# روش معادلات برآوردگر تعمیم‌یافته استوار و کاربرد آن در مدل‌هایی با برآمدهای دودویی همبسته

**مسعود یارمحمدی[1*]؛ سعید معدنی[2]**

**دریافت**: 1394/02/13
**پذیرش**: 1394/04/20

**چکیده**

روش معادلات برآوردگر تعمیم‌یافته 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 تعمیم‌یافته برای برآوردهایی کارا و نااریب پارامترهای رگرسیونی نسبت به برآورد ضرایب رگرسیونی در مدل‌های خطی تعمیم‌یافته، هنگامی‌که 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 تعمیم‌یافته 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 تعمیم‌یافته، معادلات برآوردگر تعمیم‌یافته استوار، کلاس شوئیپ، کلاس مالوس.

---

1. دانشیار، گروه آمار دانشگاه پیام نور (***نویسنده مسئول**) masyar@pnu.ac.ir
2. دانشجوی دکترا، آمار دانشگاه پیام نور s_madani2000@yahoo.com

## 1. مقدمه

روش معادلات برآوردگر تعمیم‌یافته (GEE)[1] امروزه به طور وسیعی در بیولوژی و پزشکی و صنایعی که در تحلیل داده‌های آنها همبستگی بین مشاهدات جمع‌آوری شده دیده می‌شود، به کار می‌رود. بیشترین کاربرد این روش در اندازه‌گیری‌های مکرر یا مطالعات طولی است، که در آنها یک فرد یا یک موضوع در موقعیت‌های زمانی یا مکانی به صورت مکرر مورد بررسی قرار می‌گیرند. این روش برای نخستین بار توسط لیانگ و زیگر[2] (1986) معرفی شد و سپس توسط پرنتیس[3] (1988) و ژائو و پرنتیس[4] (1990) تعمیم یافت. در این روش‌ها که به ترتیب به GEE1 و GEE2 مشهورند، اندازه‌های تکراری به صورت دوحالتی (1و0) در نظر گرفته می‌شوند، با این تفاوت که در روش $GEE1$ همبستگی بین مشاهدات به عنوان یک پارامتر مزاحم در نظر گرفته می‌شود، اما در $GEE2$ پارامتر ارتباط یا همبستگی به اندازه پارامترهای رگرسیونی مهم تلقی می‌شود و برآوردی از آن به دست می‌آید. مطالعات مروری مختلفی در مورد این دو روش صورت گرفته است. به عنوان نمونه خواننده می‌تواند به مقالات لیانگ و همکاران[5] (1992)، زیگر و لیانگ[6] (1992) مراجعه کند.

یکی از مشکلات مهم در استفاده از این روش وقوع داده‌های دورافتاده است که معمولا باعث تغییرات زیادی در برآورد ضرایب رگرسیونی می‌شود. جهت کاهش اثر این داده‌ها و مقاوم‌سازی این روش در برخورد با داده‌های دورافتاده روش‌های استوار سازی استفاده می‌شود. در رابطه با مطالعات انجام یافته در زمینه استوارسازی روش معادلات برآوردگر تعمیم‌یافته می‌توان به تحقیقات پان[7] (2001) و وانگ و لونگ[8] (2010) اشاره کرد. به طور کلی در روش GEE مدل‌های حاشیه‌ای را برای پاسخ‌ها یا خوشه‌های همبسته برازش داده و از یک برآوردگر فشرده برای ماتریس واریانس کواریانس ضرایب رگرسیونی استفاده می‌شود. اگرچه این برآوردگر نسبت به ساختار همبستگی پاسخ‌ها نسبتا استوار است، پان و سپس وانگ و لونگ، برآوردگرهای تصحیح شده استوار دیگری که برای نمونه‌های متناهی از اریبی کمتر و کارایی بالاتری برخوردار می‌باشند، معرفی کردند.

