داریم

$$\sum_{i=1}^k D'_i (A_i R_i(\alpha) A_i)^{-1} (y_i - \mu_i) = 0 \quad (1)$$

### 2-1 ساختارهای مختلف ماتریس همبستگی کاری

برای برآورد پارامتر همبستگی کاری  $\alpha$  و تعیین ماتریس همبستگی کاری  $R_i(\alpha)$  لیانگ و زیگر (1986) روش های زیر را پیشنهاد داده اند.

#### 2-1-1 ساختار همبستگی تبادل پذیر<sup>3</sup>

اگر مشاهدات برای هر فرد دارای همبستگی مشترک باشند، یعنی برای تمام زوج های  $y_{it}$  و  $y_{is}$  همبستگی ثابت وجود داشته باشد، در این صورت ساختار همبستگی تبادل پذیر داریم. پارامتر همبستگی  $\alpha$  اسکالر بوده و ماتریس همبستگی کاری برای ساختار همبستگی تبادل پذیر به صورت

$$\text{corr}(y_{it}, y_{is}) = \begin{cases} 1 & t = s \\ \alpha & t \neq s \end{cases}$$

$$R(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha & \dots & \alpha \\ \alpha & 1 & \alpha & \dots & \alpha \\ \alpha & \alpha & 1 & \dots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \dots & 1 \end{bmatrix} \quad (2)$$

$\eta_{it} = g(\mu_{it}) = x'_{it}\beta$   
بردار پارامترهای نامعلوم از مرتبه  $1 \times p$  بوده

و فرض می شود که واریانس متغیر پاسخ تابعی از میانگین  $\mu_{it}$  بوده و به صورت زیر بیان می شود:

$$\text{var}(y_{it}) = f(\mu_{it})\phi$$

به عنوان مثال اگر متغیر پاسخ دارای توزیع برنولی باشد داریم:

$$f(\mu_{it}) = \mu_{it}(1 - \mu_{it})$$

$\phi$  پارامتر پراکندگی است و اگر متغیر پاسخ دارای توزیع برنولی یا دوجمله ای باشد  $\phi = 1$  در نظر گرفته می شود (پارک و همکاران<sup>1</sup> (1998)).  
ماتریس  $V_i$ ، ماتریس کوواریانس  $Y_i$  می باشد. این ماتریس را می توان براساس ماتریس همبستگی عناصر  $Y_i$  که آن را  $R_i(\alpha)$  می نامیم به صورت زیر بیان کرد:

$$V_i = V_i(\beta, \alpha, \phi) = A_i(\beta) R_i(\alpha) A_i(\beta) \phi$$

که  $A_i(\beta)$  یک ماتریس قطری  $n_i \times n_i$  بوده و عناصر روی قطر اصلی آن به صورت  $f^{\frac{1}{2}}(\mu_{it})$  می باشد یعنی:

$$A_i = \text{diag} \left( \text{var}^{\frac{1}{2}}(y_{it}) \right) = \text{diag} \left( f^{\frac{1}{2}}(\mu_{it}) \right)$$

از آنجا که همبستگی بین عناصر  $y_i$  به درستی معلوم نیست،  $R_i(\alpha)$  را ماتریس همبستگی کاری<sup>2</sup> می نامند، و باید با برآورد پارامتر  $\alpha$ ، ماتریس همبستگی کاری  $R_i(\alpha)$  را معلوم کرد. روش های مختلفی برای تعیین  $R_i(\alpha)$  با توجه به نوع مدل به کار می رود. در بخش 2-1 روش های تعیین  $R_i(\alpha)$  را بیان کرده و در ادامه برای برآورد  $\beta$  تابع حداقل مربعات تعمیم یافته را به صورت زیر تشکیل می دهیم:

$$\sum_{i=1}^N (y_i - \mu_i(\beta))' V_i^{-1} (y_i - \mu_i(\beta))$$

1. Park, T.P., Shin, D.W., Park, C.G.

2. Working Correlation Matrix

3. Exchangeable

$$\bar{e}_{i2} = \frac{1}{n_i - 1} \sum_{j=2}^{n_i} \hat{e}_{ij}$$

$$\bar{e}_{i1} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i-1} \hat{e}_{ij}$$

برآورد می شود. ماتریس همبستگی کاری به صورت

$$R(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{T-1} \\ \alpha & 1 & \alpha & \dots & \alpha^{T-2} \\ \alpha^2 & \alpha & 1 & \dots & \alpha^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{T-1} & \alpha^{T-2} & \alpha^{T-3} & \dots & 1 \end{bmatrix} \quad (5)$$

خواهد بود.

