# برآورد پارامترهای مدل رگرسیونی به کمک آنتروپی ماکزیمم تعمیم‌یافته برای داده‌های استیلن

**منیژه صانعی طبس – غلامرضا محتشمی برزادران**

دانشکده علوم ریاضی، دانشگاه فردوسی مشهد

**چکیده:** یکی از شرایط لازم و ضروری برای مدل رگرسیونی چندگانه عدم وجود هم خطی است، وجود هم خطی بالا باعث می شود پیشگوهایی با هم بستگی بالا داشته باشیم که دراین حالت برآوردگرهای حداقل مربعات به خوبی برآورد نمی‌شوند و به دلیل بزرگی واریانس، این بر آوردگرها دقت خیلی کمی دارند. راهکارهای متعددی در هنگام برخورد با یک چنین مشکلاتی در رگرسیون مطرح شده است . گلان و همکاران(۱۹۹۲) برآوردگرهای آنتروپی ماکسیمم تعمیم یافته را معرفی نمودند. در اینجا برای یک مجموعه داده که از هم خطی بالایی برخوردار هستند هر دو نوع برآوردگرهای حداقل مربعات و آنتروپی ماکسیمم تعمیم یافته را محاسبه و به مقایسه آنها می پردازیم.

**واژه‌های کلیدی:** برآوردگرآنتروپی ماکسیمم تعمیم یافته، رگرسیون، برآوردگر حداقل مربعات، مدل های خطی

## ۱ مقدمه

هنگامیکه بر روی یک مجموعه داده رگرسیونی کار می‌کنیم شرایطی پیش می‌آید که این داده‌ها ما را با محدودیت هایی مواجه می‌سازد به عبارت دیگر داده‌هااز یک سری شرایط که انتظارش را داریم، برخوردار نیستند در برخورد با این گونه مسائل واقعا چه باید کرد؟ به عنوان مثال یک مجموعه داده که از یک فرایند یا فرایندهای اقتصادی تولید می‌شود را در نظر بگیرید. این نمونه از یک جامعه‌ای گرفته شود که خود تحت تاثیر عوامل و پیشامدهای فراوان ولی شمارا می‌باشد. با داشتن این مجموعه داده، به دنبال اطلاعاتی هستیم که به کمک آن در ابتدا بتوان رفتارافراد جامعه و سپس وضعیت سیستم را تشخیص داد. بر اساس این اطلاعات قادر هستیم رفتار اقتصادی آینده وهمچنین وضعیت آینده دستگاه را پیش بینی کنیم. اما این مجموعه داده ممکن است نواقص و ضعف هایی داشته باشد که با ابزار و روش های سنتی قادر به تحلیل آنها نیستیم مگر اینکه فرضیاتی را ایجاد کنیم. روش آنتروپی ماکسیمم تعمیم یافته رویکرد دیگری را برای حل بعضی از این مشکلات ارائه می دهد. برآوردگر آنتروپی ماکسیمم تعمیم یافته(GME)، بر اساس روش ماکسیمم آنتروپی کلاسیک جینز پایه گذاری شده است. جینز در سال ۱۹۵۷ استفاده از آنتروپی ماکسیمم را در تعیین توزیع های احتمال مجهول پیشنهاد داد. در سال ۱۹۹۶ گلان و همکاران روش آنتروپی ماکسیمم را به چارچوب رگرسیونی تعمیم دادند. آنها با بازنویسی دوباره مدل، خطاها و ضرایب رگرسیونی را به عنوان متغیرهای تصادفی گسسته بر روی تکیه‌گاه کراندار دوباره پارامتری کرده، پارامترها و خطاهای مجهول را به فرم احتمالات در نظر می گیرد. بدین ترتیب توزیع احتمال هر یک از عناصر مجهول برآورد گردیده و مجموع آنتروپی‌های توزیع ضرایب و خطاهای متناظربا محدودیت‌های سازگاری ماکسیمم می‌شوند. اگر هیچ رابطه خطی بین متغیرهای رگرسیونی نباشد، گفته می شود که متعامدند. اگر متغیرهای رگرسیونی بر هم عمود باشند، استنباط‌هایی نسبتا ساده شامل ۱–مشخص کردن اثرات نسبی متغیرهای رگرسیونی ۲– پیش بینی و یا