در این مقاله با استفاده از روش کارول و پدرسن[9] (1993) در استوار سازی رگرسیون لوژستیک، روش‌های استوارسازی معادلات برآوردگر تعمیم‌یافته را مطرح و آنها را برای مدلی شبیه‌سازی شده با پاسخ‌های دودویی همبسته به کار خواهیم برد. در این روش‌ها یک ماتریس وزن‌دهی قطری که به هر مشاهده وزنی بین صفر و یک می‌دهد، به کار می‌رود. این عمل تاثیر مشاهدات دورافتاده را روی برآورد پارامترها کم می‌کند. وزن‌دهی به مشاهدات در دو کلاس مالوس و شوئیپ انجام می‌شود. در کلاس مالوس وزن دهی بر مبنای خاصیت اهرمی مشاهدات و در کلاس شوئیپ بر مبنای مانده‌ها انجام می‌شود، برای توضیحات بیشتر به کاکیش و پرایسر[10] (1999) مراجعه شود. مطالب ارایه شده در بخش‌های بعدی به شرح زیر است: نخست معادلات برآوردگر تعمیم‌یافته را معرفی می‌شود. سپس روش کمترین مربعات دوباره وزنی شده تکراری برای حل معادلات براوردگر تعمیم‌یافته بیان می‌گردد. در ادامه معادلات

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.

برآوردگر تعمیم‌یافته استوار را مطرح کرده و سپس با استفاده از شبیه‌سازی روش معادلات برآوردگر تعمیم‌یافته استوار (REGEE)[1] با روش معمول آن (GEE) مقایسه می‌شود.

## 2. معادلات برآوردگر تعمیم‌یافته

در دو دهه اخیر توجه بسیاری از محققان به تجزیه و تحلیل داده‌های چندمتغیره و همبسته جلب شده است. در این گونه مطالعات متغیر پاسخ برای هر فرد در چندین نوبت متوالی مشاهده می‌شوند. به مطالعه‌ای که اندازه‌گیری مربوط به یک صفت در طول زمان مورد بررسی قرار می‌گیرد، مطالعه طولی می‌گویند. این گونه از اندازه‌گیری‌ها معمولا روی فرد و یا واحد اندازه‌گیری در زمان‌های مختلف در طول مطالعه انجام می‌شوند. به فرد و یا واحد نمونه‌گیری که برای وی در طول زمان چند اندازه‌گیری انجام شده است، یک خوشه گفته می‌شود. بدیهی است مشاهدات حاصل از هر فرد یا خوشه که در واقع اندازه‌های تکراری را تشکیل می‌دهند، با یکدیگر همبسته بوده، روش‌های معمول برای تحلیل داده‌های مستقل در مورد آن‌ها کارایی لازم را ندارند.

از آنجا که روش معادلات برآوردگر تعمیم‌یافته روشی برای برآورد ضرایب رگرسیونی در مدل‌های خطی تعمیم‌یافته می‌باشد، برای متغیر پاسخ $y_i, i = 1,2,\dots,n$ مدل خطی تعمیم‌یافته به شرح زیر است: .

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

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

فرض کنید در یک مطالعهٔ طولی برای هر موضوع یا فرد مورد مطالعه $n_i$ زمان اندازه‌گیری وجود داشته باشد بنابراین فرد $i$ام $(i = 1,2,\dots,k)$ در موقعیت‌های $t = 1,2,\dots,n_i$ مشاهده می‌شود. متغیر پاسخ مربوط به فرد iام در زمانt را می‌توان به صورت $y_{it}$ نمایش داد که می‌تواند 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 $p \times 1$ از متغیرهای کمکی دارد که به صورت زیر نمادگذاری می‌شود.