### 4-1-2 روش بی ساختار<sup>3</sup>

در این حالت، الگوی مشخصی برای توصیف همبستگی بین مشاهدات وجود ندارد و برآورد  $\hat{\alpha}$  در این روش به صورت

$$\hat{\alpha}_{ts} = \frac{1}{(K-P)\hat{\phi}} \sum_{i=1}^k \hat{e}_{it} \hat{e}_{is} \quad (6)$$

$$\text{corr}(y_{it}, y_{is}) = \begin{cases} 1 & t = s \\ \alpha_{ts} & t \neq s \end{cases}$$

است.

### 3- معادلات برآوردگر تعمیم یافته به روش کمترین مربعات دوباره وزنی شده تکراری<sup>4</sup> (IRLS)

جهت استوار سازی معادلات برآوردگر تعمیم یافته، ناگزیر به بیان روش IRLS برای این خانواده از معادلات می باشیم. جهت این کار فرض می کنیم  $N = \sum n_i$  و  $Y = (y'_1, y'_2, \dots, y'_k)'$  بردار  $N \times 1$  و  $X = (x'_1, x'_2, \dots, x'_k)'$  ماتریس  $n_i \times p$  پر رتبه ستونی از متغیرهای کمکی باشد. مدل کلی  $\eta = g(\mu) = X\beta$  را که در آن

است. روش GEE با ساختار همبستگی تبادل پذیر به منظور برآورد پارامتر همبستگی مشترک  $(\alpha)$ ، از مانده های پیرسن برآورد شده توسط رابطه

$$\hat{e}_{it} = (y_{it} - \hat{\mu}_{it}) / \sqrt{V(\hat{\mu}_{it})}$$

برآورد  $\alpha$  به دست آمده از این مانده ها به صورت

$$\hat{\alpha} = \frac{1}{(N^* - P)\hat{\phi}} \sum_{i=1}^k \sum_{t \neq s} \hat{e}_{it} \hat{e}_{is} \quad (3)$$

$$N^* = \sum_{i=1}^k n_i(n_i - 1)$$

می باشد.

### 2-1-2 روش 1- همبستگی<sup>1</sup>

در این حالت هر مشاهده در زمان  $t$  فقط با مشاهده در زمان بعدی یعنی  $t+1$  همبستگی داشته و سایر همبستگی ها را صفر در نظر می گیریم. به عبارت دیگر

$$\text{corr}(y_{it}, y_{it+1}) = \begin{cases} 1 & t = 0 \\ \alpha & t = 1 \\ 0 & t > 1 \end{cases}$$

در این صورت برآورد پارامتر همبستگی از رابطه

$$\hat{\alpha} = \frac{\sum_{i=1}^k \hat{e}_{it} \hat{e}_{it+1}}{(k-P)\hat{\phi}} \quad (4)$$

به دست می آید.

### 3-1-2 همبستگی اتورگرسیو<sup>2</sup>

اگر مشاهدات مکرر درون خوشه به مشاهده قبل از خود وابسته باشند از همبستگی اتورگرسیو برای برآورد پارامترها استفاده می شود.

$$\text{corr}(y_{it}, y_{is}) = \alpha^{|t-s|}$$

پارامتر همبستگی  $\alpha$  با استفاده از مانده های پیرسن  $\hat{e}_{it}$  به صورت

$$\hat{\alpha} = \frac{1}{\hat{\phi}} \frac{\sum_{i=1}^k \left\{ \sum_{j=2}^{n_i} (\hat{e}_{ij} - \bar{e}_{i2})(\hat{e}_{i,j-1} - \bar{e}_{i1}) \right\}}{\sum_{i=1}^k \left\{ \sum_{j=2}^{n_i} (\hat{e}_{i,j-1} - \bar{e}_{i1})^2 \right\}}$$

3. Unstructured  
4. Iteratively reweighted least squares

1. 1-dependence  
2. Auto regressive

نشان داده خواهد شد و  $E = D^*(y - \hat{\mu}) = Z^*$  و  $\eta = (I - H)Z^*$  بردار مانده‌ها خواهد بود.

