انتخاب متغیر با استفاده از تابع تاوان

هادی موقری [۱]، سیدمحمدابراهیم حسینی‌نسب [۲]

چکیده:

علی‌رغم کاربرد گسترده‌ی معیارهایی نظیر AIC و BIC در زمینه‌ی انتخاب متغیر، استفاده‌ی از آنها سبب بی‌ثباتی می‌شود؛ یعنی تغییرات اندک در داده‌ها ممکن است نتیجه‌ی انتخاب متغیر را به طور فاحشی تغییر دهد. از روش‌هایی که اخیراً در زمینه‌ی انتخاب متغیر بسیار مورد توجه قرار گرفته است، استفاده از توابع تاوان می‌باشد. تابع تاوان SCAD، به دلیل ویژگی‌های بهینه‌ی برآوردگر حاصل، یکی از مقبول‌ترین توابع تاوان در زمینه‌ی انتخاب متغیر می‌باشد. در این مقاله علاوه بر معرفی اجمالی برخی از توابع تاوان و بسته‌های نرم‌افزاری مرتبط با آنها، تابع تاوان SCAD و عملکرد آن برای انتخاب متغیر را مورد ارزیابی قرار می‌دهیم. در انتها نیز به تحلیل یک مجموعه داده‌ی واقعی پرداخته و نتایج را گزارش می‌کنیم.

واژه‌های کلیدی: انتخاب متغیر، تابع تاوان SCAD، AIC، BIC، LASSO.

۱ مقدمه

پیشرفت روزافزون علوم مختلف و به وجود آمدن تکنیک‌های جدید این امکان را فراهم کرده است تا بتوان حجم وسیعی از داده‌ها در مورد متغیرهای مختلف را بدست آورد. وجود این داده‌های حجیم در اکثر شاخه‌ها از مطالعات بهداشتی گرفته تا مدیریت ریسک تفکر آماری را به طور کلی دگرگون کرده است. در این گونه موارد انتخاب متغیر[۳]، خود را به عنوان یک ابزار بسیار توانمند و حیاتی برای زمینه‌های اکتشافی در علوم مختلف مطرح کرده است. فن و لی [۹] به طور مفصل به بررسی اهمیت انتخاب متغیر در بسیاری از زمینه‌های علمی پرداختند.

در بسیاری از مسایل عملی، تحلیل‌گر با تعداد زیادی از متغیرهای توضیحی مواجه است که بنا به ضرورت باید تنها تعداد محدودی از آنها را جهت حضور در مدل انتخاب کند. زیرا اگر مدل حاوی متغیرهای زاید و بی‌اهمیت باشد یا به اصطلاح دچار بیش‌برآوردی [۴] گردد، علاوه بر آنکه تعبیر و تفسیر چنین مدلی مشکل خواهد شد جمع‌آوری مشاهدات برای چنین مدلی مستلزم صرف وقت و هزینه‌ی زیاد می‌باشد. برعکس اگر مدل دچار کم‌برآوردی [۵] گردد در این صورت مدل حاصل، ممکن است داده‌های موجود را به خوبی توصیف نکند

[۱] دانشکده علوم ریاضی، دانشگاه تربیت مدرس
[۲] دانشکده علوم ریاضی، دانشگاه شهید بهشتی
[۳] Variable Selection
[۴] Overestimation
[۵] Underestimation

(برای آشنایی با اثرات تخصیص مدل نادرست به رنچر و شالج [۱۷]، صفحات ۱۶۹–۱۷۸ مراجعه شود). پیدا کردن مدلی که گرفتار مسایل بیش‌برآوردی و کم‌برآوردی نباشد در حیطه‌ی مربوط به انتخاب متغیر قرار می‌گیرد.

مدل رگرسیون خطی زیر را در نظر بگیرید:

$$y_i = \beta_1 x_{i1} + \ldots + \beta_d x_{id} + \epsilon_i, \quad i = 1, \ldots, n, \quad (1)$$

