استفاده از ماکسیمم آنتروپی تی سالیس تعمیم‌یافته برای برآورد پارامتر رگرسیون ستیغی

منیژه صانعی طبس [۱]

تاریخ دریافت: ۱۴۰۱/۱۲/۰۱
تاریخ پذیرش: ۱۴۰۲/۰۲/۰۸

چکیده:
تحلیل رگرسیون به روش کمترین توان‌های دوم مستلزم برقراری فرضیات زیربنایی است. یکی از مسائلی که تحلیل رگرسیون به این روش را با مشکلات عمده مواجه می‌سازد وجود هم خطی در بین متغیرهای رگرسیونی است. روش‌های زیادی برای حل مشکلات ناشی از وجود هم خطی معرفی‌شده‌اند. یکی از این روش‌ها رگرسیون ستیغی است. در این مقاله یک برآورد جدید برای پارامتر ستیغی به کمک ماکسیمم آنتروپی تی سالیس تعمیم‌یافته ارائه داده و آن را برآوردگر ستیغی ماکسیمم آنتروپی تی سالیس مرتبه دو تعمیم‌یافته می‌نامیم. برای مجموعه داده‌های سیمان پرتلند که از هم خطی قوی برخوردار هستند و از سال ۱۳۳۲، برآوردگرهای مختلفی برای این داده‌ها ارائه‌شده است این برآوردگر را محاسبه و با برآوردگر ستیغی ماکسیمم آنتروپی تعمیم‌یافته و برآوردگر کمترین توان‌های دوم مقایسه می‌کنیم.

واژه‌های کلیدی: رگرسیون ستیغی، ماکسیمم آنتروپی تعمیم‌یافته، آنتروپی تی‌سالیس، ماکسیمم آنتروپی تی سالیس تعمیم‌یافته

۱ مقدمه

روش برآورد حداقل مربعات بهترین برآوردگر نااریب خطی را درصورتی‌که فرضیات زیربنایی زیر برقرار باشند به دست می‌دهد: ۱- جملات خطا دارای توزیع نرمال با میانگین صفر باشند. ۲- جملات خطا دارای واریانس ثابت باشند. ۳- متغیرهای پیشگو مستقل خطی باشند. برقرار نبودن هر یک از فرضیات ذکرشده در بالا ما را در استفاده از روش OLS با مشکل مواجه می‌سازد. اگر هیچ رابطه خطی بین متغیرهای رگرسیونی برقرار نباشد، گفته می‌شود متعامدند. اگر متغیرهای رگرسیونی برهم عمود باشند، استنباط‌هایی نسبتاً ساده صورت می‌گیرد. هنگامی‌که ارتباط خطی نزدیکی بین متغیرهای پیشگو وجود دارد، گفته می‌شود هم خطی چندگانه وجود دارد. در رگرسیون چندگانه یکی از فرض‌های اساسی که بین متغیرهای پیشگو باید برقرار باشد فرض عدم وجود هم خطی است اما درصورتی‌که این فرض برقرار نباشد تحلیل رگرسیون را با مشکل مواجه می‌کند. اگر متغیرهای پیشگو دارای وابستگی زیاد باشند، واریانس برآوردگرها زیاد شده و دیگر برآوردگر حداقل مربعات برآورد خوبی نیست. محققین زیادی از جمله فرار و گلوبر [۴] ارل و کنارد [۸] و ویلان و واتس [۲۱] بلیزلی [۳] و ... از دیدگاه کلاسیک این موضوع را موردنظر قرار داده‌اند. تیستد [۱۸] مسئله هم خطی را از دیدگاه نظریه تصمیم بررسی کرد. اغلب توصیه‌شده است که این مشکل با یک سری اطلاعات اضافی مدل قابل جبران است، اما اینکه چه میزان اطلاعات لازم است تا مشکلات ناشی از هم خطی را جبران کند، مشخص نیست. روش رگرسیون ستیغی از جمله روش‌هایی است که برای مقابله با مشکل هم خطی در رگرسیون به کار گرفته‌شده است. این روش توسط ارل و کنارد [۸] معرفی شد که یک روش رایج برای برآورد پارامترهای مدل است، بدون اینکه لازم باشد برخی از متغیرهای رگرسیونی را از مدل حذف کنیم. این روش با اضافه کردن یک ثابت نامنفی کوچک (که اغلب از آن به نام پارامتر ستیغی یاد می‌شود) به قطر ماتریس همبستگی متغیرهای توضیحی این امکان را می‌دهد که واریانس برآوردگر کمترین مربعات از طریق اعمال مقداری اریبی به مدل کاهش داده شود. اگرچه برآوردگرهای حاصل اریب هستند، ولی این اریبی آن‌قدر کوچک است که دقت این برآوردگرها از دقت برآوردگرهای نااریب بیشتر است. روش ماکسیمم آنتروپی تعمیم‌یافته[۲] (GME) با یک زیربنای اقتصادی توسط جاج و گلان [۵] معرفی شد. روش (GME) برای برآورد پارامترهای مدل رگرسیونی خطی و غیرخطی بدون اعمال هیچ شرطی روی توزیع احتمال

