

## حل معادلات برآوردکننده مدل‌های رگرسیون با اندازه خطای تصادفی روی متغیر مستقل به روش بهینه‌سازی

منوچهر بابانژاد<sup>۱\*</sup>

۱- دانشیار، دانشگاه گلستان، گروه آمار، گرگان، ایران

رسید مقاله: ۱۵ دی ۱۳۹۵

پذیرش مقاله: ۱۱ خرداد ۱۳۹۶

### چکیده

اندازه‌های بعضی از متغیرها در تحلیل‌های آماری اغلب با خطاهای تصادفی همراه می‌باشند. در نتیجه بررسی اثرات این خطاها مهم و ضروری به نظر می‌رسد. این بررسی در تحلیل‌های رگرسیونی وقتی اندازه‌های متغیر مستقل دارای خطای تصادفی باشند از اهمیت بیش‌تری برخوردار است. زیرا هدف برازش یک مدل رگرسیون برآورد اثر یک متغیر مستقل بر روی یک متغیر پاسخ می‌باشد. پس اگر اندازه‌های متغیر مستقل در یک مدل رگرسیون آلوده به خطاهای تصادفی باشند ممکن است در برآورد پارامترهای آن مدل تاثیر زیادی داشته باشند. ما در این مقاله ابتدا نحوه رخ دادن خطاهای تصادفی بر روی اندازه‌های یک متغیر را مورد بررسی قرار می‌دهیم. سپس نشان می‌دهیم که وجود این خطاها بر روی اندازه‌های متغیر مستقل در مدل‌های رگرسیون بر برآورد پارامترهای آن مدل‌ها تاثیر داشته، به طوری که حل مستقیم معادلات برآوردکننده را برای برآورد پارامترها ناممکن می‌سازد. همچنین نشان می‌دهیم به روش بهینه‌سازی با حل معادلات برآوردکننده می‌توان به برآورد پارامترهای یک مدل دست یافت. در نهایت نتایج روش بهینه‌سازی را بر روی دو مثال عملی آزمون می‌کنیم و اثرات نادیده گرفتن خطاهای تصادفی را در این دو مثال نشان می‌دهیم.

**کلمات کلیدی:** خطاهای تصادفی، مدل‌های رگرسیون خطی و غیر خطی، معادلات برآوردکننده، بهینه‌سازی.

### ۱ مقدمه

اندازه‌ها یا مشاهدات بعضی از متغیرها در تحقیقات آماری اغلب با خطاهای تصادفی همراه می‌باشند [۷-۱]. این خطاها ممکن است هنگام اندازه‌گیری رخ دهند، یا ممکن است توسط افراد هنگام خود اظهاری خواسته یا ناخواسته حادث شوند [۱۲-۶]. در هر دو صورت روبرو شدن با خطاهای تصادفی روی مشاهدات بعضی از متغیرها که در یک تحلیل آماری نقش کلیدی دارند امری اجتناب ناپذیر می‌باشد [۱۵-۱۰]. به عنوان مثال در مطالعات علوم پزشکی یا اپیدمیولوژی مقدار داروی مصرف شده توسط یک بیمار برای درمان یک نوع بیماری

\* عهده‌دار مکاتبات

آدرس الکترونیکی: m.babanezhad@gu.ac.ir

همراه با خطا می باشد [۸-۹]. زیرا مطالعات نشان دادند که مقدار دارویی که پزشک برای بیمار تجویز می کند و آن مقداری که بیمار در طول درمان خود مصرف می کند متفاوت است [۸]. به دلیل اینکه اغلب بیماران در میانه دوره درمان احساس بهبودی کامل می کنند، یا برعکس احساس می کنند که تجویز پزشک ناکارآمد است. بنابراین طول درمان را کامل نمی کنند و یا داروها را دقیقاً مطابق با دستورات پزشک مصرف نمی کنند [۶-۹]. یا به طور مثال در برآورد اثر مصرف سیگار بر روی وزن نوزادان به دنیا آمده، ممکن است مادران گزارش نادرستی از مصرف روزانه سیگار خود در زمان خود اظهاری بدهند [۱۱]. بنابراین ممکن است بین میزان واقعی سیگار مصرف شده با میزان گزارش شده توسط مادران تفاوت وجود داشته باشد [۸-۹]. در نتیجه اگر محققان بخواهند اثر دقیق یک نوع دارو را روی یک نوع بیماری و یا اثر دقیق مصرف سیگار بر روی وزن نوزادان را با برازش یک مدل رگرسیون مناسب برآورد کنند باید نقش و اثر خطاهای تصادفی را در فرآیند برآورد در نظر بگیرند [۹-۱۲]. در غیر این صورت با نادیده گرفتن اثرات خطاهای تصادفی ممکن است برآورد نادرستی از اثر دارو یا اثر سیگار که اندازه های شان آلوده به خطاهای تصادفی هستند به دست آیند [۹-۱۲]. فرض کنید که متغیر مستقل  $X$  مقدار داروی تجویز شده توسط پزشک برای یک بیمار و یا میزان واقعی مصرف سیگار توسط یک مادر که یک نوزاد بدنیا آورده باشد، و فرض کنید که متغیر  $Z$  مقدار داروی مصرف شده توسط آن بیمار باشد، و یا میزان مصرف سیگار خود اظهاری آن مادر باشد. در این صورت آنچه را یک محقق از روی داده های نمونه گیری شده برآورد می کند اثر  $Z$  است. زیرا داده های مشاهده شده اندازه های متغیر  $Z$  هستند و اندازه های متغیر  $X$  در دسترس وی قرار ندارند.