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

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

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

مشاهده می‌شود. مجموعه‌ای از مشاهدات همبسته مربوط به فرد $i$ام را خوشه نامیده و فرض می‌شود که مشاهدات درون خوشه‌ها همبسته‌اند، ولی بین یک خوشه و خوشه دیگر همبستگی وجود ندارد. مدل‌های بررسی شده به روش $(GEE)$ مدل‌هایی با میانگین حاشیه‌ای[2] نامیده می‌شوند. در یک مدل حاشیه‌ای برای مطالعات طولی امید حاشیه‌ای متغیر پاسخ یعنی $E(y_{it}) = \mu_{it}$ از طریق یک تابع پیوند به نام $g$ به $x_{it}$ به صورت زیر مربوط می‌شود:

1. Resistant Generalized Estimating Equations

2. Marginal Mean Model

$$\eta_{it} = g(\mu_{it}) = x'_{it}\beta$$

$\beta$ بردار پارامترهای نامعلوم از مرتبه $p \times 1$ بوده و فرض می‌شود که واریانس متغیر پاسخ تابعی از میانگین $\mu_{it}$ بوده و به صورت زیر بیان می‌شود:

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

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 دوجمله‌ای باشد $\phi = 1$ در نظر گرفته می‌شود (پارک و همکاران[1] (1998)). ماتریس $V_i$، ماتریس کوواریانس $Y_i$ می‌باشد. این ماتریس را می‌توان براساس ماتریس همبستگی عناصر $Y_i$ که آن را $R_i(\alpha)$ می‌نامیم به صورت زیر بیان کرد:

$$\begin{aligned} V_i &= V_i(\beta,\alpha,\phi) \\ &= A_i(\beta)R_i(\alpha)A_i(\beta)\phi \end{aligned}$$

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

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

از آنجا که همبستگی بین عناصر $y_i$ به درستی معلوم نیست، $R_i(\alpha)$ را ماتریس همبستگی کاری[2] می‌نامند، و باید با برآورد پارامتر $\alpha$، ماتریس همبستگی کاری $R_i(\alpha)$ را معلوم کرد. روش‌های مختلفی برای تعیین $R_i(\alpha)$ با توجه به نوع مدل به کار می‌رود. در بخش 2-1 روش‌های تعیین $R_i(\alpha)$ را بیان کرده و در ادامه برای برآورد $\beta$ تابع حداقل مربعات تعمیم‌یافته را به صورت زیر تشکیل می‌دهیم:

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

در این رابطه $\mu_i$ بردار میانگین پاسخ با عناصر $\mu_{it}$ است و داریم:

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

برای برآورد پارامترهای $\beta$ روش مینیمم کردن تابع حداقل مربعات تعمیم‌یافته منجر به معادلات برآوردگر تعمیم‌یافته به صورت زیر شده که با حل آن می‌توان $\beta$ 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

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

### 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) روش‌های زیر را پیشنهاد داده‌اند.

#### 2-1-1 ساختار همبستگی تبادل‌پذیر[3]

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

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

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

1. Park, T.P., Shin, D.W., Park, C.G.
2. Working Correlation Matrix
3. Exchangeable

است. روش GEE با ساختار همبستگی تبادل‌پذیر به منظور برآورد پارامتر همبستگی مشترک $(\alpha)$، از مانده‌های پیرسن برآورد شده توسط رابطهٔ

$\hat{e}_{it} = \left(y_{it} - \hat{\mu}_{it}\right) / \sqrt{V\left(\hat{\mu}_{it}\right)}$ استفاده می‌کند. برآورد $\alpha$ به‌دست آمده از این مانده‌ها به صورت

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

می‌باشد.

### 2-1-2 روش 1 – همبستگی[1]

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

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

در این صورت برآورد پارامتر همبستگی از رابطه

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

به دست می‌آید.

### 2-1-3 همبستگی اتورگرسیو[2]

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

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

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

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

$$\bar{\hat{e}}_{i2} = \frac{1}{n_i - 1} \sum_{j=2}^{n_i} \hat{e}_{ij}$$
$$\bar{\hat{e}}_{i1} = \frac{1}{n_i - 1} \sum_{j=1}^{n_{i-1}} \hat{e}_{ij}$$

برآورد می‌شود. ماتریس همبستگی کاری به صورت

$$\left(R(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{T-1} \\ \alpha & 1 & \alpha & \dots & \alpha^{T-2} \\ \alpha^2 & \alpha & 1 & \dots & \alpha^{T-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha^{T-1} & \alpha^{T-2} & \alpha^{T-3} & \dots & 1 \end{bmatrix} \right. \qquad (5)$$

خواهد بود.