[۱] گروه آمار، دانشکده ریاضی آمار و علوم کامپیوتر، دانشگاه سیستان و بلوچستان (نویسنده مسئول: manijesanei@math.usb.ac.ir)
[2] Generalized Maximum Entropy

خطاها مورداستفاده واقع‌شده است. این روش که بر اساس ماکسیمم آنتروپی شانون و اصل ماکسیمم آنتروپی جینز [۹] پایه‌گذاری شده است مدل خطی را به‌گونه‌ای بازنویسی می‌کند که پارامترها و خطاهای مجهول را به شکل احتمالات در نظر گرفته و آنان را برآورد می‌کند. مزیت روش GME این است که در این روش لازم نیست هیچ‌گونه فرض یا توزیع احتمالی را برای باقیمانده‌ها (حتی در مواقعی که نمی‌خواهیم استنباط آماری انجام دهیم) در نظر بگیریم. در رگرسیون متغیر کیفی نیز این روش کاربرد خود را پیداکرده است. مرکانی و همکاران [۱] روش ماکسیمم آنتروپی تعمیم‌یافته را در رگرسیون لوژستیک اعمال کرده‌اند. همان‌طور که پیش‌تر گفته شد مدت‌های طولانی است که رگرسیون ستیغی به‌عنوان راه‌حلی برای مشکل هم خطی و برآورد پارامترهای مدل رگرسیونی مطرح‌شده است. ماسدو و همکاران [۱۱] بر اساس روش GME برآوردگر ستیغی ماکسیمم آنتروپی تعمیم‌یافته را ارائه داده‌اند. اشکال اساسی اغلب روش‌های سنتی که برای برآورد پارامتر ستیغی به‌کاررفته‌اند، وابستگی پارامترهای مجهولی است که از روی داده‌ها باید برآورد شوند اما در روش ماکسیمم آنتروپی تعمیم‌یافته، این وابستگی از بین می‌رود چون این روش مستلزم برقراری هیچ فرضی نیست. GME پارامترهای مدل رگرسیونی را بدون اینکه هیچ محدودیتی روی جملات احتمالی و خطا در نظر بگیرد برآورد می‌کند. اما بعد از شانون [۱۵] تاکنون اندازه‌های آنتروپی تعمیم‌یافته زیادی معرفی گردید که هرکدام در حالت خاص آنتروپی شانون را در بر می‌گیرند. از جمله پرکاربردترین آن‌ها آنتروپی تی سالیس [۱۹] است. اصل ماکسیمم آنتروپی که تکیه‌بر آنتروپی شانون دارد توزیع‌های نمایی را نتیجه می‌دهد. آن‌گونه که مطالعات تجربی نشان می‌دهد در بسیاری از مواقع توزیع جامعه تحت بررسی از نوع توزیع‌های توانی بوده (بسیاری از مدل‌های اقتصادی) بنابراین استفاده از آنتروپی تی سالیس به‌عنوان یک رده کلی‌تر، به‌جای آنتروپی شانون پیشنهادشده و موضوع آنتروپی تی سالیس ماکسیمم بررسی‌شده است. بر این اساس می‌توان با جایگزین کردن آنتروپی تی سالیس به‌جای آنتروپی شانون در روش ماکسیمم آنتروپی تعمیم‌یافته، این روش را برای آنتروپی مرتبه α تعمیم داده و رده دیگری از برآوردها به نام $GME\alpha$ را معرفی کرد؛ به عبارت دیگر مجموع آنتروپی تی سالیس مرتبه α دو جمله احتمالی و جمله خطا را ماکسیمم کنیم. با توجه به کاربردهای زیاد آنتروپی تی سالیس و به‌خصوص آنتروپی تی سالیس مرتبه دو در بسیاری از علوم از جمله شیمی، فیزیک، زیست آزمایشگاهی و اقتصاد صانعی و محتشمی [۱۶] برآوردگرهای ماکسیمم آنتروپی تی سالیس تعمیم‌یافته، معرفی کردند. آن‌ها با قرار دادن $\alpha = ۲$ روش ماکسیمم آنتروپی تعمیم‌یافته را با قرار دادن آنتروپی تی سالیس مرتبه دو به جای آنتروپی شانون تعمیم داده و مجموع آنتروپی تی سالیس مرتبه دو هر دو جمله آنتروپی و جمله خطا را ماکسیمم کرده‌اند. برآوردهای به‌دست‌آمده از این روش را برآوردگرهای ماکسیمم آنتروپی تی سالیس مرتبه دو تعمیم‌یافته ($GMET۲$) نامیده‌اند. به دلیل اهمیت مسئله هم خطی صانعی و محتشمی [۱۷] به‌کمک برآوردگرهای $GMET۲$ برآورد جدیدی از پارامتر ستیغی ارائه دادند. آن‌ها با ترکیب اثر ستیغی و برآوردگر ماکسیمم آنتروپی تی سالیس مرتبه دو تعمیم‌یافته، برآورد جدیدی از پارامتر ستیغی به نام برآورد ستیغی ماکسیمم آنتروپی تی سالیس مرتبه دو تعمیم‌یافته $Ridge - GMET۲$ معرفی کردند. در ادامه این مقاله بعد از معرفی روش $GMET$ و برآوردگرهای $GMET۲$ برای یک مدل رگرسیونی که از هم خطی بالایی برخوردار است برآوردگرهای $GMET۲$ را محاسبه و با برآوردگرهای OLS و GME مقایسه می‌کنیم. در ادامه برای یک مجموعه داده که از هم خطی بالایی برخوردار هستند این روش را اعمال نموده و برآوردگر ستیغی ماکسیمم آنتروپی تی سالیس مرتبه دو تعمیم‌یافته $Ridge - GMET۲$ را محاسبه می‌کنیم.