به طوریکه $x_{id}, \ldots, x_{i1}, y_i$ به ترتیب متغیر پاسخ و d متغیر توضیحی مرتبط با آن در نمونه‌ی iام می‌باشند. علاوه بر آن، $\beta_1, \ldots, \beta_d$ ضرایب رگرسیونی و ϵ_iها نیز خطای تصادفی مدل و مستقل از یکدیگر با میانگین صفر و واریانس σ^2 می‌باشند. هدف از انتخاب متغیر، تعیین زیر مجموعه‌ای از متغیرهای توضیحی $x_1, \ldots, x_d$ است به گونه‌ای که در توصیف متغیر پاسخ y بهترین عملکرد را در مقابل سایر مدل‌های زیر مجموعه داشته باشد. تا کنون معیارها و روش‌های متنوعی برای انتخاب متغیر ارایه شده‌اند که معروفترین آنها معیار اطلاع آکائیک (AIC[6]) و معیار اطلاع بیزی (BIC[7]) است که به ترتیب توسط آکائیک [۲] و شوارتز [۱۸] معرفی شدند. علی‌رغم کاربرد گسترده‌ای که این معیارها در زمینه‌ی انتخاب متغیر دارند استفاده‌ی از آنها باعث بی‌ثباتی در انتخاب متغیر می‌شود [۳]. زیرا در صورت استفاده از این معیارها در مورد هر ضریب رگرسیونی تنها قادر به اتخاذ دو تصمیم هستیم؛ یا متغیر مرتبط با آن ضریب در مدل حضور پیدا می‌کند یا نه و این باعث می‌شود زمانی که مجموعه‌ی داده‌ها دستخوش تغییرات اندک می‌شود، نتیجه‌ی انتخاب مدل به طرز فاحشی تغییر کند. یک راه غلبه بر این مشکل استفاده از تابع تاوان[8] برای انتخاب متغیر می‌باشد. لازم به ذکر است که معیارهای اطلاع نیز به نوعی با مفهوم تاوانیدن در ارتباط هستند [۱۲]. در صورت استفاده از توابع تاوان دو فرآیند انتخاب مدل و برآورد ضرایب رگرسیونی به طور همزمان صورت می‌پذیرد در حالی که این نکته در مورد معیارهایی نظیر AIC صادق نیست و این باعث می‌شود برای مدل‌های با بعد نسبتاً بزرگ، فرآیند انتخاب متغیر بسیار وقت‌گیر باشد.

بخش دوم این مقاله به رابطه‌ی بین معیارهای اطلاع و مفهوم تاوانیدن می‌پردازد. بخش سوم به معرفی انواع توابع تاوان و امکانات نرم افزاری موجود برای استفاده از آنها اختصاص دارد. تابع تاوان SCAD[9] در بخش چهارم معرفی می‌شود و در ادامه نیز روش‌های انتخاب پارامتر نظم در بخش پنجم بیان می‌شوند. همچنین مطالعات شبیه‌سازی، شامل مقایسه‌ی بین انواع روش‌های انتخاب متغیر و نیز بررسی تاثیر پارامتر نظم بر عملکرد تابع تاوان در بخش ششم آورده می‌شود. سرانجام در بخش هفتم روش‌های معرفی شده را بر روی یک مجموعه داده‌ی واقعی اجرا می‌کنیم.

۲ معیار اطلاع و مفهوم تاوانیدن

تعداد زیادی از معیارهای انتخاب مدل بر اساس نظریه‌ی اطلاع پدید آمده‌اند. یکی از پرکاربردترین این معیارها،

[6] Akaike Information Criterion
[7] Bayesian Information Criterion
[8] Penalty Function
[9] Smoothly Clipped Absolute Deviation

AIC است که به صورت زیر معرفی می‌شود [۲]:

$$AIC = -۲\ell + ۲p. \qquad (۲)$$

در رابطه‌ی فوق، ℓ لگاریتم تابع درستنمایی و p تعداد ضرایب رگرسیونی موجود در مدل می‌باشند. هر مدل زیر مجموعه با AIC کمتر، بر سایر مدل‌های زیر مجموعه برتری دارد. به دنبال معرفی AIC معیارهای متعدد دیگری پدید آمدند که معروفترین آنها معیار اطلاع شوارتز (SIC) یا معیار اطلاع بیزی (BIC) است و به صورت زیر تعریف می‌شود [۱۸]:

$$BIC = -۲\ell + p\ln n. \qquad (۳)$$

در عبارت فوق n حجم نمونه است. برای آشنایی با سایر معیارهای اطلاع و نقش نظریه‌ی اطلاع در معرفی آنها به بورنهام و اندرسون [۴] مراجعه کنید. اگر جملات خطا در مدل (۱) دارای توزیع نرمال باشند معیارهای (۲) و (۳)، به صورت مجموع توان‌های دوم باقیمانده‌ی تاوانیده‌ی زیر حاصل می‌شوند (با فرض معلوم بودن $\sigma^۲$):

$$\frac{RSS_p}{\sigma^۲} + pF, \qquad (۴)$$

که در آن RSS_p مجموع توان‌های دوم باقیمانده‌ی مدل با بعد p می‌باشد [۱۲]. F که یک مقدار از پیش تعیین شده است، میزان تاوانی است که برای بعد مدل در نظر گرفته می‌شود. واضح است که معیار اطلاع آکائیک و بیزی را می‌توان از رابطه‌ی (۴) به ترتیب با قرار دادن $F = ۲$ و $F = \ln n$ به دست آورد.