#### 4- معادلات برآوردگر تعمیم‌یافته‌ی استوار

معادلات برآوردگر تعمیم‌یافته استوار توسط پرایسر و کاکیش<sup>1</sup> (1996) و (1999) به صورت زیر تعریف شده است:

$$\sum_{i=1}^k D'_i(X_i, \beta) V_i^{-1}(\alpha, \beta) [W_i(X_i, X, Y_i, \alpha, \beta) (Y_i, \mu_i(\beta)) - C_i] = 0 \quad (8)$$

در این رابطه اگر  $W_i = I$  و  $C_i = 0$  در نظر گرفته شوند معادلات برآوردگر تعمیم‌یافته مطرح شده توسط لیانگ و زیگر (1992) معرفی شده در رابطه (1) به دست می‌آید.

$W_i$  یک ماتریس قطری برای  $i$  امین خوشه شامل وزن‌های  $w_{it}$ ،  $t = 1, 2, \dots, n_i$  می‌باشد. همواره  $w_{it}$ ها بین 0 و 1 بوده و برای بیشتر مشاهدات وزنی نزدیک به 1 در نظر گرفته می‌شود اما برای مشاهداتی که تأثیر بیشتری روی برآورد  $\beta$  دارند وزن پایین‌تری در نظر گرفته می‌شود.

کمیت  $\psi_i = w_i(y_i - \mu_i)$  که به عامل ناریب‌سازی<sup>2</sup> نامیده می‌شود، باید به‌گونه‌ای محاسبه شود که معادلات برآوردگر تعمیم‌یافته در رابطه (8) ناریب شود. پائین آوردن وزن مشاهدات پرنفوذ و دورافتاده ممکن است براساس نفوذ متغیرهای کمکی باشد یعنی  $W_{it} = W_{it}(h_{it})$ ، که در کلاس مالوس این‌گونه است، و یا ممکن است علاوه بر متغیرهای کمکی به متغیر پاسخ هم وابسته باشد یعنی  $W_{it} = W_{it}(x_{it}, X, y_{it}, \alpha, \beta)$ ، که کلاس شوئیپ بدین صورت می‌باشد. در کلاس مالوس چون وزن‌های  $W_{it}$  غیر تصادفی‌اند  $C_i = 0$  می‌باشد. و در

$\mu' = [\mu_1, \mu_2, \dots, \mu_k]$  می‌باشد در نظر گرفته و با توجه به بسط مرتبه اول سری تیلور حول نقطه  $\hat{\mu}$  به رابطه زیر می‌رسیم:

$$g(\mu) \cong g(\hat{\mu}) + g'(\hat{\mu})(y - \hat{\mu})$$

اگر  $Z^* = g(\hat{\mu}) + g'(\hat{\mu})(y - \hat{\mu})$  در نظر

گرفته شود و  $Z^*$  را بردار پاسخ کاری بنامیم در حالت ماتریسی رابطه زیر را خواهیم داشت

$$(Z^* = X\hat{\beta} + D^*(y - \mu)) \quad (7)$$

در این رابطه  $D^* = \partial\eta/\partial\mu$  یک ماتریس قطری بلوکی  $N \times N$  است که بلوک  $i$ ام آن با  $D_i^*$  نشان داده می‌شود و  $D_i^* = \text{diag}\{(\partial\eta_{it})/(\partial\mu_{it})\}$  می‌باشد. برآورد  $\beta$  را می‌توان به روش IRLS به دست آورد، بدین صورت که یک برآورد اولیه  $\hat{\beta}$  که به روش معروف کمترین توان دوم خطا به دست آمده است را در رابطه (7) قرار داده و  $Z^*$  را به دست می‌آوریم.  $V^{-1} = W^*$  بوده و بلوک  $i$ ام آن مربوط به خوشه  $i$ ام می‌باشد. با استفاده از رابطه  $W_i^* = D_i^{*-1} A_i^{-1} R_i^{-1}(\hat{\alpha}) A_i^{-1} D_i^{*-1}$  و پس از برآورد  $\hat{\alpha}$  و تعیین  $R_i(\alpha)$  به روش‌های بیان شده در بخش (1-2)، می‌توان  $W_i^*$  و برآورد جدید  $W^*$  را به دست آورد، سپس  $\hat{\beta}_{new}$  را با استفاده از رابطه  $\hat{\beta}_{new} = (X'W^*X)^{-1}X'W^*Z^*$  محاسبه کرده و مراحل بالا را تا جایی که اختلاف  $\hat{\beta}$ ها کم و به تقریب قابل قبولی برسد تکرار می‌کنیم. حال رابطه زیر را در نظر می‌گیریم و در آن ماتریس تصویر  $H$  را تعریف می‌کنیم.