هدف این مقاله بررسی میزان تاثیر پذیری در مقدار و همچنین در فرایند برآورد پارامترهای مدل های رگرسیون خطی و غیر خطی می باشد وقتی مشاهده های متغیر مستقل  $X$  آلوده به خطاهای تصادفی باشند. ابتدا نشان می دهیم که وجود چنین خطاهایی سبب پیچیده شدن معادلات برآورد کننده مدل رگرسیون می گردند، به طوری که پارامترهای اضافی وارد مدل می شوند و این پارامترها امکان حل مستقیم معادلات برآورد کننده را برای برآورد پارامترهای اصلی سخت می کنند. سپس نشان می دهیم که با روش بهینه سازی با استفاده از یک الگوریتم مناسب می توان با حل معادلات برآورد کننده پارامترهای اصلی مدل رگرسیون را برآورد کرد. در آخر نتایج روش بهینه سازی را بر روی دو مثال عملی آزمون می کنیم، و اثر نادیده گرفتن خطاهای تصادفی را در تحلیل اثرات داروی تجویز شده روی یک نوع بیماری و مصرف سیگار روی وزن نوزادان نشان می دهیم.

## ۲ خطای تصادفی روی اندازه های یک متغیر

چگونگی رخ دادن خطای تصادفی روی اندازه های یک متغیر بستگی به نوع آن متغیر دارد. به عبارت دیگر رابطه بین خطای تصادفی روی یک متغیر به پیوسته یا گسسته بودن آن متغیر متفاوت می باشد. اگر متغیر  $X$  یک متغیر پیوسته باشد، رابطه بین خطای تصادفی و متغیر  $X$  به صورت زیر بیان می شود [۴-۱]:

$$X = Z \pm \varepsilon \quad (1)$$

که در آن  $X$  را متغیر واقعی،  $Z$  را متغیر مشاهده شده و  $\varepsilon$  را عبارت خطای تصادفی گویند. معادله (۱) را معادله خطای تصادفی کلاسیک گویند. در این حالت اغلب فرض بر این است که عبارت خطای تصادفی  $\varepsilon$  دارای توزیع نرمال با میانگین صفر و واریانس  $\sigma^2$  می باشد و همچنین  $E(\varepsilon|Z) = E(\varepsilon) = 0$ ؛ یعنی عبارت خطای تصادفی  $\varepsilon$  مستقل از متغیر مشاهده شده  $Z$  می باشد. از فرض اخیر نتیجه می شود که  $E(X|Z) = Z$ ؛ یعنی متغیر  $Z$  یک مقدار ناریب برای متغیر  $X$  می باشد. در مثال اثر یک نوع دارو روی یک نوع بیماری، متغیر  $X$  مقدار داروی تجویز شده توسط پزشک، متغیر  $Z$  مقدار داروی مصرف شده توسط بیمار و  $\varepsilon$  عبارت خطای تصادفی بین  $X$  و  $Z$  می باشد.

در حالتی که متغیر  $X$  گسسته باشد، برخلاف حالت پیوسته خطای تصادفی به دلیل نادرست طبقه بندی شدن روی متغیر  $X$  رخ می دهد. در این حالت بدون کاستن از کلیت بحث، فرض می کنیم که متغیر گسسته  $X$  یک متغیر دوتایی باشد. در این صورت رابطه بین متغیر  $X$  و متغیر  $Z$  با احتمال شرطی زیر بیان می شود [۱۲-۸]:

$$\Pi_{x|z} = P(X = x | Z = z) \quad (2)$$

که در آن  $z = 0, 1$  و  $x = 0, 1$ . احتمال شرطی (۲)، احتمال خطای مشاهده ات  $X$  طبقه بندی شده در مشاهدات  $Z$  نامیده می شود. در مثال میزان مصرف سیگار، بطور مثال کمیت  $\Pi_{10}$  بیانگر احتمالی است که یک مادر واقعا سیگار می کشد ( $x = 1$ ) در صورتی که اظهار کرده که سیگار نمی کشد  $z = 0$ .