۳ معرفی چند تابع تاوان

روش معمول برای برآورد ضرایب رگرسیونی در مدل (۱)، استفاده از توان‌های دوم باقیمانده می‌باشد که منجر به $\hat{\beta}_{\rm ols} = (\boldsymbol{X}'\boldsymbol{X})^{-۱}\boldsymbol{X}'\boldsymbol{y}$ می‌گردد که در آن $\boldsymbol{X}$ و $\boldsymbol{y}$ به ترتیب ماتریس مشاهدات و بردار متغیر پاسخ هستند. اما زمانی که بین متغیرهای توضیحی همبستگی وجود داشته باشد، این برآوردگر با نقایصی همراه خواهد شد که مهمترین آنها فاصله گرفتن $\hat{\beta}_{\rm ols}$ از مقدار واقعی β و در نتیجه بزرگ شدن طول بردار $\hat{\beta}_{\rm ols}$ است [۱۳]. برای رفع این مشکل هورل و کنارد [۱۳] استفاده از تابع هدف زیر را جهت برآورد ضرایب رگرسیونی پیشنهاد دادند:

$$\min_{\boldsymbol{\beta}}\{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\} \quad s.t \quad \sum_{j=۱}^{d}\beta_j^۲ < c. \qquad (۵)$$

در رابطه‌ی فوق β بردار ضرایب رگرسیونی است. شرط $\sum_{j=۱}^{d}\beta_j^۲ < c$ با کوچک کردن ضرایب رگرسیونی به سمت مبدا، از بزرگ شدن طول بردار برآورد جلوگیری می‌کند. مساله‌ی می‌نیمم سازی مشروط فوق، معادل با می‌نیمم کردن تابع غیر شرطی زیر می‌باشد:

$$\min_{\boldsymbol{\beta}}\{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda\sum_{j=۱}^{d}\beta_j^۲\}. \qquad (۶)$$

در عبارت فوق، λ را که متناظر با c و با استفاده از مشاهدات بدست می‌آید (مونت‌گمری [۱۶]، صفحات ۳۴۳–۳۳۹)، پارامتر نظم[۱۰] می‌نامند. با استفاده از تابع (۶) برآوردگر ستیغی[۱۱] به صورت $\hat{\beta}_{\rm ridge} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-۱}\boldsymbol{X}'\boldsymbol{y}$ حاصل می‌شود که در آن $\boldsymbol{I}$ یک ماتریس واحد است. فرانک و فریدمن [۱۰]

[۱۰] tuning parametr
[۱۱] Ridge

با استفاده از ایده‌ی به کار رفته در رگرسیون ستیغی، خانواده‌ی توابع $L_\gamma = \sum_{j=1}^{d} |\beta_j|^\gamma$ را جهت برآورد β در تابع هدف (۶) معرفی کردند. واضح است که برآوردگر ستیغی به ازای $\gamma = 2$ حاصل می‌شود. اما توابع L_γ تنها به ازای $0 < \gamma \leq 1$ توانایی انتخاب متغیر را دارند ([۸] و [۱۹]) یعنی با برابر صفر قرار دادن ضرایب کم اهمیت و بی‌اهمیت سبب حذف آنها از مدل می‌شوند. بنابراین رگرسیون ستیغی نمی‌تواند به عنوان یک روش انتخاب متغیر مطرح باشد. تیب‌شیرانی [۱۹] تابع L_1 را جهت انتخاب متغیر در تابع هدف (۶) به کار برد و برآوردگر حاصل از آن را تحت عنوان LASSO[۱۲] معرفی نمود.