$$\begin{aligned} \hat{\eta} &= X\hat{\beta}_{new} \\ &= X(X'W^*X)^{-1}X'W^*Z^* \\ &= HZ^* \end{aligned}$$

عناصر قطر اصلی  $H$  که با  $h_{it}$  نشان داده می‌شوند متناسب با خاصیت اهرمی مشاهدات و تأثیر آنها روی مقادیر برازش داده شده هستند. اگر  $p = \text{tr}(H)$  در نظر گرفته شود متوسط  $h_{it}$  با  $\frac{p}{N}$

1. Preisser and Qaqish.

2. Debiasing factor

مدل  $Z^* = X\hat{\beta} + D^*(\psi - C)$  به دست آمده واریانس  $\hat{\beta}_R$  توسط رابطه زیر به دست خواهد آمد

$$\left( \sum_{i=1}^k D_i' V_i^{-1} \Gamma_i D_i \right)^{-1} \left\{ \sum_{i=1}^k D_i' V_i^{-1} (\psi_i - C_i) (\psi_i - C_i)' V_i^{-1} D_i \right\} \left( \sum_{i=1}^k D_i' V_i^{-1} \Gamma_i D_i \right)^{-T}$$

برای اثبات و توضیحات بیشتر به پرایسر و کاکیش ([11],[12]) مراجعه شود.

#### 4-1 معادلات برآوردگر تعمیم یافته استوار برای

مدلهایی با پاسخ دودویی همبسته

روش "کاهش وزن مشاهده" در کلاسهای شوئیپ و مالوس در پاسخهای دودویی همبسته به آسانی قابل استفاده است، زیرا توزیعهای حاشیه‌ای و توزیعهای دومتغیره  $y_i$  فقط به  $\alpha$  و  $\beta$  مربوط اند. یک خوشه به اندازه دلخواه را در نظر می‌گیریم و بدون از دست دادن کلیت مسئله فرض می‌کنیم اولین دو عضو بردار پاسخ  $y_1$  و  $y_2$  باشند.  $\pi_{jk}$  را به صورت  $\pi_{jk} = p_r(y_1 = j, y_2 = k)$  تعریف می‌کنیم. توزیع دومتغیره یک توزیع چندجمله‌ای با عناصر احتمال  $(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$  بوده و توسط میانگینهای حاشیه‌ای  $\mu_1 = p_r(y_1 = 1)$  و  $\mu_2 = p_r(y_2 = 1)$  معین می‌شوند و اگر همبستگی بین  $y_1$  و  $y_2$  را با  $\rho$  نشان دهیم داریم:

$$\pi_{10} = \mu_1 - \pi_{11} \quad \text{و} \quad \pi_{11} = \mu_1 \mu_2 + \rho \sqrt{V_1^2 V_2^2}$$

$$\pi_{00} = 1 - \mu_1 - \mu_2 + \pi_{11} \quad \text{و} \quad \pi_{01} = \mu_2 - \pi_{11}$$

تابع واریانس توسط رابطه‌ای  $V_t = V(\mu_t)$  تابع  $\mu_t(1 - \mu_t)$  تعریف شده و  $\phi = 1$  در نظر گرفته می‌شود. فرض کنیم تابع  $W_t$  وزن مشاهدات و تابعی از مانده  $r_t$  باشد. برای حل معادلات برآوردگر تعمیم یافته استوار در روش «کاهش وزن مشاهده» کلاس شوئیپ، نیاز به تعیین  $C_t$  دارد که می‌تواند از توزیع برنولی حاشیه‌ای به دست آورد که به صورت

کلاس شوئیپ باید  $C_i$  را طوری تعیین کرد که معادلات برآوردگر رابطه (8) نااریب باشند. پرایسر و کاکیش<sup>1</sup> (1996) و (1999) قضیه زیر را در رابطه با معادلات برآوردگر استوار رابطه (8) مطرح کردند.