برآورد ۳– انتخاب یک مجموعه مناسب از متغیرها برای مدل صورت می‌گیرد. متاسفانه در بسیاری از کاربردهای رگرسیون متغیرهای رگرسیونی متعامد نیستند. هنگامیکه ارتباط خطی نزدیکی بین متغیرهای رگرسیونی وجود دارد گفته می‌شود که مسئله هم‌خطی چندگانه وجود دارد. وجود هم‌خطی بالا باعث می‌شود پیشگوهایی با هم‌بستگی بالا داشته باشیم که دراین حالت برآوردگرهای حداقل مربعات خیلی ضعیف برآورد می‌شوند و به دلیل بزرگی واریانس، این بر آوردگرها دقت خیلی کمی دارند. آدنیز و همکاران(۲۰۱۱) برای برآورد ضرایب رگرسیونی مدل رگرسیونی خطی به دلیل وجود هم خطی چندگانه شدید بین متغیرهای پیشگو، ضمن اعمال محدودیت‌های علامت پارامترهای رگرسیونی، از روش GME برای برآورد این پارامترهااستفاده کردند. در اینجا برای یک مجموعه داده که از هم‌خطی بالایی برخوردار هستند هر دو نوع برآوردهای حداقل مربعات(OLS) و آنتروپی ماکسیمم تعمیم‌یافته(GME) را محاسبه و به مقایسه آنها می‌پردازیم. داده‌های مربوط به درصد تبدیل هپتان نرمال به استیلن و سه متغیر رگرسیونی (این محموعه داده را از کتاب رگرسیون خطی کاربردی تالیف داگلاس مونت گومری، ۱۹۹۰ فصل هشتم انتخاب کرده‌ایم ) در این کتاب برآوردهای حداقل مربعات، محاسبه شده و با توجه به وجود هم خطی بالایی که در این داده‌ها مشاهده شده‌است راهکاری که برای برخورد با این مشکل ارائه داده محاسبه ضرایب رگرسیونی به روش رگرسیون مضرس بوده است. در این مقاله یک راهکار دیگر برای این مسئله معرفی و ضرایب رگرسیونی را به روش آنتروپی ماکسیمم تعمیم یافته محاسبه می کنیم. در ادامه ابتدا روش ذکرشده را که ابزار اصلی کار به‌شمار می‌آید، معرفی نموده و در بخش سوم به مبحث هم خطی چندگانه پرداخته و روش های تشخیص وجود هم خطی چندگانه را مورد بررسی قرار می دهیم. سپس برای داده های استیلن هردو نوع برآوردگرها را محاسبه نموده و به مقایسه آنها می پردازیم.

## ۲ روش آنتروپی ماکسیمم تعمیم یافته

فرض کنید $Y$ یک بردار $T$ بعدی از مشاهدات ونیز$\beta$، بردار $K$ بعدی از پارامترهای مجهول را در نظر بگیرید. $Y$ و $\beta$ از طریق رابطه زیر با هم مربوط می شوند:

$$Y = X\beta + e, \tag{۱}$$

که در آن $X$ یک ماتریس $(T \times K)$ از متغیرهای رگرسیونی و $e$ نیز یک بردار $(T \times ۱)$ از خطاها می باشند. $\beta = (\beta_۱, \beta_۲, ..., \beta_K)$ بردار $K$ تایی از ضرایب رگرسیونی است که باید برآورد شوند. برای تعیین برآوردهایی از پارامترهای رگرسیونی با استفاده از روش آنتروپی ماکسیمم تعمیم یافته مدل خطی (۱) را دوباره بازنویسی می کنیم بطوریکه برآوردهاو خطاهای مجهول به فرم احتمالات نوشته می شوند زیرا عناصر تابع آنتروپی ماکسیمم، احتمالات هستند. فرض کنید هر $\beta_k$ یک متغیر تصادفی گسسته با $M$ برآمد ممکن $(z_{k۱}, z_{k۲}, ..., z_{kM})$هر یک با احتمال متناظر $(p_{k۱}, p_{k۲}, ..., p_{kM})$ باشد به طوریکه $\sum_{m=۱}^{M} p_{km} = ۱$. براین اساس هر پارامتر $\beta_k$ را می‌توان به صورت زیر نوشت:

$$\beta_k = \sum_{m=۱}^{M} z_{km} p_{km}, \qquad k = ۱, ۲, ..., K, \ \ , M \geq ۲, \tag{۲}$$