### ۳ معادلات برآورد کننده مدل رگرسیون

همان طور که گفته شد در تحلیل رگرسیون هدف برآورد اثر یک متغیر مستقل مانند  $X$  بر روی یک متغیر پاسخ مانند  $Y$  می باشد، و این اثر با برآورد پارامترهای مدل تعیین می شود. در این بخش به برآورد پارامترهای مدل با حل معادلات برآورد کننده دو مدل رگرسیون خطی و غیر خطی می پردازیم که در آن ها اندازه های متغیر مستقل  $X$  آلوده به خطاهای تصادفی می باشند.

### ۳-۱ معادله برآورد کننده مدل رگرسیون خطی

فرض کنید علاقه مند به برآورد اثر یک متغیر مستقل  $X$  روی یک متغیر پاسخ  $Y$  با برازش مدل رگرسیون خطی زیر باشیم:

$$E(Y|X) = \beta_0 + \beta_1 X \quad (3)$$

برآورد اثر متغیر مستقل  $X$  روی متغیر پاسخ  $Y$  با برآورد ضرایب  $\beta_0$  و  $\beta_1$  مدل (۳) با حل معادله برآورد کننده این مدل دست می آید. در این حالت فرض بر این است که متغیرهای  $X$  و  $Y$  متغیرهای پیوسته هستند. ضرایب  $\beta_0$  و  $\beta_1$  به راحتی با حل معادله برآورد کننده مدل (۳) قابل برآورد می باشند. ولی اگر اندازه های متغیر  $X$  آلوده به خطای تصادفی باشند، در این حالت ما در واقع اثر  $Z \pm \varepsilon$  را روی  $Y$  برآورد می کنیم. بنابراین در عمل با مدل زیر سر و کار خواهیم داشت:

$$E(Y|Z) = \theta_0 + \theta_1 Z \quad (4)$$

یعنی در واقع به جای ضرایب  $\beta_0$  و  $\beta_1$  ضرایب  $\theta_0$  و  $\theta_1$  را برآورد می کنیم. ضرایب  $\theta_0$  و  $\theta_1$  با حل معادله برآورد کننده مدل (۴) که به صورت زیر نوشته می شود قابل برآورد می باشد:

$$E\{\Phi(\theta_0, \theta_1)\} = E\left\{\left(\frac{1}{Z}\right)(Y - \theta_0 - \theta_1 Z)\right\} = 0 \quad (5)$$

برای برآورد ضرایب  $\beta_0$  و  $\beta_1$  باید این ضرایب را وارد معادله (۵) کرد. با استفاده از خاصیت های امید ریاضی شرطی داریم:

$$\begin{aligned} E\{\Phi(\theta_0, \theta_1)\} &= E\left\{\left(\frac{1}{Z}\right)(Y - \theta_0 - \theta_1 Z)\right\} \\ &= E\{E\left(\frac{1}{Z}\right)(Y - \theta_0 - \theta_1 Z) | X, Z\} = 0 \end{aligned}$$

در تحلیل رگرسیونی با حضور خطاهای تصادفی اغلب فرض بر این است که خطای تصادفی روی متغیر  $X$  "غیر تشخیص دهنده" می باشد، یعنی متغیر پاسخ  $Y$  به شرط  $X$  مستقل از  $Z$  است. بنابراین با توجه به این فرض رابطه اخیر به صورت زیر خواهد بود:

$$\begin{aligned} E\{\Phi(\theta_0, \theta_1)\} &= E\left\{\left(\frac{1}{Z}\right)(E(Y | X) - \theta_0 - \theta_1 Z)\right\} \\ &= E\left(\frac{1}{Z}\right)(\beta_0 + \beta_1 X - \theta_0 - \theta_1 Z) = 0 \end{aligned} \quad (6)$$

همان طور که در رابطه (۶) مشاهده می شود وجود خطای تصادفی سبب شد که در این معادله پارامترهای  $\beta_0$ ،  $\beta_1$  و عبارت خطای  $\varepsilon$  اضافه شوند. در نتیجه امکان حل مستقیم این معادله برای برآورد پارامترها وجود نداشته باشد. قابل بیان است که داشتن اطلاعات از  $\varepsilon$  مستلزم داشتن مقدار واریانس آن یعنی  $\sigma^2$  می باشد. بنابراین معادله (۶) علاوه بر پارامترهای  $\beta_0$  و  $\beta_1$  شامل پارامتر  $\sigma^2$  نیز می باشد. در معادله (۶) پارامترهای  $\beta_0$ ،  $\beta_1$  و  $\sigma^2$  مجهول اند که باید به روش بهینه سازی برآورد شوند.