#### قضیه

اگر  $\psi_i = w_i(y_i - \mu_i)$  را تعریف کرده و  $C_i = E[\Psi_i]$  در نظر گرفته و فرض کنیم.

(1)  $\hat{\alpha}$  سازگاری از نوع  $\sqrt{k}$  به شرط وجود  $\beta$  و  $\phi$  دارد.

(2)  $\hat{\phi}$  سازگاری از نوع  $\sqrt{k}$  به شرط وجود  $\beta$  دارد.

$$\text{var}(\psi_i) < \infty \quad (3)$$

(4)  $\Psi_i$  در  $\beta$  مطلقاً پیوسته بوده و مشتق آن نسبت به  $\mu_i$  با  $\Psi_i$  نشان داده می‌شود. و برای هر  $y_i$

$$E\|\psi_i\| < \infty$$

آنگاه تحت شرایط مشخصی  $(\hat{\beta}_R - \beta)$  به  $k^{\frac{1}{2}}$  طور مجانبی به توزیع نرمال با میانگین صفر و ماتریس کوواریانس  $V_R$  که از رابطه زیر به دست می‌آید، میل می‌کند.

$$\lim_{k \rightarrow \infty} \left( \sum_{i=1}^k D_i' V_i^{-1} \Gamma_i D_i \right)^{-1} \left\{ \sum_{i=1}^k D_i' V_i^{-1} \text{Var}(\psi_i) V_i^{-1} D_i \right\} \quad (9)$$

$$\left( \sum_{i=1}^k D_i' V_i^{-1} \Gamma_i D_i \right)^{-T} \hat{C}_i = \frac{\partial}{\partial \mu_i} C_i$$

$$\hat{\psi}_{ki} = \frac{\partial}{\partial \mu_i} \psi_i(\mu_i)$$

$$\Gamma_i = E \hat{\psi}_{ki} - \hat{C}_i$$

شرایط اضافی لازم این است که مشتقهای  $(\partial \hat{\alpha}(\beta, \phi)) / \partial \beta$  و  $(\partial \hat{\alpha}(\beta, \phi)) / \partial \phi$  و  $(\partial \hat{\phi}(\beta)) / \partial \beta$  با تغییر کران‌دار باشند.

برآورد  $\beta$  در معادلات برآوردگر (8) با روش کمترین مربعات دوباره وزنی شده تکراری توسط

در این مدل  $\beta_0 = -2$  و  $\beta_1 = 0.8$  و همبستگی بین مشاهدات پاسخ را  $\rho = 0.3$  یا  $\rho = 0.7$  در نظر گرفته و با استفاده از ساختار همبستگی تبادلی پذیر (رابطه (2)) شبیه‌سازی داده‌ها را 1000 بار تکرار می‌کنیم. در این راستا جهت مقایسه روش‌های استوار (REGEE) با روش کلاسیک (GEE)، داده‌هایی در حالت‌های بدون آلودگی (0%) درصد آلودگی، 3% و 5% آلودگی برای دو کلاس مالوس و شوپ تولید می‌کنیم. سپس جهت مقایسه روش‌های مذکور از میانگین مربعات خطا استفاده می‌شود. تابع وزن دهی به کار رفته در تمام روش‌ها  $W(v) = \exp\{-(v/a)^2\}$  بوده، که توسط هولاند و ولش<sup>1</sup> (1997) پیشنهاد شده است. در روش کاهش وزن مشاهده شوپ  $v = r_{it}$  و وزن تابعی از مانده پیرسن خواهد بود. در روش کاهش وزن مشاهده‌ی مالوس  $v = h_{it}$  و وزن دهی بر اساس خاصیت اهرمی هر مشاهده انجام می‌شود.  $a$  مقدار ثابتی است که ثابت میزان‌سازی<sup>2</sup> نامیده شده و باید آن را به گونه‌ای انتخاب نمود که بیشترین کارایی برای برآوردگر حاصل شود. عموماً انتخاب مقادیر بزرگ برای  $a$  نقاط بالقوه بانفوذ را کم‌وزن نمی‌کند. در روش کاهش وزن مشاهده مالوس پیشنهاد می‌شود که  $a$  3 یا 4 برابر  $\frac{P}{N}$  انتخاب شود. در روش کاهش وزن مشاهده شوپ با بررسی کارایی مدل با توجه به مقادیر متفاوت  $a$ ، مقادیر  $a = 3$  تا  $a = 5$  برای این کار مناسب تشخیص داده شده‌اند.