تکیه گاه پارامتری یا مقادیر بردار $Z_k$ بر اساس اطلاع پیشین و یا تجارب قبلی تعیین می شود. به طور مشابه برای برآورد توزیع خطاها نیز می توان به این ترتیب عمل کرد که فرض می کنیم تکیه گاه خطا شامل $J \geq ۲$

نقطه و $\nu$ نیز یک ماتریس $(T \times TJ)$ از نقاط تکیه گاه بوده و $W$ نیز یک بردار $TJ \times ۱$ از وزن های متناظر نقاط تکیه گاه باشند به طوریکه $\sum_{j=۱}^{J} w_{tj} = ۱$. در این صورت هر بردار $e_t$ به صورت زیر بازنویسی می شود:

$$e_t = \sum_{j=۱}^{J} \nu_{tj} w_{tj}, \quad t = ۱, ۲, ..., T, J \geq ۲, \tag{۳}$$

برای تعیین کران های تکیه گاه خطاها، گلان وهمکاران(۱۹۹۶) از قانون $۳\sigma$ که توسط پوکلشیم(۱۹۹۴) مطرح شده بود استفاده کردند. آنها کران پایین را $-۳\sigma$ قرار داده و برای کران بالا از $+۳\sigma$ استفاده کردند که مقدار $\sigma$ نیز همان انحراف معیار نمونه‌ای بردار $Y$ می باشد. بر این اساس رابطه (۱) به صورت زیر نوشته می‌شود:

$$Y = X\beta + e = XZP + \nu W, \tag{۴}$$

جینز در سال ۱۹۵۷ نشان داده است که برای منابع عدم قطعیت مستقل، آنتروپی جمع‌پذیر است. (جزئیات این خاصیت در مقاله (۱۹۹۲) کاپور وکواسان آورده شده است) بنابراین با فرض اینکه وزن های مجهول پارامترها و تکیه گاه خطاهای مدل رگرسیونی از یکدیگر مستقلند، می توان با حل مسئله بهینه‌سازی آنتروپی ماکسیمم پارامترها وخطاهای مجهول را برآورد کرد. پس با مسئله آنتروپی ماکسیمم تعمیم یافته زیرمواجه‌هستیم:

$$\begin{aligned} \max_{P,W} H(P,W) &= -P' lnP - W' lnW \\ &= -\sum_{k=۱}^{K} \sum_{m=۱}^{M} p_{km} lnp_{km} - \sum_{t=۱}^{T} \sum_{j=۱}^{J} w_{tj} lnw_{tj}, \end{aligned} \tag{۵}$$

$$\text{s.t}$$

$$\begin{cases} \sum_{k=۱}^{K} \sum_{m=۱}^{M} x_{tk} z_{km} p_{km} + \sum_{j=۱}^{J} w_{tj} \nu_{tj} = Y_t, & t = ۱, ۲, ..., T. \\ \sum_{m=۱}^{M} p_{km} = ۱, & k = ۱, ۲, ..., K. \\ \sum_{j=۱}^{J} w_{tj} = ۱, & t = ۱, ۲, ..., T. \end{cases} \tag{۶}$$

برای حل مسئله بهینه‌سازی بالا با توجه به محدودیت‌های ذکرشده معادله لاگرانژ راتشکیل می‌دهیم که حل این معادله این نتایج را به دست می دهد:

$$\hat{p}_{km} = \frac{e^{-\sum_{t=۱}^{T} \lambda_t x_{tk} z_{km}}}{\sum_{m=۱}^{M} e^{-\sum_{t=۱}^{T} \lambda_t x_{tk} z_{km}}}, \qquad \hat{w}_{tj} = \frac{e^{-\lambda_t \nu_{tj}}}{\sum_{j=۱}^{J} e^{-\lambda_t \nu_{tj}}}.$$

با جایگذاری جواب های $\hat{p}_{km}$ و $\hat{w}_{tj}$ به ترتیب در هر یک از معادلات (۲) و (۳) برآوردهای آنتروپی ماکسیمم تعمیم یافته $GME$ به صورت زیربه دست می آیند:

$$\beta_k^{GME} = \sum_{m=۱}^{M} \hat{p}_{km} z_{km}, k = ۱, ۲, ..., K, e_t^{GME} = \sum_{j=۱}^{J} \hat{w}_{tj} \nu_{tj}, t = ۱, ۲, ..., T. \tag{۷}$$

با توجه به اینکه برآوردهای $GME$ ضرایب $\beta_k$ و $e$ به ضرایب لاگرانژ بستگی دارند و در این جا برای $\lambda_t$ فرم بسته ای نداریم پس برای $\beta, p, w, e$ نیز هیچ فرم بسته ای نداریم و به ناچار برای محاسبه آنهابه روش های عددی متوسل می شویم. برای محاسبات عددی و برآورد پارامترهاو نیز برای حل برخی از مسائل بهینه سازی از نرم افزارهایی از جمله $GAMS$ استفاده می کنیم.