۲ برآوردگر رگرسیون ستیغی

در این بخش برآوردگر رگرسیون ستیغی[3] موردنظر قرار می‌گیرد. مدل رگرسیون خطی

$$Y = X\beta + e \qquad (۱)$$

که در رابطه (۱) ، Y بردار $N \times ۱$ از مشاهدات و β یک بردار $K \times ۱$ از پارامترهای مجهول و X یک ماتریس $(N \times K)$ از متغیرهای توضیحی و معلوم است. e نیز بردار $(N \times ۱)$ خطاهای تصادفی است. به‌طوری‌که

$$E(e|X) = ۰, E(ee'|X) = \sigma^۲ I$$

که I ماتریس همانی $(N \times N)$ است. برآوردگر کمترین توان‌های دوم β به‌صورت

$$\hat{\beta}_{OLS} = (X'X)^{-۱}X'Y \qquad (۲)$$

و برآوردگر رگرسیون ستیغی که توسط ارل و کنارد [۸] پیشنهاد شد عبارت است از:

$$\hat{\beta}_{RR} = (X'X + \eta I)^{-۱}X'Y \qquad (۳)$$

[3] Ridge regression

که $\eta \geq ۰$ نشان‌دهنده پارامتر ستیغی است. وقتی‌که $\eta \to ۰$ برآوردگر RR به برآوردگر OLS نزدیک می‌شود درحالی‌که وقتی $\eta \to \infty$ برآوردگر RR به بردار صفر میل می‌کند. پس یک افزایش و کاهش بین واریانس و اریبی لازم است. آن‌ها نشان دادند که برآوردگر RR برای یک دامنه مقادیر η (وقتی‌که ملاک مقایسه دو برآوردگر MSE باشد) یک کران بالا برای برآوردگر کمترین توان‌های دوم است. مقدار MSE برآوردگر RR عبارت است از:

$$MSE(\hat{\beta}_{RR}) = \sigma^۲ \sum_{i=۱}^{k} \frac{\lambda_i}{(\lambda_i+\eta)^۲} + \eta^۲ \beta^{'} (X^{'}X+\eta I)^{-۲}\beta \quad (۴)$$

که در (۴) λ_i ها مقادیر ویژه ماتریس همبستگی $X^{'}X$ هستند.

۳ برآوردگر ستیغی ماکسیمم آنتروپی تی سالیس مرتبه دو تعمیم‌یافته

آنتروپی تی‌سالیس مرتبه α یکی از اشکال تعمیم‌یافته آنتروپی شانون است که وقتی $\alpha \to ۱$ همان آنتروپی شانون را نتیجه می‌دهد. به ازای $\alpha = ۲$ آنتروپی تی سالیس مرتبه ۲ حاصل شده که کاربردهای خاص خودش را در بسیاری از علوم پیداکرده است.