### ۳-۲ معادله برآورد کننده مدل رگرسیون غیرخطی

فرض کنید که علاقه مند به برآورد اثر متغیر مستقل  $X$  روی متغیر پاسخ  $Y$  (در این حالت  $X$  و  $Y$  متغیرهای گسسته و دوتایی می باشند) با برازش مدل رگرسیون غیرخطی زیر باشیم:

$$\text{logit } P(Y = 1 | X = x) = \beta_0 + \beta_1 x \quad (7)$$

برآورد اثر  $X$  روی  $Y$  از طریق حل معادله برآورد کننده مدل (۷) با برآورد ضرایب  $\beta_0$  و  $\beta_1$  به دست می آید. اگر اندازه های متغیر  $X$  دارای خطای تصادفی باشند ما در واقع اثر متغیر  $Z$  را روی  $Y$  برآورد می کنیم. بنابراین در عمل با مدل زیر سر و کار خواهیم داشت:

$$\text{logit } P(Y = 1 | Z = z) = \theta_0 + \theta_1 z \quad (8)$$

یعنی در واقع به جای برآورد ضرایب  $\beta_0$  و  $\beta_1$ ، ضرایب  $\theta_0$  و  $\theta_1$  را برآورد می کنیم. برای برآورد ضرایب  $\theta_0$  و  $\theta_1$  باید معادله برآورد کننده مدل (۸) که به صورت زیر نوشته شده است را حل کرد:

$$E\{\Phi(\theta_0, \theta_1)\} = E\left\{\left(\frac{1}{Z}\right)(Y - \text{expit}(\theta_0 + \theta_1 Z))\right\} = 0 \quad (9)$$

$$\text{که در آن } \text{expit}(\theta_0 + \theta_1 z) = \frac{e^{\theta_0 + \theta_1 z}}{1 + e^{\theta_0 + \theta_1 z}}$$

برای برآورد کردن ضرایب  $\beta_0$  و  $\beta_1$ ، با استفاده از خاصیت احتمال شرطی داریم:

$$P(Y = 1 | Z = z) = \int P(Y = 1 | Z = z, X = x) f(x | z) dx \\ = \int P(Y = 1 | X = x) f(x | z) dx$$

تساوی دوم معادله فوق از خاصیت خطای تصادفی که بیان می‌کند خطای تصادفی روی  $X$  "غیر تشخیص دهنده" می‌باشد، یعنی متغیر پاسخ  $Y$  به شرط  $X$  مستقل از  $Z$  باشد به دست آمده است. با توجه به دوتایی بودن متغیرهای  $X$  و  $Z$  عبارت اخیر به صورت زیر خواهد بود:

$$P(Y = 1 | Z = z) = P(Y = 1 | X = 1)P(X = 1 | Z = z) + P(Y = 1 | X = 0)P(X = 0 | Z = z) \quad (10)$$

که در آن  $z = 0, 1$  و احتمالات  $\Pi_{x|z} = P(X = x | Z = z)$  همان احتمالات خطای مشاهدات  $X$  طبقه بندی شده در مشاهدات  $Z$  می‌باشند. در نتیجه با حل معادله (۱۰) به روش بهینه‌سازی می‌توانیم جواب‌های بهینه را برای ضرایب  $\beta_0$  و  $\beta_1$  به دست آوریم.

#### ۴ الگوریتم بسط ریشه‌های منحصر به فرد (Euniroot Algorithm)

همان‌طور که می‌دانیم حل یک معادله به روش بهینه‌سازی در واقع یافتن جواب بهینه با اطلاعات موجود بین متغیرها در آن معادله می‌باشد [۱۹-۱۶]. ما در این مقاله برای حل معادلات برآورد کننده (۶) و (۱۰) از بسط الگوریتم ریشه‌های منحصر به فرد (Euniroot Algorithm) استفاده می‌کنیم. ما این الگوریتم را از بسط الگوریتم uniroot به دست آوردیم که برای یک تابع یک متغیره نوشته شده است [۱۹]. الگوریتم بسط ریشه‌های منحصر به فرد برای هر پارامتر مجهول یک بازه‌ای ایجاد می‌کند به طوری که مقادیر معادله برآورد کننده در دو نقطه انتهایی این بازه دارای علامت‌های مختلف باشند. در این صورت ریشه معادله برای آن پارامتر در آن بازه به دست می‌آید [۱۹]. همان‌طور که ملاحظه کردیم معادله برآورد کننده (۶) تابعی از سه مجهول  $\beta_0$ ،  $\beta_1$  و  $\sigma^2$ ؛ و معادله برآورد کننده (۱۰) تابعی از سه مجهول  $\beta_0$ ،  $\beta_1$  و  $\Pi_{x|z}$  می‌باشد. لذا از آنجایی که هر دو معادله برآورد کننده (۶) و (۱۰) توابعی سه متغیره هستند، الگوریتم بسط ریشه‌های منحصر به فرد برای یافتن سه ریشه مناسب برای سه پارامتر در هر معادله سه بازه با توجه به دامنه تغییرات پارامترها ایجاد می‌کند. همان‌طور که می‌دانیم  $\sigma^2$  یک عدد مثبت و  $\Pi_{x|z}$  عددی بین صفر و یک می‌باشد. در نتیجه با انجام این الگوریتم بعد از تعیین بازه‌ها با تعداد تکرار مناسب، تمام ریشه‌های دو معادله (۶) و (۱۰) در صورت وجود به دست می‌آیند. مراحل انجام الگوریتم بسط ریشه‌های منحصر به فرد به صورت زیر می‌باشد:

- تعیین سه بازه برای سه پارامتر به طوری که هر بازه مقادیر مجاز ریشه را برای هر پارامتر (با توجه به دامنه تغییرات آن پارامتر) در بر داشته باشد. معادله (۶) تابعی از  $\beta_0$ ،  $\beta_1$  و  $\sigma^2$  بوده که بازه اولیه برای این پارامترها به ترتیب  $(-\infty, +\infty)$ ،  $(-20, +20)$  و  $(0, +\infty)$  می‌باشند. معادله (۱۰) تابعی از  $\beta_0$ ،  $\beta_1$  و  $\Pi_{x|z}$  بوده که بازه اولیه برای این پارامترها به ترتیب  $(-\infty, +\infty)$ ،  $(-5, +5)$  و  $(0, 1)$  می‌باشند.

- (۲) یافتن اعدادی مانند  $i$  در بازه های مرحله اول با تکرار مناسب تا زمانی که مقدار معادله برآورد کننده (۶) یا (۱۰) به ازای آن ها برابر صفر شود، در غیر این صورت عددی مانند  $i$  به دست آید که علامت معادله به ازای آن مخالف علامت معادله به ازای نقطه انتهایی (یا اولیه) باشد.
- (۳) تعیین اعدادی مانند  $k$  از بازه به دست آمده از مرحله دوم تا زمانی که مقدار معادله به ازای آن برابر صفر شود.
- (۴) مراحل اول، دوم و سوم آنقدر تکرار شوند تا ریشه قابل قبول برای هر پارامتر با توجه به دامنه تغییرات آن ها به دست آید.
- با اجرای الگوریتم بالا جواب های بهینه قابل قبول پارامترهای معادلات برآورد کننده های (۶) و (۱۰) بدست می آیند.

## ۵ کاربردهای عملی

برای نشان دادن نقش خطاهای تصادفی بر روی مشاهدات متغیر مستقل در این بخش به بیان دو مثال عملی در دو حالت متغیر مستقل پیوسته و گسسته می پردازیم. در این دو مثال با روش بهینه سازی با استفاده از الگوریتم بسط ریشه های منحصر به فرد با حل معادلات برآورد کننده (۶) و (۱۰) و با لحاظ کردن خطاهای تصادفی بر روی مشاهدات متغیر مستقل، پارامترهای مدل خطی (۳) و مدل غیرخطی (۷) را برآورد می کنیم. همچنین برآوردهای بدست آمده از حالت های وجود خطاهای تصادفی را با حالت های نادیده گرفتن خطاهای تصادفی بر روی مشاهدات متغیر مستقل مقایسه می کنیم.

### ۵-۱ خطاهای تصادفی بر روی مشاهدات متغیر مستقل پیوسته

برای کنترل فشار خون ۱۵۲ نفر که به یک کلینیک مراجعه کردند، به آن ها داروی لوزارتان تجویز شده است که به مدت سه هفته روزی دو عدد قرص (۱۰۰ گرمی) مصرف کنند [۸]. بر اساس پروتکل، کلینیک بعد از سه هفته با بیماران برای مراجعه مجدد تماس گرفته تا فشار خون آن ها را بعد از مصرف دارو اندازه بگیرد. در بررسی ها ملاحظه شد که بعضی از بیماران داروها را کامل مصرف نکردند. لذا بین میزان داروی تجویز شده و میزان واقعی مصرف شده توسط بیماران تفاوت وجود دارد. ما ابتدا با اعمال مدل خطی ساده (۴) و حل مستقیم معادله برآورد کننده (۵)، اثر داروی لوزارتان تجویز شده را بر روی فشار خون برآورد کردیم. نتایج در جدول (۱) آمده اند.

**جدول ۱.** برآورد ضرایب و فاصله اطمینان ۹۵ درصد میزان پایین آمدن فشار خون ۱۵۲ بیمار

در اثر مصرف داروی لوزارتان با نادیده گرفتن خطاهای تصادفی

| ضرایب      | برآورد | فاصله اطمینان ۹۵ درصد |
|------------|--------|-----------------------|
| $\theta_0$ | -۱۰/۸۳ | (-۱۶/۴۳ ؛ -۵/۲۳)      |
| $\theta_1$ | -۶/۱۵  | (-۱۲/۲۴ ؛ -۰/۰۷)      |

نتایج در جدول (۱) همان نتایج هوتهوبور و وانستیلانت (۲۰۰۵) [۸] می‌باشد. با لحاظ کردن خطاهای تصادفی روی مشاهدات مصرف داروی لوزارتان این داده‌ها را با برازش یک مدل رگرسیون خطی ساده با همان متغیرهای کمکی در مدل دوباره تحلیل کردیم. و با حل معادله برآوردکننده (۶) نتایج متفاوتی به دست آوردیم که در جدول (۲) آمده‌اند. جدول (۲) نشان می‌دهد که با لحاظ کردن خطاهای تصادفی روی مشاهدات مصرف داروی لوزارتان متوسط میزان پایین آمدن فشار خون بیماران برابر با  $4/28-$  و فاصله اطمینان ۹۵ درصدی میزان پایین آمدن فشار خون برابر با  $(-0/13؛ -8/69)$  بوده است.