## ۳ مسئله وجود هم خطی چندگانه

مدل رگرسیون چند متغیره(۱) را در نظر بگیرید. که در آن $Y$ یک بردار $T \times ۱$ ازپاسخ هاو $X$ یک ماتریس $T \times K$ از متغیرهای رگرسیونی و $\beta$ یک بردار $K \times ۱$ از ثابت های معلوم و e یک بردار $T \times ۱$ از خطاهای تصادفی ($e_i \sim N(۰, \sigma^۲)$) باشند. می‌توان هم خطی چندگانه را به صورت بستگی خطی ستون های ماتریس $X$ تعریف کرد. بردارهای $X_۱, ..., X_K$ به صورت خطی وابسته‌اند، هرگاه مجموعه‌ای از مقادیر ثابت $t_۱, ..., t_K$ که همه صفر نیستند، وجود داشته باشند که $\sum_{j=۱}^{K} t_j X_j = M$. گفته‌می شود هم خطی چندگانه وجود دارد. وجود هم خطی یک مشکل از شرایط بیمارگونه در ماتریس $X'X$ است که می تواند تحلیل حداقل مربعات معمولی مدل رگرسیونی را به نمایشی نامناسب تبدیل کند. هم خطی چندگانه به دلایل زیادی می تواند بوجود بیاید مثلا شیوه جمع آوری داده ها، اختصاص دادن مدل، یک مدل با متغیرهای رگرسیونی بیش از حد و نیز شرایط و قیدهای داخل مدل و یا جمعیت مورد نمونه گیری و ...

روش معمول در برخورد با هم خطی چندگانه در این زمینه کنارگذاشتن بعضی متغیرهای رگرسیونی است. روش حداقل مربعات، برآوردهای ضعیفی از هریک از پارامترهای منفرد به دست می دهد اما این دلیلی بر ضعیف بودن مدل پیش بینی شده نمی باشد زیرا علیرغم اینکه پارامترهای $\beta_j$ هریک به طور ضعیف برآورد می شوند، ولی ترکیب خطی $\sum_{j=۱}^{K} \beta_j x_{tj}$ کاملا خوب برآورد می شود. به این معنی که علیرغم برآوردهای نامناسب تک پارامترها، مشاهدات آینده می توانند به صورت دقیق پیش بینی شوند. برای آشکارکردن هم خطی چندگانه تکنیک های متعددی پیشنهاد شده است که بعضی از این معیارهاعبارتند از: ۱− محک تشخیص ماتریس هم بستگی ۲− عامل تورم واریانس ($VIF$) ۳− مقادیر ویژه $X'X$. روش های متعددی برای برخورد با مشکلات ناشی از هم خطی چندگانه پیشنهاد شده است. شامل جمع آوری داده های اضافی، تخصیص مدل مجدد و بکارگیری روش های برآورد غیر از حداقل مربعات که به طور ویژه برای از بین بردن مشکلات ناشی از هم خطی چندگانه طراحی شده‌اند. برای اطلاعات بیشتر در این زمینه به کتاب رگرسیون مونت گومری مراجعه شود.

در اینجا به روش آنتروپی ماکسیمم تعمیم یافته هر یک از ضرایب رگرسیونی را برآورد می کنیم. **مثال ۱**
داده های مربوط به درصد تبدیل هپتان نرمال به استیلن و سه متغیر رگرسیونی نمونه ای از داده های فرایند شیمیایی است که مناسبت یک سطح پاسخ درجه دوم بر حسب هر سه متغیر رگرسیونی به عنوان یک مدل پیشنهادی برای آن مورد توجه قرار می گیرد. متغیر$x_۱$ درجه حرارت راکتور، $x_۲$ نسبت $H_۲$ به هپتان نرمال و $x_۳$ زمان تماس بر حسب ثانیه. چون دو متغیر رگرسیونی زمان تماس و درجه حرارت راکتور به شدت به هم وابسته اند، مشکلات هم خطی بالقوه در این داده ها وحود دارد.