1-5 مقایسه مقدار اریبی روش‌های معادلات برآوردگر تعمیم‌یافته استوار با روش GEE در برآورد  $\beta_1$

ابتدا میزان اریبی روش‌های REGEE را در برآورد  $\beta_1$  نسبت به روش GEE بررسی می‌کنیم.

اینجا  $C_t = V_t(w_t^{(1)} - w_t^{(0)})$  وزن مشاهده  $t$ ام در خوشه مورد نظر زمانی که  $y_t = j$  است می‌باشد، چون هر  $w_t$  تابعی از مانده‌های مربوط به مشاهده  $t$ ام است و به بردار کامل مانده‌های خوشه مربوط نمی‌شود، چنین بر می‌آید که  $\Psi$  و همچنین  $\Gamma$  ماتریس‌های قطری خواهند بود. لذا می‌توان نشان داد که:

$$\text{cov}(\psi_t, \psi_{t'}) = \rho_{it'} V_t^{\frac{1}{2}} V_{t'}^{\frac{1}{2}} b_t b_{t'}$$

و

$$\text{var}(\psi_t) = V_t b_t^2$$

و

$$\Gamma = \text{Diag}\{-b_t\}$$

که در آن  $b_t = (1 - \mu_t)w_t^{(1)} + \mu_t w_t^{(0)}$  و

$\rho_{tt'}$  همبستگی بین  $y_t$  و  $y_{t'}$  خواهد بود.

5- مقایسه معادلات برآوردگر تعمیم‌یافته با معادلات برآوردگر تعمیم‌یافته استوار با استفاده از روش شبیه‌سازی

به منظور مقایسه روش معادلات برآوردگر تعمیم‌یافته با روش معادلات برآوردگر تعمیم‌یافته استوار از شبیه‌سازی استفاده کرده و مدلی با یک متغیر کمکی و متغیر پاسخ دودویی که در آن تعداد اعضای درون خوشه‌ها مساوی و برابر 2 باشد را در نظر می‌گیریم. اگر ماتریس طرح به صورت زیر بیان شود:  $k$  تعداد خوشه‌ها و  $i$  تعداد آنها را نشان می‌دهد.

$$\begin{pmatrix} 1 & \Delta_i \\ 1 & -\Delta_i \end{pmatrix}$$

$$\Delta_i = \frac{i}{k}$$

$$i = 1, 2, \dots, k$$

با توجه به توزیع متغیر پاسخ مدلی با یک متغیر کمکی و با پیوند لوجیت به صورت زیر خواهیم داشت:

$$\begin{aligned} \text{logit}(\pi_{it}) &= \beta_0 + \beta_1 x_{it} \\ i &= 1, 2, \dots, k \\ t &= 1, 2 \end{aligned}$$

1. Holland and Welsch

2. Tuning Constant



درصد آلودگی به نقاط دورافتاده این میزان بین 40 تا 80 درصد کمتر از روش GEE می باشد. همان طور که ملاحظه می شود زمانی که  $\rho = 0.7$  در نظر گرفته شود، میزان اریبی هر دو روش بیشتر از زمانی است که  $\rho = 0.3$  باشد .