### 2-1-4 روش بی ساختار[3]

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

$$\hat{\alpha}_{ts} = \frac{1}{(K-P)\hat{\phi}} \sum_{i=1}^{k} \hat{e}_{it}\hat{e}_{is} \qquad (6)$$
$$corr(y_{it}, y_{is}) = \begin{cases} 1 & t = s \\ \alpha_{ts} & t \neq s \end{cases}$$

است.

## 3- معادلات برآوردگر تعمیم‌یافته به روش کمترین مربعات دوباره وزنی شده تکراری[4] (IRLS)

جهت استوار سازی **معادلات برآوردگر تعمیم‌یافته، ناگزیر به** بیان **روش** IRLS **برای این خانواده از معادلات می‌باشیم. جهت این کار** فرض می‌کنیم: $N = \Sigma n_i$ و $Y = \left(y_1', y_2', \dots, y_k'\right)'$ یک بردار $N \times 1$ و $X = \left(\mathrm{x}_1', \mathrm{x}_2', \dots, \mathrm{x}_k'\right)'$ یک ماتریس $n_i \times p$ پر رتبه ستونی از متغیرهای کمکی باشد. مدل کلی $\eta = g(\mu) = X\beta$ را که در آن

1. 1-dependence
2. Auto regressive
3. Unstructured
4. Iteratively reweighted least squares

$\mu' = [\mu_1, \mu_2, \dots, \mu_k]$ می‌باشد در نظر گرفته و با توجه به بسط مرتبهٔ 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

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

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 $D_i^* = diag\{(\partial\eta_{it})/(\partial\mu_{it})\}$ 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 رابطهٔ $W_i^* = D_i^{*-1}A_i^{-1}R_i^{-1}(\hat{\alpha})A_i^{-1}D_i^{*-1}$ و پس از برآورد $\hat{\alpha}$ و تعیین $R_i(\alpha)$ به روش‌های بیان شده در بخش (2-1)، می‌توان $W_i^*$ و برآورد جدید $W^*$ را به دست آورد، سپس $\hat{\beta}_{new}$ را با استفاده از رابطهٔ $\hat{\beta}_{new} = (X'W^*X)^{-1}X'W^*Z^*$ محاسبه کرده و مراحل بالا را تا جایی که اختلاف $\hat{\beta}$ها کم و به تقریب قابل قبولی برسد تکرار می‌کنیم. حال رابطهٔ 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 تاثیر آنها روی مقادیر برازش داده شده هستند .اگر $p = tr(H)$ در نظر گرفته شود متوسط $h_{it}$ با $\frac{P}{N}$ نشان داده خواهد شد و $E = D^*(y - \hat{\mu}) = Z^* - \eta = (I - H)Z^*$ بردار مانده‌ها خواهد بود.

## 4- معادلات برآوردگر تعمیم‌یافته‌ی استوار

معادلات برآوردگر تعمیم‌یافته استوار توسط پرایسر و کاکیش[1] (1996) و (1999) به صورت زیر تعریف شده است:

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

در این رابطه اگر $W_i = I$ و $C_i = 0$ در نظر گرفته شوند معادلات برآوردگر تعمیم‌یافته مطرح شده توسط لیانگ وزیگر (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 ناااریب‌سازی[2] نامیده می‌شود، باید به‌گونه‌ای محاسبه شود که معادلات برآوردگر تعمیم‌یافته در رابطه (8) ناااریب 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

1. Preisser and Qaqish.
2. Debiasing factor

کلاس شوئیپ باید $C_i$ را طوری تعیین کرد که معادلات برآوردگر رابطه (8) نااریب باشند. پرایسر و کاکیش[1] (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.