مدل درجه دوم کامل را به این داده ها برازش داده ایم.

$$\begin{aligned} y = {} & \beta_۰ + \beta_۱ T + \beta_۲ H + \beta_۳ c + \beta_{۱۲} TH + \beta_{۱۳} Tc + \beta_{۲۳} cH \\ & + \beta_{۱۱} T^۲ + \beta_{۲۲} H^۲ + \beta_{۳۳} c^۲ + \epsilon \end{aligned} \tag{۸}$$

جدول ۱ نتایج حاصل از برازش مدل (۱۰) و برآوردهای حداقل مربعات را نشان می دهد.

با توجه به اینکه $F_{۹,۶,۰.۹۷۵} = ۵.۵۲$ در سطح $\alpha = ۰.۰۵$ فرض $H_o : \beta_j = ۰; \forall j$ رد می شود. همچنین $R^۲ = ۰.۹۹۸$ و بنابراین ۹۹.۸٪ تغییرات داده ها توسط مدل توضیح داده می شود. مقادیر ویژه برای داده های استیلن عبارتند از $\lambda_۱ = ۴.۲۰۸۴, \lambda_۲ = ۲.۱۶۲۶, \lambda_۳ = ۱.۱۳۸۴, \lambda_۴ = ۱.۰۴۱۳, \lambda_۵ = ۰.۳۸۴۵, \lambda_۶ = ۰.۰۴۹۵, \lambda_۷ = ۰.۰۱۳۶, \lambda_۸ = ۰.۰۰۵۱, \lambda_۹ = ۰.۰۰۰۱$ و نیز عدد k برابر $k = \frac{۴.۲۰۸۴}{۰.۰۰۰۱} = ۴۲۰۸۴$ که حاکی از وجود هم خطی چندگانه شدید می‌باشد.

جدول ۱: برآوردهای حداقل مربعات

| جمله | ضریب | رگرسیونی | خطای معیار | $t_\circ$ | درجات آزادی | $SS$ | $F_\circ$ |
|---|---|---|---|---|---|---|---|
| $\beta_\circ$ | ۳۵.۹۲ | ۱.۱۰۵ | | ۳۲.۵۰ | | | |
| $T$ | ۳.۸۶ | ۴.۶۳۸ | | ۰.۸۳ | | | |
| $H$ | ۲.۷۴ | ۰.۳۰۷۴ | | ۸.۹۲ | | | |
| $C$ | −۸.۲۰۱ | ۶.۱۸۴۸ | | −۱.۳۲ | | | |
| $TH$ | −۶.۴۵۶۸ | ۱.۴۶۶۰ | | −۴.۴۰ | ۶ | $SSE = ۴.۸۷۵۶$ | $MSE = ۰.۸۱۲۶$ |
| $TC$ | −۲۶.۹۸۱۸ | ۲۱.۰۲۲۴ | | −۱.۲۸ | | | |
| $HC$ | −۳.۷۶۸۳ | ۱.۶۵۵۴ | | −۲.۲۸ | ۹ | $SSR = ۲۱۱۸.۸۳۸۲$ | $MSR = ۲۳۵.۴۲۶۵$ |
| $T^۲$ | −۱۲.۵۲۳۷ | ۱۲.۳۲۳۹ | | −۱.۰۲ | | | |
| $H^۲$ | −۰.۹۷۲۱ | ۰.۳۷۴۶ | | −۲.۶۰ | ۱۵ | $SST = ۲۱۲۳.۷۱۳۸$ | $F_\circ = \frac{MSR}{MSE} = ۲۸۹.۷۲$ |
| $C^۲$ | −۱۱.۵۹۴۳ | ۷.۷۰۷۰ | | −۱.۵ | | | |

در اینجا به کمک روش آنتروپی ماکسیمم تعمیم یافته ضرایب رگرسیونی را برآورد می کنیم. همانگونه که شرح آن در بخش دوم گذشت این برآوردگرها به صورت ضرایبی از احتمالات وخطاها هستند که هریک از عناصر $p_i$ و $w_i$ از ماکسیمم آنتروپی تعمیم یافته به دست می آیند.