**جدول ۲.** برآورد ضرایب و فاصله اطمینان ۹۵ درصد میزان پایین آمدن فشار خون ۱۵۲ بیمار در اثر مصرف

داروی لوزارتان با لحاظ کردن خطاهای تصادفی

| ضرایب     | برآورد | فاصله اطمینان ۹۵ درصد |
|-----------|--------|-----------------------|
| $\beta_0$ | -۱۲/۷۶ | $(-17/56؛ -7/96)$     |
| $\beta_1$ | -۴/۲۸  | $(-8/69؛ -0/13)$      |

$$\sigma^2 = 0/25$$

## ۵-۲ خطاهای تصادفی بر روی مشاهدات متغیر مستقل گسسته

برای برآورد اثر مصرف سیگار روی میزان ریسک به دنیا آمدن نوزادانی با وزن کم (وزن کم‌تر از دو و نیم کیلوگرم)، تعداد ۱۸۹ مادر مورد مطالعه قرار گرفتند [۱۱]. از گزارش خود اظهاری مادران دریافتند که ۷۴ نفر (۳۹ درصد) سیگار مصرف می‌کنند. از آنجا که اطلاعات مصرف سیگار توسط خود مادران گزارش شده است، مشاهده شد که بعضی از مادران اطلاعات نادرست می‌دهند. به عبارت دیگر مادری گزارش داده که سیگار مصرف نمی‌کند در حالی که مصرف می‌کند، یا مصرف نمی‌کند به اشتباه در پرسشنامه گزارش می‌دهد که مصرف می‌کند. در نتیجه متغیر  $X$  که مصرف واقعی سیگار ( $X=1$ ) یا عدم مصرف سیگار ( $X=0$ ) می‌باشد به خطا کلاس‌بندی شده است. بنابراین متغیر  $X$  دارای خطای تصادفی می‌باشد و این خطاها با احتمالات شرطی معادله (۲) که در آن متغیر  $Z$  مصرف سیگار گزارش (مشاهده) شده با مقادیر (عدم مصرف سیگار)  $Z=0$  و (مصرف سیگار)  $Z=1$  بیان می‌شود. متغیر پاسخ  $Y$ ؛ وزن نوزدان، یک متغیر دوتایی است، که در آن برای وزن کم‌تر از دو و نیم کیلوگرم  $Y=1$ ، و برای وزن بیش‌تر و مساوی با دو و نیم کیلوگرم  $Y=0$  می‌باشد. بنابراین با توجه به متغیر پاسخ که یک متغیر دوتایی می‌باشد، با اعمال مدل‌های غیرخطی (۸) و (۹) و حل معادله برآوردکننده (۱۰) می‌توانیم میزان ریسک به دنیا آوردن یک نوزاد با وزن کم‌تر از دو و نیم کیلوگرم توسط مادران سیگاری را برآورد کنیم. ابتدا با برازش مدل (۸)، بدون در نظر گرفتن اثر خطا میزان ریسک به دنیا آوردن یک نوزاد با وزن کم برابر با  $2/02$  و فاصله اطمینان ۹۵ درصد  $(3/78؛ 1/09)$  برآورد شده است [۱۱]. نتایج در جدول (۳) آمده‌اند.

**جدول ۳.** برآورد ریسک به دنیا آمدن نوزادان با وزن کم تر از دو و نیم کیلو گرم در اثر مصرف سیگار ۱۸۹ مادر با نادیده گرفتن اثرات خطاهای تصادفی

| ضرایب      | برآورد | فاصله اطمینان ۹۵ درصد |
|------------|--------|-----------------------|
| $\theta_0$ | ۰/۳۴   | (۰/۲۲ ؛ ۰/۵۱)         |
| $\theta_1$ | ۲/۰۲   | (۱/۰۹ ؛ ۳/۷۸)         |

ما با لحاظ کردن خطاهای تصادفی، این داده ها را دوباره تحلیل کردیم. با حل معادله برآورد کننده (۱۰) به روش بهینه سازی با استفاده از الگوریتم بسط ریشه های منحصر به فرد میزان ریسک متولد شدن یک نوزاد با وزن کم تر از دو و نیم کیلو گرم برای یک مادر سیگاری برابر با ۱/۸۵ و با ۹۵ درصد فاصله اطمینان برابر با (۱/۰۱ ؛ ۳/۴۳) به دست آمده است. نتایج در جدول (۴) آمده اند.