نتایج این بررسی ها در جداول (1) و (2) بیان شده است. در این جداول اثر تعداد خوشه ها، همبستگی بین مشاهدات، میزان آلودگی داده ها به نقاط دورافتاده را روی برآورد  $\beta_1$  بررسی کرده ایم که نتایج حاصل از آن به صورت زیر است:

جدول 1. مقایسه میزان اریبی روش کاهش وزن مشاهده شوئیپ با  $\alpha = 3$  (GEE)

		$\rho = 0/3$		$\rho = 0/7$	
		GEE	شوئیپ	GEE	شوئیپ
درصد نقاط دورافتاده	تعداد خوشه				
0%	k=200	0/011	0.044	-0.112	-0.020
0%	k=100	0.020	0.094	-0.092	0.060
0%	k=50	0.073	0.339	-0.056	0.290
3%	k=50	-0.109	0.047	-0.209	-0.035
5%	k=50	-0.194	-0.050	-0.279	-0.114
10%	k=50	-0.343	-0.309	-0.403	-0.368

جدول (2) نشان می دهد که زمانی که بین مشاهدات، نقاط دورافتاده وجود ندارد تفاوتی بین روش کاهش وزن مشاهده مالوس و روش GEE در برآورد  $\beta_1$  وجود ندارد. وجود نقاط دورافتاده به میزان کمی اریبی روش مالوس را نسبت به روش GEE کم می کند ضمناً تغییر همبستگی نیز در نتایج

جدول (1) نشان می دهد که روش کاهش وزن مشاهده شوئیپ در برآورد  $\beta_1$ ، زمانی که بین داده ها نقاط دورافتاده و پرنفوذ وجود ندارد اریبی بیشتری از روش GEE دارد اما زمانی که در مشاهدات نقاط دورافتاده وجود داشته باشد به مقدار زیادی اریبی آن از روش GEE کمتر است. می توان گفت در یک مدل

جدول 2. مقایسه میزان اریبی روش کاهش وزن مشاهده مالوس با روش  $\alpha = \frac{4P}{N}$  (GEE)

		$\rho = 0.3$		$\rho = 0.7$	
		GEE	مالوس	GEE	مالوس
درصد نقاط دورافتاده	تعداد خوشه				
0%	k=200	0.011	0.011	-0.112	-0.112
0%	k=100	0.020	0.021	-0.092	-0.091
0%	k=50	0.073	0.078	-0.056	-0.050
3%	k=50	-0.109	-0.106	-0.209	-0.207
5%	k=50	-0.194	-0.192	-0.279	-0.278
10%	k=50	-0.343	-0.342	-0.403	-0.402

بیان شده تغییری ایجاد نکرده است. اما زمانی که  $\rho = 0.3$  به  $\rho = 0.7$  تغییر می کند میزان اریبی هر دو روش بیشتر از حالت قبل است.

با تعداد خوشه  $k=50$  و 5% آلودگی به نقاط دورافتاده اریبی روش های REGEE بین 15 تا 30 درصد کمتر از روش (GEE) است و برای مدلی با  $k=50$  و 3

## 2-5 مقایسه میانگین مربعات خطا در برآورد $\beta_1$ در معادلات برآوردگر تعمیم یافته و روش های استوار آن

در این بخش می خواهیم با استفاده از میانگین مربعات خطا روش معادلات برآوردگر تعمیم یافته را با روش های استوار آن مقایسه کنیم. همچنین می خواهیم با تغییر در اندازه خوشه ها به این سؤال پاسخ دهیم که افزایش تعداد خوشه ها چه تاثیری در میانگین مربعات خطا دارد؟ جدول (3) نتایج حاصل از این بررسی را نشان می دهد.

می شود. و در روش کاهش وزن مشاهده شوئیپ همواره این میزان هنگامی که نقاط دورافتاده در داده ها وجود داشته باشد کمتر از روش GEE است. به طور کلی از نتایج جدول (3) داریم که میزان کارایی روش شوئیپ نسبت به روش مالوس با افزایش میزان آلودگی از روندی صعودی برخوردار می باشد. بررسی مشابه نشان داده است که با افزایش عناصر ماتریس طرح میزان کارایی روش استوار به مقدار قابل ملاحظه ای افزایش می یابد (پرایسر و کاکیش، 1999).