تعریف ۱.۳. *برای متغیر تصادفی X با تابع احتمال مربوطه $P = (p_۱, ...p_n)$ آنتروپی تی‌سالیس و آنتروپی تی سالیس مرتبه دو به ترتیب عبارت‌اند از:*

$$T_\alpha(P) = \frac{۱}{\alpha - ۱}[\sum_{i=۱}^{n} p^\alpha(x_i) - ۱], \alpha > ۰, \alpha \neq ۱$$

$$T_۲(P) = \sum_{i=۱}^{n} p^۲(x_i) - ۱ = \sum_{i=۱}^{n} p(x_i)[p(x_i) - ۱]. \quad (۵)$$

روش GME بر اساس آنتروپی شانون تعریف‌شده است. [۱۶] ضمن جایگزینی آنتروپی شانون با آنتروپی تی سالیس مرتبه دو روش GME را تعمیم داده و یک خانواده جدید از برآوردگرهای نظریه اطلاع معرفی و آن را برآوردگرهای ماکسیمم آنتروپی تی سالیس تعمیم‌یافته ($GMET۲$) [۴] نام‌گذاری کردند. روش یافتن این برآوردگرها همانند برآوردگر GME بوده با این تفاوت که به‌جای آنتروپی شانون از آنتروپی تی سالیس مرتبه دو استفاده می‌شود. در این مقاله بر اساس برآوردگر $GMET۲$ برآوردگر دیگری برای پارامتر ستیغی معرفی کرده و آن را با سایر برآوردگرهای ستیغی ذکرشده در بخش قبل مقایسه می‌کنیم. بدین منظور ابتدا روش یافتن برآوردگرهای $GMET۲$ توضیح داده می‌شود.

۱.۳ برآوردگر ماکسیمم آنتروپی تی سالیس مرتبه دو تعمیم‌یافته

برای به دست آوردن برآوردگر $GMET۲$ همانند روش یافتن GME، مدل خطی (۱) را دوباره بازنویسی می‌کنیم به‌طوری‌که برآوردها و خطاهای مجهول به فرم احتمالات نوشته می‌شوند زیرا عناصر تابع ماکسیمم آنتروپی، احتمالات هستند. فرض کنید هر β_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 به‌طوری‌که $\sum_{m=۱}^{M} p_{km} = ۱$. بر این اساس هر پارامتر β_k را می‌توان به‌صورت زیر نوشت:

$$\beta_k = \sum_{m=۱}^{M} z_{km} p_{km}, \qquad k = ۱, ۲, ..., K, \quad , M \geq ۲, \quad (۶)$$

تکیه‌گاه پارامتری یا مقادیر بردار Z_k بر اساس اطلاع پیشین و یا تجارب قبلی تعیین می‌شود. به طور مشابه برای برآورد توزیع خطاها نیز می‌توان به‌این‌ترتیب عمل کرد که فرض می‌کنیم تکیه‌گاه خطا شامل $J \geq ۲$ نقطه و ν نیز یک ماتریس $(T \times TJ)$ از نقاط تکیه‌گاه بوده و W نیز یک بردار $TJ \times ۱$ از وزن‌های متناظر نقاط تکیه‌گاه باشند به‌طوری‌که $\sum_{j=۱}^{J} w_{tj} = ۱$. در این صورت هر بردار e_t به‌صورت زیر بازنویسی می‌شود:

$$e_t = \sum_{j=۱}^{J} \nu_{tj} w_{tj}, \quad t = ۱, ۲, ..., T, J \geq ۲, \qquad (۷)$$

برای تعیین کران‌های تکیه‌گاه خطاها، گلان و همکاران [۶] از قانون ۳σ که توسط پولکیشم [۱۴] مطرح‌شده بود استفاده کردند. آن‌ها کران پایین را -۳σ قرار داده و برای کران بالا از $+۳\sigma$ استفاده کردند که مقدار σ نیز همان انحراف معیار نمونه‌ای بردار Y می‌باشد. بر این اساس هر پارامتر β_k و e_t به ترتیب به‌صورت (۶) و (۷) بازنویسی می‌شوند. پس مسئله ماکسیمم آنتروپی تی سالیس تعمیم‌یافته به‌صورت:

$$\begin{aligned} \max_{P,W} H(P,W) &= -P'(P' - ۱) - W'(W' - ۱) \\ &= -\sum_{k=۱}^{K}\sum_{m=۱}^{M} p_{km}(p_{km} - ۱) \\ &\quad - \sum_{t=۱}^{T}\sum_{j=۱}^{J} w_{tj}(w_{tj} - ۱), \end{aligned} \qquad (۸)$$