$$Y = X\beta + e = XZP + \nu W \qquad (۹)$$

هر ضریب رگرسیونی $\beta_k$ به صورت یک متغیر تصادفی گسسته‌که از $M \geq ۲$ برآمد ممکن تشکیل شده است و به صورت زیر نمایش داده می شود:

$$\beta_k = z'_k p_k = [z_{k۱}, ..., z_{kM}] \begin{bmatrix} p_۱ \\ p_۲ \\ \vdots \\ p_{kM} \end{bmatrix}$$

به طور مشابه $\nu$ نیز یک ماتریس $T \times TJ$ از مقادیر معلوم تکیه گاه و $W$ یک بردار $TJ \times ۱$ از وزن های احتمالی بطوریکه $\sum_{j=۱}^{J} w_{tj} = ۱; \forall t$ که $J$ تعداد نقاط تکیه گاه انتخاب شده برای هر خطای $w_t$ می باشد. برای انتخاب تکیه گاه پارامتری دو مقدار $M = ۵$ و $M = ۷$ را در نظر می گیریم. برای هردو تکیه‌گاه خطا و ضرایب لازم است تعداد J و M نقطه به ترتیب انتخاب کنیم. اول تکیه گاه پارامتری را در نظر بگیرید. فرض کنیم M=5 یعنی برای هر پارامتر تکیه‌گاه ۵ نقطه‌ای داشته باشیم. تکیه‌گاه حول صفر متقارن است. پس میانگین پیشین هر پارامتر صفر است. لازم است که برا ی خطاها نیز تکیه گاه در نظر گرفته شود کران های خطا را $\pm ۳\sigma$ در نظر میگیریم که برای $\sigma$ دو امکان است ۱) $\hat{\sigma} = ۰.۹۰۱$ که از روش حداقل مربعات به دست آمده است و ۲) انحراف معیار نمونه ای Y که برابر است با $S = ۱۱.۹$. کمپبل و هیل (۲۰۰۶) نشان داده اند که اغلب مناسبتر است از انحراف معیارنمونه ای Y استفاده شود که ما هم از آن استفاده می کنیم. با قراردادن مقدار $\sigma = ۱۱.۹$ برای $M = ۵$ و $M = ۷$ کران های خطابه ترتیب عبارتند از: $\nu = (-۳۵.۷, -۱۷.۸۵, ۰, ۱۷.۸۵, ۳۵.۷)$ و $\nu = (-۳۵.۷, -۲۳.۸, -۱۱.۹, ۰, ۱۱.۹, ۲۳.۸, ۳۵.۷)$.

برآوردگرهای نقطه‌ای OLS و GME در جدول ۲ آورده شده‌اند. همانطور که ملاحظه میکنید نتایج برای $M = ۵$ خیلی با $M = ۷$ متفاوت نمی باشد. ولی با برآوردگرهای OLS چه از نظر مقدار و چه از نظر علامت خیلی متفاوت است . همه مقادیر ضرایب رگرسیونی GME از نظر مقدار از مقادیرمتناظر ضرایب رگرسیونی OLS کمتر می باشند.

جدول ۲: برآوردگرهای OLS و GME (n=16)

| پارامترها و متغیرها | برآوردگرهای OLS | برآوردگرهای M=5 GME | برآوردگرهای M=7 GME |
|---|---|---|---|
| constant $\beta_o$ | ۳۵.۹۲ | ۳۱.۳۵ | ۳۱.۶۱ |
| T $\beta_۱$ | ۳.۸۶ | ۱.۹۰ | ۱.۹۸ |
| H $\beta_۲$ | ۲.۷۴ | ۰.۴۶ | ۰.۴۹ |
| C $\beta_۳$ | −۸.۲۰ | −۴.۸۳ | −۵ |
| TH $\beta_۴$ | −۶.۴۶ | −۱.۴۴ | −۱.۵۳ |
| TC $\beta_۵$ | −۲۶.۹۸ | −۱.۷ | −۱.۶۸ |
| HC $\beta_۶$ | −۳.۷۷ | ۰.۲۸ | ۰.۳ |
| T2 $\beta_۷$ | −۱۲.۵۲ | ۲.۹۴ | ۳.۰۲ |
| H2 $\beta_۸$ | −۰.۹۷ | ۰.۳۴ | ۰.۳۴ |
| C2 $\beta_۹$ | −۱۱.۶ | −۱.۹۰ | −۱.۹۸ |