روش LASSO پنجره‌ی جدیدی را به روی مبحث انتخاب متغیر گشود به طوری که بسیاری از روش‌های انتخاب متغیر که بعد از سال ۱۹۹۶ ارایه شدند به نوعی با LASSO در ارتباط هستند و هر یک با هدف اصلاح یکی از نقاط ضعف LASSO ایجاد شدند. به عنوان مثال، به دلیل اینکه در روش LASSO ضرایب بزرگ نیز تاوانیده می‌شوند (به سمت مبدا منقبض می‌شوند)، ضرایب کوچکتر فرصت حضور در مدل را پیدا می‌کنند و این سبب بیش‌برآوردی می‌گردد. فن و لی [۸] برای رفع این نقیصه تابع تاوان SCAD را معرفی کردند (این تابع در بخش بعد معرفی می‌شود). مشکل عمده‌ی دیگر LASSO در ارتباط با مسایلی است که در آنها تعداد متغیرهای توضیحی (d) بسیار بزرگتر از حجم نمونه (n) است. در این حالت LASSO اجازه‌ی حضور حداکثر n متغیر را در مدل نهایی می‌دهد. این مشکل به وسیله‌ی زو و هستی [۲۲] رفع شده است. مشکل دیگر برآوردگر LASSO ناسازگاری آن است. زو [۲۱] شرط لازم برای سازگاری این برآوردگر را بیان و فرم اصلاح شده‌ای از برآوردگر LASSO را که دیگر با مشکل ناسازگاری دست به گریبان نباشد، ارایه نمود.

به موازات معرفی انواع توابع تاوان، الگوریتم‌ها و روش‌های مختلفی برای محاسبه‌ی برآوردگرهای فوق ارایه شدند. به عنوان مثال شوتینگ الگوریتم[۱۳] [۱۱] و رگرسیون کمترین زاویه [۱۴] [۷] جهت محاسبه‌ی برآوردگر LASSO مطرح شدند. الگوریتم تقریب درجه‌ی دوم موضعی (LQA[۱۵]) نیز در رابطه با تابع تاوان SCAD، توسط فن و لی [۸] ارایه شد. نرم افزارهای آماری نیز با ایجاد بسته‌های مختلف امکان اجرای اکثر این الگوریتم‌ها را فراهم کرده‌اند. برای مثال، بسته‌ی brdgrun و lasso در نرم‌افزار S-Plus به ترتیب با استفاده از شوتینگ الگوریتم و الگوریتم پیشنهادی تیب‌شیرانی [۱۹]، برای محاسبه‌ی برآوردگر LASSO ایجاد شدند. بسته‌ی lars در نرم‌افزار R، برای استفاده از روشِ رگرسیون کمترین زاویه ایجاد شد. روش PROC SCADLS نیز در نرم افزار SAS امکان اجرای الگوریتم LQA را برای تابع تاوان SCAD فراهم کرده است [۶].

---

[۱۲] Least Absolute Shrikage and Selection Opertor
[۱۳] Shooting Algorithm
[۱۴] Least Angle Regression
[۱۵] Local Quadratic Approximation

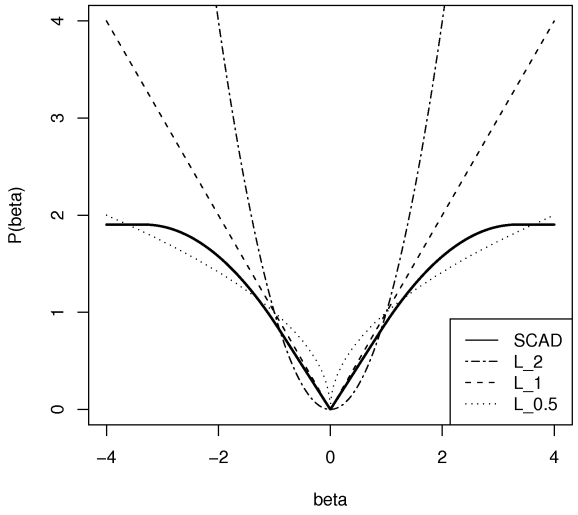
۴ تابع تاوان SCAD

فن و لی [۸] ویژگی‌های یک تابع تاوان خوب را به صورت (i) نااریبی، (ii) تُنکی[۱۶] و (iii) پیوستگی بیان کردند. هر کدام از این ویژگی‌ها در ارتباط با یکی از روش‌های معرفی شده‌ی قبلی می‌باشند. همانطور که در بخش سوم اشاره شد، روش LASSO با تاوانیدن ضرایب بزرگ سبب بیش‌برآوردی می‌گردد و در نتیجه LASSO از ویژگی نااریبی بی‌بهره است. رگرسیون ستیغی از ویژگی تنکی بی‌بهره است زیرا توانایی خارج کرن ضرایب بی‌اهمیت از مدل را ندارد. ویژگی سوم نیز به معیارهایی نظیر AIC اشاره دارد که استفاده‌ی از آنها موجب بی‌ثباتی در انتخاب متغیر می‌شود [۳]. دلیل این بی‌ثباتی نیز عدم استفاده از یک تابع تاوان پیوسته است [۸].