به‌طوری‌که

[4] Generalized maximum Tsallis entropy

$$\begin{cases} \sum_{k=1}^{K}\sum_{m=1}^{M} x_{tk}z_{km}p_{km} + \sum_{j=1}^{J} w_{tj}\nu_{tj} = Y_t, & t = 1,2,...,T. \\ \sum_{m=1}^{M} p_{km} = 1, & k = 1,2,...,K. \\ \sum_{j=1}^{J} w_{tj} = 1, & t = 1,2,...,T. \end{cases} \quad (9)$$

است. برای حل مسئله بهینه‌سازی (۸) با توجه به محدودیت‌های ذکرشده معادله لاگرانژ را تشکیل داده که حل این معادله نتایج زیر:

$$\hat{p}_{km} = \frac{1-M}{M} + \frac{1}{2M}\sum_{t=1}^{T}\lambda_t x_{tk}\sum_{m=1}^{M} z_{km} - \frac{1}{2}\sum_{t=1}^{T}\lambda_t x_{tk} z_{km}, \quad (10)$$

$$m = 1,2,...,M, k = 1,2,...,K.$$

$$\hat{w}_{tj} = -\frac{1}{2}\lambda_t\nu_{tj} + \frac{1}{2J}\lambda_t(\sum_{j=1}^{J}\nu_{tj}), \quad j = 1,...,J.$$

را داریم. با جایگذاری جواب‌های $\hat{p}_{km}$ و $\hat{w}_{tj}$ به ترتیب در هر یک از معادلات (۶) و (۷)برآوردهای ماکسیمم آنتروپی تی‌سالیس تعمیم‌یافته $GMET2$ به‌صورت زیر حاصل می‌شوند:

$$\beta_k^{GMET2} = \sum_{m=1}^{M}\hat{p}_{km}z_{km}, k = 1,2,...,K,$$

$$e_t^{GMET2} = \sum_{j=1}^{J}\hat{w}_{tj}\nu_{tj}, t = 1,2,...,T.$$

۲.۳ برآوردگر ستیغی ماکسیمم آنتروپی تی‌سالیس مرتبه دو تعمیم‌یافته

بر اساس برآوردگر $GMET2$ که در بالا به دست آمد، برآوردگر ستیغی ماکسیمم آنتروپی تی‌سالیس تعمیم‌یافته[5] ($Ridge-GMET2$) به‌صورت:

$$\hat{\eta}_{Ridge-GMET2} = min_{\eta\in[a_1,a_2]}||Z\hat{P} - (X'X + \eta I)^{-1}X'Y||_\infty \quad (11)$$

تعریف می‌شوند. بردار $\hat{P}$ در رابطه (۱۱)، همان بردار رابطه (۱۰) به‌دست‌آمده از روش $GMET2$ که رابطه (۸) را ماکسیمم می‌کند است. بازه ستیغی در نظر گرفته‌شده بازه $[a_1, a_2]$ است که در آن a_1 و a_2 به ترتیب کران‌های بالا و پایین این بازه هستند.

۴ مثال عددی

در این بخش بر اساس برآوردگر ستیغی ماکسیمم آنتروپی تی سالیس مرتبه دو تعمیم‌یافته که در بخش قبل ارائه شد برآوردگر رگرسیون ستیغی را برای مجموعه داده سیمان پرتلند که از درجه هم خطی بالایی برخوردار است محاسبه می‌کنیم. با توجه به وجود هم خطی در این داده‌ها بعد از وود و همکاران [۲۲] برآوردهای مختلفی برای برآورد پارامتر ستیغی ارائه‌شده است از جمله می‌توان به هالد [۷] کاکرانلار، [۱۰]، لیو [۱۳] و مونیز و کیبریا [۱۲] اشاره کرد. در این مجموعه برای اندازه‌گیری میزان سختی سیمان پرتلند متغیر پاسخ Y (میزان گرمای آزادشده از هر گرم سیمان) تعریف‌شده است که خود تحت تأثیر چهار متغیر رگرسیونی X_1: (آلومینات تری کلسیم)، X_2: (سیلیکات تری کلسیم)، X_3 (آلومینو فریت تتراکلسیم) و X_4 (بتا دی کلسیم سیلیکات) است و به ترتیب چهار ستون ماتریس X را تشکیل می‌دهند. این مجموعه داده در زیر ارائه‌شده است.

مدل رگرسیونی

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e. \quad (12)$$

را به این مجموعه داده‌ها برازش داده و برآوردهای OLS محاسبه‌شده‌اند.