3) $\mathrm{var}(\psi_\mathrm{i}) < \infty$

4) $\psi_\mathrm{i}$ در $\beta$ مطلقاً پیوسته بوده و مشتق آن نسبت به $\mu_i$ با $\psi_\mathrm{i}$ نشان داده می‌شود. و برای هر $y$، $E\|\psi_i\| < \infty$

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

$$\lim_{k\to\infty}\left(\sum_{i=1}^{k} D_i' V_i^{-1}\Gamma_i D_i\right)^{-1} \left\{\sum_{i=1}^{k} D_i' V_i^{-1}\mathrm{Var}(\psi_\mathrm{i}) V_i^{-1} D_i\right\} \left(\sum_{i=1}^{k} D_i' V_i^{-1}\Gamma_i D_i\right)^{-T} \quad (9)$$

$$\dot{C}_i = \frac{\partial}{\partial \mu_i} C_i$$

$$\dot{\psi}_{\mathrm{ki}} = \frac{\partial}{\partial \mu_i}\psi_\mathrm{i}(\mu_i)$$

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

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

برآورد $\beta$ در معادلات برآوردگر (8) با روش کمترین مربعات دوباره وزنی شده تکراری توسط مدل $Z^* = X\hat{\beta} + D^*(\psi - \mathrm{C})$ به دست آمده واریانس $\hat{\beta}_R$ توسط رابطه زیر به دست خواهد آمد

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

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

### 1-4 معادلات برآوردگر تعمیم‌یافته استوار برای مدل‌هایی با پاسخ دودوئی 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 $\beta$ مربوط‌اند .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:

$\pi_{11} = \mu_1\mu_2 + \rho v_1^{\frac{1}{2}} v_2^{\frac{1}{2}}$ و $\pi_{10} = \mu_1 - \pi_{11}$

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

تابع واریانس توسط رابطه‌ای $V_t = V(\mu_t) = \mu_t(1 - \mu_t)$ 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 تعمیم‌یافته 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

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

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

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

و

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

و

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

که در آن $b_t = \left(1-\mu_t\right)\mathrm{w}_\mathrm{t}^{(1)} + \mu_t \mathrm{w}_\mathrm{t}^{(0)}$ و $\rho_{tt'}$ همبستگی بین $y_t$ و $y_{t'}$ خواهد بود.

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

به منظور مقایسه روش معادلات برآوردگر تعمیم‌یافته با روش معادلات برآوردگر تعمیم‌یافتهٔ 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:

$$logit(\pi_{it}) = \beta_0 + \beta_1 \mathrm{x_{it}}$$
$$i = 1,2,\dots k$$
$$t = 1,2$$

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 تبادل‌پذیر (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 شویپ 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[1] (1997) پیشنهاد شده است. در روش کاهش وزن مشاهده شوئیپ $v = r_{it}$ و وزن تابعی از مانده پیرسن خواهد بود. در روش کاهش وزن مشاهده‌ی مالوس $v = h_{it}$ و وزن‌دهی بر اساس خاصیت اهرمی هر مشاهده انجام می‌شود. $a$ مقدار ثابتی است که ثابت میزان‌سازی[2] نامیده شده و باید آن را به گونه‌ای انتخاب نمود که بیشترین کارایی برای برآودگر حاصل شود. عموما 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 مشاهدهٔ شوئیپ با بررسی کارایی مدل با توجه به مقادیر متفاوت a، مقادیر $a = 3$ تا $a = 5$ برای این کار مناسب تشخیص داده شده‌اند.

### 5-1 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 .

1. Holland and Welsch
2. Tuning Constant

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

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

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

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

جدول (1) نشان می‌دهد که روش کاهش وزن مشاهده شوئیپ در برآورد $\beta_1$، زمانی که بین داده‌ها 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. می‌توان گفت در یک مدل

جدول (2) نشان می‌دهد 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 می‌کند ضمنا تغییر همبستگی نیز در نتایج

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

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

با تعداد خوشه=k50 و 5% آلودگی به نقاط دورافتاده اریبی روش‌های REGEE بین 15تا 30 در صد کمتر از روش (GEE) است و برای مدلی با k=50 و 3

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

### 2-5 مقایسه میانگین مربعات خطا در برآورد $\beta_1$ در معادلات برآوردگر تعمیم‌یافته و روش‌های استوار آن

در این بخش می‌خواهیم با استفاده از میانگین مربعات خطا روش معادلات برآوردگر تعمیم‌یافته را با روش‌های استوار آن مقایسه کنیم. همچنین می‌خواهیم با تغییر در اندازه خوشه‌ها به این سؤال پاسخ دهیم که افزایش تعداد خوشه‌ها چه تاثیری در میانگین مربعات خطا دارد؟ جدول (3) نتایج حاصل از این بررسی را نشان می‌دهد.