با توجه به این سه ویژگی، فن و لی [۸] تابع تاوان SCAD را به صورت زیر معرفی کردند:

$$P_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j| & 0 \le |\beta_j| < \lambda \\ \frac{(a^2-1)\lambda^2 - (|\beta_j|-a\lambda)^2}{2(a-1)} & \lambda \le |\beta_j| < a\lambda \\ \frac{(a+1)\lambda^2}{2} & |\beta_j| \ge a\lambda \end{cases}, \qquad (7)$$

به طوریکه $0 < \lambda$ و $a > 2$. پارامتر a شکل تابع تاوان SCAD را کنترل می‌کند به طوری که هر چه $a \to \infty$ شکل تابع SCAD به شکل L_1 میل می‌کند و عملکرد تابع تاوان SCAD مشابه عملکرد LASSO خواهد شد. نمودار SCAD به همراه نمودار توابع L_1، $L_{0.5}$ و L_2 در شکل ۱ نشان داده شده است. تابع تاوان SCAD برخلاف LASSO، برای ضرایبی که از یک مقدار به خصوص $(a\lambda)$ بزرگتر باشند تاوان ثابتی را در نظر می‌گیرد (شکل ۱ را ببینید.) . در تابع (۷) دو پارامتر a و λ وجود دارد. فن و لی [۸] با یک مطالعه‌ی بیزی نشان دادند که مقدار ۳/۷ برای a یک انتخاب مناسب است. به همین دلیل تابع SCAD را تنها با اندیس λ نشان‌گذاری می‌کنند.



شکل ۱. نحوه‌ی تاونیدن ضرایب، توسط توابع تاوان مختلف.

فن و لی [۸] برای برآورد β از مجموع توان‌های دوم باقیمانده‌ی تاوانیده‌ی زیر:

$$\frac{1}{n}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + \sum_{j=1}^{d} P_\lambda(|\boldsymbol{\beta}_j|), \qquad (8)$$

یا به طور معادل

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + n\sum_{j=1}^{d} P_\lambda(|\boldsymbol{\beta}_j|), \qquad (9)$$

استفاده کردند. که در آنها $P_\lambda(|\beta_j|)$ در (۷) تعریف شده است.

مشکل استفاده از تابع تاوان SCAD عدم مشتق‌پذیری آن در مبدا مختصات می‌باشد. به همین منظور فن و لی [۸] ابتدا تابع تاوان SCAD را با یک تابع درجه‌ی دوم تقریب زده و به دنبال آن با استفاده از الگوریتم تکراری نیوتن—

[۱۶] Sparsity

رافسون به می‌نیمم کردن تابع هدف (۸) و یا به طور متناظر تابع هدف (۹) پرداختند. در مورد ویژگی‌های همگرایی این الگوریتم که با عنوان تقریب درجه‌ی دوم موضعی (LQA) شناخته می‌شود، به هانتر و لی [۱۴] مراجعه کنید. با استفاده از این الگوریتم برآوردگر SCAD در تکرار $(k+1)$ام به صورت زیر حاصل می‌شود:

$$\beta^{(k+1)} = (\boldsymbol{X}'\boldsymbol{X} + n\boldsymbol{\Sigma}_\lambda(\beta^k))^{-1}\boldsymbol{X}'\boldsymbol{y}, \qquad (10)$$

که در آن $\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}^{(k)}) = diag\{\frac{P'_\lambda(|\beta_1^{(k)}|)}{|\beta_1^{(k)}|}, \ldots, \frac{P'_\lambda(|\beta_d^{(k)}|)}{|\beta_d^{(k)}|}\}$ و $P'_\lambda(|\boldsymbol{\beta}^{(k)}|)$ مشتق تابع تاوان SCAD در نقطه‌ی $\boldsymbol{\beta}^{(k)}$ می‌باشد. این برآوردگر با برآوردگر ستیغی شباهت زیادی دارد.