بر اساس اطلاعات جدول۱هیچ‌یک از ضرایب در سطح α مقدار۰/۰۵ معنی‌دار نبوده و مقدار $\hat{\sigma}^2$ برابر ۵/۹۸۲۹ به‌دست‌آمده است. همچنین مقدار R^2 برابر ۰/۹۸۲۴ است که نشان می‌دهد ۹۸/۲۴ درصد تغییرات توسط متغیرهای رگرسیونی پیشگو توضیح داده می‌شود. مقادیر ویژه برای ماتریس X در جدول۲آورده شده‌اند.

بنابراین عدد شرطی که جذر تقسیم کوچک‌ترین مقدار ویژه بر بزرگ‌ترین مقدار ویژه است، در این مجموعه داده برابر ۸۲/۷۷ است. بزرگ بودن عدد شرطی ناشی از وجود هم خطی زیاد بین متغیرهای رگرسیونی است. بازه ستیغی را به‌صورت [۰, ۱] تعریف کرده و برآوردگر ستیغی ماکسیمم آنتروپی تی سالیس مرتبه دو تعمیم‌یافته ($\hat{\eta}$) را به دست می‌آوریم. برای محاسبات از نرم‌افزار متلب کمک می‌گیریم. الگوریتم محاسبه برآوردگرهای $Ridge-GME$ و $Ridge-GMET2$ به این صورت است که ابتدا مقادیر برآوردهای GME و $GMET2$ را محاسبه

[5] Ridge Gneralized Maximum Tsallis Entropy

می‌شود . سپس با مینیمم کردن رابطه روی بازه ستیغی $[\circ,\text{۱}]$ به ازای مقادیر برآورد شده مقدار $\hat{\eta}$ را که این بازه را مینیمم می‌کند، تعیین می‌کنیم. به ازای این مقدار مینیمم کننده، برآورد $Ridge-GME$ و $Ridge-GMET\text{۲}$ محاسبه می‌شود. این برآوردها در جدول ۳ آورده شده‌اند. مقادیر MSE مربوط به هر سه نوع برآوردگر OLS ، $Ridge-GME$ و $Ridge-GMET\text{۲}$ در جدول ۴ خلاصه‌شده است.

$$X=\begin{bmatrix} ۷ & ۲۶ & ۶ & ۶۰\\ ۱ & ۲۹ & ۱۵ & ۵۲\\ ۱۱ & ۵۶ & ۸ & ۲۰\\ ۱۱ & ۳۱ & ۸ & ۴۷\\ ۷ & ۵۲ & ۶ & ۳۳\\ ۱۱ & ۵۵ & ۹ & ۲۲\\ ۳ & ۷۱ & ۱۷ & ۶\\ ۱ & ۳۱ & ۲۲ & ۴۴\\ ۲ & ۵۴ & ۱۸ & ۲۲\\ ۲۱ & ۴۷ & ۴ & ۲۶\\ ۱ & ۴۰ & ۲۳ & ۳۴\\ ۱۱ & ۶۶ & ۹ & ۱۲\\ ۱۰ & ۶۸ & ۸ & ۱۲ \end{bmatrix}, \sum_{i=۱}^{۴} x_i=\begin{bmatrix} ۹۹\\ ۹۷\\ ۹۵\\ ۹۷\\ ۹۸\\ ۹۷\\ ۹۷\\ ۹۸\\ ۹۶\\ ۹۸\\ ۹۸\\ ۹۸\\ ۹۸ \end{bmatrix}, Y=\begin{bmatrix} ۷۸/۵\\ ۷۴/۳\\ ۱۰۴/۳\\ ۸۷/۶\\ ۹۵/۹\\ ۱۰۹/۲\\ ۱۰۲/۷\\ ۷۲/۵\\ ۹۳/۱\\ ۱۱۵/۹\\ ۸۳/۸\\ ۱۱۳/۳\\ ۱۰۹/۴ \end{bmatrix}$$

جدول ۱: جدول آنالیز واریانس داده‌های سیمان پرتلند به روش کمترین توان‌های دوم

جمله	پارامتر	خطای معیار	آماره‌های توزیع تی	مقادیر بحرانی آزمون
$\beta_\circ$	۶۲/۴۰۵۴	۷۰/۰۷۱۰	۰/۸۹	۰/۳۹۹۱
$\beta_۱$	۱/۵۵۱۱	۰/۷۴۴۸	۲/۰۸	۰/۰۷۰۸
$\beta_۲$	۰/۵۱۰۲	۰/۷۲۳۸	۰/۷۰	۰/۵۰۰۹
$\beta_۳$	۰/۱۰۱۹	۰/۷۵۴۷	۰/۱۴	۰/۸۹۵۹
$\beta_۴$	−۰/۱۴۴	۰/۷۰۹۱	−۰/۲	۰/۸۴۴۱