جدول 3. میانگین مربعات خطا برای روش های GEE و REGEE

GEE		شوئیپ		مالوس	
MSE	برآورد $b_1$	MSE	برآورد $b_1$	MSE	برآورد $b_1$
% نقاط دورافتاده	تعداد خوشه ها				
0%	K=100	0.820	0.116	0.894	0.266
0%	K=200	0.811	0.056	0.818	0.060
0%	K=400	0.809	0.027	0.808	0.028
0%	K=600	0.807	0.018	0.818	0.019
5%	K=100	0.593	0.137	0.598	0.138
5%	K=200	0.578	0.095	0.581	0.093
5%	K=400	0.572	0.074	0.574	0.073
5%	K=600	0.5734	0.066	0.575	0.065

### نتیجه گیری

با توجه به نتایج شبیه سازی از مقایسه روش های استوار سازی معادلات برآوردگر تعمیم یافته در دو کلاس شوئیپ و مالوس به نظر می رسد که:

**الف:** روش کاهش وزن مشاهده شوئیپ در برآورد  $\beta_1$ ، زمانی که بین داده ها نقاط دورافتاده و پرنفوذ وجود ندارد اریبی بیشتری از روش GEE دارد اما زمانی که در مشاهدات نقاط دورافتاده وجود داشته باشد به مقدار زیادی اریبی آن از روش GEE کمتر است.

**ب.** میانگین مربعات خطا در روش کاهش وزن مشاهده شوئیپ زمانی که در داده ها نقاط دورافتاده

جدول (3) نشان می دهد که میانگین مربعات خطا در روش کاهش وزن مشاهده شوئیپ زمانی که در داده ها نقاط دورافتاده وجود ندارند بیشتر از روش GEE است اما زمانی که به میزان 5% نقاط دورافتاده در داده ها وجود داشته باشد میانگین مربعات خطای این روش کمتر از روش GEE می شود. در روش کاهش وزن مشاهده مالوس چه در زمانی که در داده ها نقاط دورافتاده وجود ندارد و چه زمانی که نقاط دورافتاده به میزان 5% وجود دارد تفاوت چندانی بین میانگین مربعات خطای این روش و GEE در برآورد  $\beta_1$  وجود ندارد. در هر 2 روش با افزایش تعداد خوشه ها از میانگین مربعات خطا کم

روش استوار دیده نمی شود، ولی در حالت کلی کلاس شوئیپ با توجه به میزان کاهش اریبی آن در برآورد  $\beta_1$  و نیز از منظر میانگین مربعات خطا مناسب تر می باشد.

#### References

- Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biomtrika*; 73: 13-22.
- Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observations, 44: 1033-1048.
- Zhao, I. and Prentice, R.L. (1990). Correlated binary regression analysis quadratic exponential model, *Biometrika*, 77: 642-648.
- Liang, K.Y., Zeger, S.L. and Qaqish, B. (1992). Multivariate regression analysis for categorical data. *Journal of statistical society, Series B*, 54: 3-40.
- Zeger, S.L. and Liang, K.Y. (1992). An overview of methods for the analysis of Longitudinal data. *Statistics in medicine*, 11: 1825-1839.
- Pan W. (2001). On the robust variance estimator in generalised estimating equations. *Biometrika*; 88(3):901-906. DOI:10.1093/biomet/88.3.901.
- Wang, M. and Long, Q. (2010). Modified robust variance estimator for generalized estimating equations with improved

وجود ندارند بیشتر از روش GEE است اما زمانی که به میزان 5% نقاط دورافتاده در داده ها وجود داشته باشد میانگین مربعات خطای این روش کمتر از روش GEE می شود. از طرف دیگر با توجه به معیارهای اریبی و MSE تفاوت چندانی بین دو small-sample performance. *Statistics in medicine*, 30 1278-1291.

Carroll, R.J. and Pederson, S. (1993). On robustness in the logistic regression model. *J.R statist. Soc, B*, 55 693-706.

Qaqish, B.F. and Preisser, J.S. (1999) Resistant fits for regression with correlated outcomes an estimating equations approach, *Journal of Statistical Planning and Inference*, 75, 415-431.

Park, T.P., Shin, D.W. and Park, C.G. (1998). A generalized estimating equations approach for order group effects with repeated measurements, *Biometrics*, 83, 688-694.

Preisser, J.S. and Qaqish, B.F. (1996). Deletion diagnostics for Generalised Estimating Equations. *Biometrika*, 83, 551-562.

Preisser, J.S. and Qaqish, B.F. (1999). Robust regression for clustered data with application to binary responses. *Biometrika* 55, 574-579.

Holland, P.W. and Welsch, R.E. (1977). Robust Regression Using Iteratively Reweighted Least-Squares. *Communications in Statistics – Theory and Methods* 6(9), 813-827.