**جدول 3.** میانگین مربعات خطا برای روش‌های GEE و REGEE

| | مالوس | | شوئیپ | | GEE | | |
|---|---|---|---|---|---|---|---|
| MSE | برآورد $b_1$ | MSE | برآورد $b_1$ | MSE | برآورد $b_1$ | تعداد خوشه‌ها | %نقاط دورافتاده |
| 0.116 | 0.820 | 0.266 | 0.894 | 0.116 | 0.820 | K=100 | 0% |
| 0.056 | 0.811 | 0.060 | 0.818 | 0.056 | 0.811 | K=200 | 0% |
| 0.027 | 0.809 | 0.028 | 0.808 | 0.027 | 0.809 | K=400 | 0% |
| 0.018 | 0.808 | 0.019 | 0.818 | 0.018 | 0.807 | K=600 | 0% |
| 0.138 | 0.592 | 0.138 | 0.598 | 0.137 | 0.593 | K=100 | 5% |
| 0.095 | 0.5781 | 0.093 | 0.581 | 0.095 | 0.578 | K=200 | 5% |
| 0.074 | 0.574 | 0.073 | 0.574 | 0.074 | 0.572 | K=400 | 5% |
| 0.066 | 0.573 | 0.065 | 0.575 | 0.066 | 0.5734 | K=600 | 5% |

جدول (3) نشان می‌دهد که میانگین مربعات خطا در روش کاهش وزن مشاهدهٔ شوئیپ زمانی که در داده‌ها نقاط دورافتاده وجود ندارند بیشتر از روش GEE است اما زمانی که به میزان 5% نقاط دورافتاده در داده‌ها وجود داشته باشد میانگین مربعات خطای این روش کمتر از روش GEE می‌شود. در روش کاهش وزن مشاهده مالوس چه در زمانی که در داده‌ها 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 خوشه‌ها از میانگین مربعات خطا کم می‌شود. و در روش کاهش وزن مشاهده شوئیپ همواره این میزان هنگامی‌که نقاط دورافتاده در داده‌ها وجود داشته باشد کمتر از روش GEE است. به طور کلی از نتایج جدول (3) داریم که میزان کارایی روش شویپ نسبت به روش مالوس با افزایش میزان آلودگی از روندی صعودی برخوردار می‌باشد. بررسی مشابه نشان داده است که با افزایش عناصر ماتریس طرح میزان کارایی روش استوار به مقدار قابل ملاحظه‌ای افزایش می‌یابد (پرایسر و کاکیش، 1999).

### نتیجه‌گیری

با توجه به نتایج شبیه‌سازی از مقایسه روش‌های استوار سازی معادلات برآوردگر تعمیم‌یافته در دو کلاس شوئیپ و مالوس به نظر می‌رسد که:

**الف**: روش کاهش وزن مشاهده شوئیپ در برآورد $\beta_1$، زمانی که بین داده‌ها 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 داده‌ها نقاط دورافتاده

وجود ندارند بیشتر از روش GEE است اما زمانی که به میزان 5% نقاط دورافتاده در داده‌ها وجود داشته باشد میانگین مربعات خطای این روش کمتر از روش GEE می‌شود. از طرف دیگر با توجه به معیارهای اریبی و MSE تفاوت چندانی بین دو روش استوار دیده نمی‌شود، ولی در حالت کلی کلاس شوئیپ با توجه به میزان کاهش اریبی آن در برآورد $\beta_1$ و نیز از منظر میانگین مربعات خطا مناسب‌تر می‌باشد.

**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 eachb 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 forgeneralized estimating equations withimproved 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 anestimating equations approach, Journal of Statistical Planning andInference, 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.