جدول ۲: مقادیر ویژه

$\lambda_۱$	$\lambda_۲$	$\lambda_۳$	$\lambda_۴$	$\lambda_۵$
۲۱۱/۳۶۷۵	۷۷/۲۳۶۱	۲۸/۴۵۹۷	۱۰/۲۶۷۴	۰/۰۳۴۹

[۱۷] برای یک مجموعه داده شبیه‌سازی‌شده مقادیر میانگین مربعات خطا را برای هر سه نوع برآوردگر $Ridge-GME$ و $Ridge-GMET\text{۲}$ و برآورد حداقل مربعات محاسبه کرده و نتیجه این شده است که $Ridge-GMEt\text{۲}$ کمترین میانگین مربعات خطا را داراست. در این مقاله برای مجموعه داده یک مثال عددی مثال داده‌های سیمان پرتلند $ridge-GMET\text{۲}$ محاسبه و با دو نوع برآوردگر دیگر مقایسه می‌شود.

جدول ۳: برآوردهای OLS، $Ridge-GME$، $Ridge-GMET۲$

پارامتر	OLS	$Ridge-GME$	$Ridge-GMET۲$
$\beta_۰$	۶۲/۴۰۵	۳۱/۲۳۳۹	۳۴/۲۷۵۷
$\beta_۱$	۱/۵۵۱	۱/۸۶۸۵	۱/۸۴۰۵
$\beta_۲$	۰/۵۱۰	۰/۸۳۲۴	۰/۸۰۰۱
$\beta_۳$	۰/۱۰۲	۰/۴۲۷۱	۰/۳۹۷۹
$\beta_۴$	−۰/۱۴۴	۰/۱۷۱۴	۰/۱۴۰۱

جدول ۴: مقادیرMSE برآوردهای مختلف ستیغی .

OLS	$Ridge-GME$	$Ridge-GMET۲$
۵/۹۸	۶/۹۸۰۷	۷/۴۲۶۱

۵ بحث و نتیجه‌گیری

علیرغم توانایی بالای روش ماکسیمم آنتروپی، روش ماکسیمم آنتروپی تعمیم‌یافته با چارچوب رگرسیونی مطرح شد. وجود هم خطی بالا بین متغیرهای توضیحی یکی از مشکلات عمده‌ای است که در تحلیل رگرسیونی با آن مواجه می‌شویم. یکی از روش‌هایی که برای مقابله با این مشکل مطرح گردیده است، روش رگرسیون ستیغی است. محققین برآوردهای زیادی برای پارامتر ستیغی ارائه داده‌اند. در این مقاله بر اساس برآورد ماکسیمم آنتروپی تی‌سالیس مرتبه دو تعمیم‌یافته برآورد دیگری برای پارامتر ستیغی یافته و آن را برآورد ستیغی ماکسیمم آنتروپی تی سالیس مرتبه دو تعمیم‌یافته نامیده‌ایم. قبلاً به کمک شبیه‌سازی هر سه برآوردگر ذکرشده در جدول ۴ را محاسبه و نشان دادیم که میانگین مربعات خطا برای برآوردگر ستیغی ماکسیمم آنتروپی تی سالیس مرتبه دو تعمیم‌یافته از دو برآوردگر دیگر کمتر است نتایج کامل در [۱۷] ارائه‌شده است. یک مجموعه داده سه نوع برآوردگر ستیغی ماکسیمم آنتروپی تعمیم‌یافته، حداقل مربعات و برآوردگر ستیغی ماکسیمم آنتروپی تی سالیس مرتبه دو تعمیم‌یافته را محاسبه و مقادیر MSE شان را برآورد کرده‌ایم اگرچه مقدار MSE روش OLS نسبت به دو روش $Ridge-GME$ و $Ridge-GMET۲$ کمتر به‌دست‌آمده ولی این تفاوت ناچیز و با توجه به اینکه مشکل هم خطی در داده‌ها وجود دارد و از طرف دیگر کوچک بودن حجم نمونه در این مثال (کارایی روش GME و نیز $GMET۲$ در حالتی که حجم نمونه کوچک است در [۱۶] موردبحث و بررسی قرار داده‌ایم) و روش ماکسیمم آنتروپی تعمیم‌یافته مستلزم برقراری فرضیات زیربنایی نیست این اختلاف ناچیز قابل‌چشم‌پوشی بوده و این دو برآوردگر جدید را ترجیح می‌دهیم.