**جدول ۴.** برآورد ریسک به دنیا آمدن نوزادان با وزن کم تر از دو و نیم کیلو گرم در اثر مصرف سیگار ۱۸۹ مادر با لحاظ کردن اثرات خطاهای تصادفی

| ضرایب     | برآورد | فاصله اطمینان ۹۵ درصد |
|-----------|--------|-----------------------|
| $\beta_0$ | ۰/۳۲   | (۰/۱۰ ؛ ۱/۰۱)         |
| $\beta_1$ | ۱/۸۵   | (۱/۰۱ ؛ ۳/۴۳)         |

$$\Pi_{\theta_0} = 0/85; \Pi_{\theta_1} = 0/45; \Pi_{\beta_0} = 0/03; \Pi_{\beta_1} = 0/92$$

## ۶ نتیجه و جمع بندی

اگرچه مطالعات زیادی نشان دادند که وجود خطاهای تصادفی در اندازه های یک متغیر امری اجتناب ناپذیر است، با این حال محققان اغلب در تجزیه و تحلیل داده ها وجود خطاهای تصادفی را نادیده می گیرند و توجهی به میزان تاثیر گذاری آنها در برآورد پارامترها ندارند. همچنین بعضی از مطالعات بیان کردند که محققان اغلب آگاهانه وجود خطاهای تصادفی را نادیده می گیرند [۷-۳]. این ممکن است به این دلیل باشد که لحاظ کردن خطاهای تصادفی سبب پیچیده تر شدن فرایند برآورد پارامترها می شوند و همچنین تاکنون الگوریتم ساده و بسته ای که اثر پارامترهای خطای تصادفی را در فرایند برآورد پارامترها حل کند وجود ندارد [۱۰ و ۱۲]. برای مثال اگر مدل رگرسیون در نظر گرفته خطی آمیخته و یا غیر خطی آمیخته باشد، الگوریتم بکار رفته ما در این مقاله (Euniroot Algorithm) هم قادر به حل معادلات برآورد کننده این مدل ها نبوده و به الگوریتم پیچیده تری نیاز است، و تا آنجا که ما می دانیم تاکنون حتی به آن ها پرداخته نشده است. ما در این مقاله اثر خطاهای تصادفی بر روی یک متغیر مستقل در دو مدل رگرسیون خطی و غیر خطی را مورد بحث قرار دادیم. ما نشان دادیم که چگونه اندازه های خطاهای تصادفی وارد فرایند برآورد پارامترهای این مدل ها می شوند و چگونه می توان پارامترهای این مدل ها را در حضور خطاهای تصادفی برآورد کرد. همچنین با دو مثال عملی نشان دادیم که نادیده گرفتن اثر خطای تصادفی چگونه به برآورد اریبی از پارامترها در مدل های رگرسیون خطی و غیر خطی منجر می شوند. با علم به اینکه یک الگوریتم بسته برای حل اثرات خطای تصادفی در تحلیل رگرسیونی وجود ندارد، این تحقیق با