## بحث و نتیجه‌گیری

در جدول ۲ برآوردگرهای OLS و نیز برآوردگرهای GME برای دو مقدار متفاوت M=5 و M=7 آورده شده‌اند. نتایج نشان می‌دهند که تعداد بیشتر از ۵ نقطه برای تکیه‌گاه تاثیری در بهبود برآوردگرها ندارد. هر دو نوع برآوردگرهای OLS و GME هم از نظر بزرگی و هم از نظر علامت با یکدیگر متفاوت بوده و با کمی دقت ملاحظه می شود که دراین مثال همه برآوردگرهای GME از نظر مقدار از برآوردگرهای OLS متناظر کمتر می باشند. همچنین براساس این جدول

$$\hat{\beta}'_{OLS}\hat{\beta}_{OLS} = ۲۴۵۶.۱ > \hat{\beta}'_{GMEM۷}\hat{\beta}_{GMEM۷} = ۱۰۴۶.۷۶۲$$
$$> \hat{\beta}'_{GMEM۵}\hat{\beta}_{GMEM۵} = ۱۰۲۷.۳۷۴$$

حال برای مقایسه دو نوع برآوردگر یادشده میزان مخاطره میانگین مربعات خطا(MSE) را برای هردو آنها محاسبه می‌کنیم بدین منظور از روش بوت استرپ کمک می گیریم.

$$\hat{mse}(\hat{\beta}_{OLS}) = ۴۴۷۶.۴۵۲ > \hat{mse}(\hat{\beta}_{GMEM۷}) = ۲۴۴۱.۴۷۶$$
$$> \hat{mse}(\hat{\beta}_{GMEM۵}) = ۲۴۳۲.۹۴۹$$

ملاحظه می‌کنید که مقادیر mse برای برآوردگرهای GME از مقدار متناظر آن برای برآوردگر OLS خیلی کمتر می‌باشد، که این نکته با توجه به وجود هم خطی بالا در داده های استیلن قابل توجه و تامل می باشد.

## مراجع


Akdeniz, F., Cabuk, A. and Guler. H. (2011). *Generalized maximum entropy estimators: Applications to the Portland cement dataset*, The open statistics and probability Journal, 3, 13-20.

Campbell, R. C. and Hill, R. C. (2005). *A monte carlo study of the effect of design characteristics on the inequality restricted maximum entropy estimator*, Review of Applied Economics, 1, 53-84.

Campbell, R. C. and Hill, R. C. (2006). *Imposing parameter inequality restrictions using the principle of maximum entropy* , Journal of statistical computation and simulation, 76, 985-1000.

Csiszar, I. (1991). *Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems*, The Annals of statistics, 19, 2032-2066.

Golan, A., Judge, G. and Miller, D. (1996). *Maximum Entropy Econometrics: Robust estimation with limited data*, John wiley Sons, New york.

Golan, A. (2001). *A simultaneous estimation and variable selection rule*, Journal of econometrics, 101, 165-193.

Golan, A. and Perloff, J. M. (2002). *Comparison of maximum entropy and higher-order entropy estimators*, Journal of econometrics, 107, 195-211.

Jaynes, E. T. (1957). *Information theory and statistical mechanics.* Physical Reviews, 106, 620-630.

Judge, G. G. and Golan, A. (1992). *Recovering information in the case of ill-posed inverse problems with noise*, Mimo Department Of Agricultural And Natural Resources, University of California, Berekeley, CA.

Kapur, J. N. and Kesavan, H. K. (1992). *Entropy optimization principles with applications*, Academic press, London.

Koski, T. and Persson, L. E. (1992). *Some properties of generalized exponential entropies with applications to data compression*, Information sciences, 63, 103-132.

Kullback, S. (1959). *Information Theory And Statistics.* Johnwiley, New York.

Lee, A. (1997). *A mont carlo study of a generalized maximum entropy estimator of the binary choice model*, Advances in econometrics, 12, 183-197.

Mittelhammer, R. Cardel, S. and Marsh, L. T. (2002). *The data constrained GME estimator of the GML: Asymptotic theory and inference*, Working paper, Washington state university, Pullman, WA.

Pukelsheim, F. (1994). *The three-sigma Rule*, The American statistician, 48, 4, 88-91.

Shannon, C. E. (1948). *A Mathimatical Theory Of Communication.* Bell System Tech., 27, 379-423.