مراجع

[۱] مرکانی، مهسا. صانعی طبس ، منیژه. نادری، حبیب. احمدزاده، حامد و جمالزاده، جواد. برآورد پارامترهای مدل رگرسیونی لوژستیک به کمک ماکسیمم آنتروپی تعمیم‌یافته، اندیشه آماری، سال بیست و ششم، شماره ۲ پاییزو زمستان ، شماره ۵۲، ۱۴۰۰، ص ۱−۸

[2] Akdeniz, F., Cabuk, A. and Guler, H. (2011), Generalized maximum entropy estimators: applications to the Portland Cement data set. *The Open Statistics and Probability Journal*, **3**, 13-20.

[3] Belsley, D.A. (1982), Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise. *Journal of Econometrics*, **20**, 211-253.

[4] Farrar, d. E. and Glauber, R. R. (1967). Multicollinearity in Regression analysis: the problem revisited. *Review of Economics and Statistics*, **49**, 92-107.

[5] 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.

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

[7] Hald, A. (1952), *Statistical Theory with Engineering Applications*, John Wiley Sons, New York.

[8] Horel, A. E. and Kenard. R. W. (1970), Ridge Regression: Applications to Nonorthogonal Problems, *Technometrics*, **12**, 69-82.

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

[10] Kacranlar, S., Sakalliogglu, S., Akdeniz, F., Styan, G. P. H. and Werner, H. J. (1999), A new biased estimator in linear regression and a detailed analysis of the widely-analysed dataset on Portland cement, *The Indian Journal of Statistics, Series B, Senkhiya*, **61**, 443–459.

[11] Macedo, P., Scotto, M. and Silva, E. (2010), On the Choice of the Ridge Parametr: A Maximum Entropyy Approach, *Communication in Statistics Simulations and Computations*, **39**, 1628-1638.

[12] Muniz, G. and Kibria, B. M. G. (2009), On some ridge regression estimator: an emperical comparisons, *Communication in Statistics Simulations and Computations*, **38**, 621-630.

[13] Liu, K. (2003), Using Liu-type estimator to combat collinearity, *Communications in Statistics– Theory and Methods*, **32**, 1009–1020.

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

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

[16] Sanei Tabass, M. and Mohtashami Borzadaran, GH. R. (2017), The Generalized Maximum Tsallis Entropy Estimators and Applications to the Portland Cement Data set, *Communications in Statistics - Simulation and Computation* , **46**, 3284-3293.

[17] Sanei Tabass, M. and Mohtashami Borzadaran, GH. R. (2022), On the choice of the Ridge parameter: a generalized maximum Tsallis entropy approach, *Communications in Statistics - Simulation and Computation* , DOI: 10.1080/03610918.2022.2082475.

[18] Thisted, R.A. (1982). Decision-theoretic regression diagnostics. *Statistical Decision Theory and Related Topics III*, **2**, 363-382.

[19] Tsallis, C. (1988). Possible generalizations of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, **52**, 479-487.

[20] Renyi, A. (1961). On measures of entropy and information. *Proc. Berekeley Symposium, Statist. Probability*, **1**, 547-561.

[21] Willan. A.R. and D.G. Watts. (1978). Meaningful multicollinearity measures. *Technometrics*, **20**, 407-412.

[22] Woods, H. Steinor, H. and Starke, H. (1932). Effect of composition of Portland cement on heat evolved during hardening. *Industerial and Engineering Chemistry*, **24**, 1207-1214.

Using the generalized maximum Tsallis entropy for estimating the Ridge regression parameter

Manije Sanei Tabas[1]

Abstract:

Regression analysis using the method of least squares requires the establishment of basic assumptions. One of the problems of regression analysis in this way faces major problems is the existence of collinearity among the regression variables. Many methods to solve the problems caused by the existence of the same have been introduced linearly. One of these methods is ridge regression. In this article, a new estimate for the ridge parameter using generalized maximum Tsallis entropy is presented and we call it the Ridge estimator of generalized maximum Tsallis entropy. For the cement dataset Portland, which have strong collinearity and since 1332, different estimators have been presented for these data, this estimator is calculated and We compare the generalized maximum Tsallis entropy ridge estimator, generalized maximum entropy ridge estimator and the least squares estimator.

Keywords: Ridge Regression, Generalized Maximum Entropy, Tsallis Entropy, Generalized Maximum Tsallis Entropy.

[1] **foot print**