استفاده از الگوریتم بسط ریشه‌های منحصر به فرد (Algorithm Euniroot) که از بسط الگوریتم uniroot به دست آمده است، با حل معادلات برآوردکننده با بیش از یک متغیر با روش بهینه‌سازی توانست با به دست آوردن جواب بهینه به استنباط صحیح‌تری از پارامترهای دو مدل رگرسیون خطی و غیرخطی در حضور خطاهای تصادفی به پردازد. مزیت الگوریتم uniroot این است که همواره در حل معادلات برآورده‌کننده مدل‌های رگرسیونی به جواب بهینه منجر می‌شود. زیرا همواره تعیین جواب بهینه برای پارامترهای مدل‌های رگرسیونی مورد بحث بوده است [۱۸-۲۰]. از نتایج جدول (۲) در مقایسه با جدول (۱) مشاهده می‌شود که با در نظر گرفتن خطای تصادفی با واریانسی برابر با  $0.25$ ، فشار خون به میزان  $4/28$  (میلی متر بر جیوه) به ازای یک واحد مصرف لوزاتان پایین آمده است. در حالی که با نادیده گرفتن خطای تصادفی تحلیل رگرسیونی نشان داد که فشار خون به میزان  $6/15$  (میلی متر بر جیوه) به ازای یک واحد مصرف لوزاتان پایین آمده است [۸]. این تفاوت نشان می‌دهد اگر خطاهای تصادفی نادیده گرفته شوند تحلیل نادرستی از اثر دارو به دست می‌آیند که ممکن است با اعتماد به این تحلیل سلامت افراد در معرض خطر قرار گیرد. همچنین از نتایج جدول (۴) در مقایسه با جدول (۳) درمی‌یابیم که با در نظر گرفتن خطای تصادفی میزان ریسک بدنیا آمدن یک نوزاد با وزن کم‌تر از دو و نیم کیلوگرم از یک مادر سیگاری برابر با  $85$  درصد برآورد شده است. در حالی که با نادیده گرفتن خطای تصادفی، تحلیل رگرسیونی داده‌ها نشان داد که میزان ریسک به دنیا آمدن یک نوزاد با وزن کم‌تر از دو و نیم کیلوگرم توسط یک مادر سیگاری برابر با  $2/02$  فولد برآورد شده است [۱۱]. می‌دانیم که نوزادان با وزن کم‌تر از دو و نیم کیلوگرم بیشتر در معرض مرگ قرار دارند. همچنین از نتایج جدول (۴) ملاحظه می‌شود که  $35$  درصد مادران سیگاری بودند، در صورتی که اظهار داشتند که سیگاری نیستند. این نشان می‌دهد که اگر خطاهای کلاس بندی شدن یک متغیر را نادیده بگیریم تحلیل آماری نادرستی به دست خواهیم آوردیم و هر چه احتمال خطاهای کلاس بندی شدن بیش تر باشد اریبی برآوردها بیش تر خواهد بود. همچنین لازم به ذکر است که الگوریتم بسط ریشه‌های منحصر به فرد بکار رفته در این مقاله برای حل معادلات برآورده‌کننده مدل‌های رگرسیون خطی آمیخته و یا غیر خطی آمیخته با خطای تصادفی کارایی ندارد. در تحقیقات آینده باید به دنبال یک الگوریتم دیگر برای کنترل اثر خطاهای تصادفی در این مدل‌ها باشیم.

## منابع

[۲۰] مهدی الله دادی، م.، میش مست نهی، ح.، (۱۳۹۶). ناحیه جواب جدید برای حل مدل برنامه‌ریزی خطی بازه‌ای. مجله تحقیق در

عملیات در کاربردهای آن، (۲) ۱۴، ۱۲۱-۱۱۱.

- [1] Buonaccorsi, J. P., (2010). Measurement error: Models, Methods, and Application. CRC Press.
- [2] Buonaccorsi, J. P., Laake, P., Veierod, M. B., (2005). On the effect of misclassification on bias of perfectly measured covariates in regression. Biometrics, 61, 831-836.
- [3] Carroll, R. J., Ruppert, D., Stefanski, L. A., Crainiceanu, C. M., (2006). Measurement Error in Nonlinear Models, Second Edition. CRC Press.
- [4] Gustafson, P., (2003). Measurement Error and Misclassification in Statistics and Epidemiology Impacts and Bayesian Adjustments. Press/CRC.
- [5] Schennach, S. M., (2004). Estimation of nonlinear models with measurement error. Econometrica, 72, 33-75.

- [6] Carroll, R. J., Chen, X., Hu, Y., (2010). Identification and inference in nonlinear models using two samples with non-classical measurement errors. *Journal of Nonparametric Statistics*, 22, 379–399.
- [7] Hu, Y., (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: a general solution. *Journal of Econometrics*, 144, 27–61.
- [8] Goetghebeur, E., Vansteelandt, S., (2005). Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Statist. Meth. Med. Res.*, 14, 397-415.
- [9] Chen, X., Hong, H., Tamer, E., (2006). Measurement error models with auxiliary data. *Review of Economic Studies*, 72, 343-366.
- [10] Chen, X., Hong, H., Nekipelov, D., (2011). Nonlinear models of measurement errors. *Journal of Economic Literature*, 49, 901-937.
- [11] Jedrychowski, W., (1998). Exposure misclassification error in studies on prenatal effects of tobacco smoking in pregnancy and the birth weight of children. *Journal of Exposure Analysis and Environmental Epidemiology*, 8(3), 347-57.
- [12] Hu, Y., Sasaki, Y., (2015). Closed form estimation of nonparametric models with non-classical measurement errors. *Journal of Econometrics*, 185, 392–408.
- [13] Lewbel, A., (2007). Estimation of average treatment effects with misclassification. *Econometrica*, 75, 537-551.
- [14] Mahajan, A., (2006). Identification and estimation of regression models with misclassification. *Econometrica*, 74, 631-665.
- [15] Schennach, S., (2004a). Estimation of nonlinear models with measurement error. *Econometrica*, 72, 33-75.
- [16] Small, C. G., Wang, J., (2003). *Numerical Methods for Nonlinear Estimating Equations*. OxfordUniversity Press, New York.
- [17] Varadhan, R., Gilbert, P., (2009). BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software*, 32(4).URL. <http://www.jstatsoft.org/v32/i04>.
- [18] Nocedal, J., Wright, S. J., (1999). *Numerical Optimization*. Springer.
- [19] Nash, J. C., Varadhan, R., (2011). Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software*, 43, 1–14.