۵ انتخاب پارامتر نظم

اولین قدم در استفاده از هر تابع تاوانی انتخاب مقدار مناسب برای پارامتر نظم است. اگر مقدار پارامتر نظم خیلی بزرگ باشد متغیرهای بیشتری از مدل حذف می‌شوند که این، سبب کم برآوردی مدل می‌شود. برعکس، اگر مقدار پارامتر نظم خیلی کوچک باشد، باعث بیش برآوردی خواهد شد (جدول ۱ را ملاحظه کنید). برخلاف معیارهای اطلاع که در آن تاوان بُعد (F در رابطه‌ی ۴)، یک مقدار ثابت و از پیش تعیین شده است، مقدار پارامتر نظم در روش‌هایی مانند LASSO با استفاده از مشاهدات محاسبه می‌شود. در این بخش دو روش را که کاربرد فراوانی برای این منظور دارند، معرفی می‌کنیم.

۱.۵ اعتبارسنجی متقابل

در روش اعتبارسنجی متقابل (CV[۱۷]) مشاهدات به دو گروه تقسیم می‌شوند. با کمک گروه اول، مدل را برازش داده و توسط گروه دوم، مدل برازش شده، ارزیابی می‌شود. به عنوان مثال اگر n مشاهده به صورت (x_i, y_i) موجود باشند و با استفاده از $(n-1)$ تا از این مشاهدات، یک مدل برازش دهیم یا به عبارت دیگر تنها یکی از این مشاهدات را کنار بگذاریم، در این صورت معیار CV به صورت زیر تعریف می‌شود:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_\lambda^{-i})^2,$$

که در آن $\hat{\boldsymbol{y}}_\lambda^{-i}$ بیانگر مقدار برازش شده پس از کنار گذاشتن iامین مشاهده است. در این صورت λی مورد نظر متناظر با λیی خواهد بود که این معیار را می‌نیمم کند.

محاسبه‌ی $CV(\lambda)$ بسیار وقت‌گیر است. به همین دلیل کراون و واهبا [۵] معیار اعتبارسنجی متقابل تعمیم یافته (GCV[۱۸]) را به عنوان جایگزینی برای $CV(\lambda)$ به صورت زیر معرفی کردند:

$$GCV(\lambda) = \frac{(\boldsymbol{y} - \hat{\boldsymbol{y}}_\lambda)'(\boldsymbol{y} - \hat{\boldsymbol{y}}_\lambda)}{n(1 - \frac{tr(\boldsymbol{A}(\lambda))}{n})^2}, \qquad (11)$$

که در آن $\hat{\boldsymbol{y}}_\lambda$ بردار مقادیر برازش شده تحت ماتریس $\boldsymbol{A}(\lambda)$ است که با توجه برآوردگر SCAD به صورت زیر بدست می‌آید:

$$\begin{aligned}\hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\beta} \\ &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + n\boldsymbol{\Sigma}_\lambda)^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &\equiv \boldsymbol{A}(\lambda)\boldsymbol{y}. \qquad (12)\end{aligned}$$

[۱۷] Cross Validation
[۱۸] Generalized Cross Validation

ماتریس $\boldsymbol{A}(\lambda)$ را ماتریس تصویری[۱۹] $\boldsymbol{y}$ می‌گویند.

۲.۵ معیار BIC

لنگ و همکاران [۱۵] نشان دادند که اگر در روش LASSO پارامتر نظم با استفاده از معیارهایی نظیر CV و GCV انتخاب شود، برآوردگر حاصل ناسازگار خواهد بود. نکته‌ای مشابه با آن را ونگ و همکاران [۲۰] در مورد تابع تاوان SCAD بیان نمودند و نشان دادند که برآوردگر حاصل دچار بیش برآوردی می‌شود. آنها به جای بکارگیری معیار GCV، استفاده از معیار BIC را به صورت زیر پیشنهاد دادند:

$$BIC(\lambda) = -2\ell + DF_\lambda \ln n. \qquad (۱۳)$$

در اینجا نیز λی مورد نظر عبارت از λیی خواهد بود که به ازای آن معیار فوق می‌نیمم می‌شود. نکته‌ی حایز اهمیت در مورد معیار (۱۳)، چگونگی تعریف DF_λ می‌باشد. اگر چه ونگ و همکاران [۲۰]، DF_λ را به صورت $tr(\boldsymbol{A}(\lambda))$ تعریف کردند، با این حال اگر بخواهیم از تعریف معمول معیار BIC استفاده نماییم، DF_λ برابر تعداد ضرایب غیر صفر در برآوردگر SCAD می‌باشد [۶]. عملکرد این دو، در برآورد β تقریباً مشابه است. امکان استفاده از هر یک از این دو گزینه در روش PROC SCADLS از نرم‌افزار SAS فراهم شده است.

[۱۹] Projection Matrix
[۲۰] Median of Relative Model Error

۶ مطالعات شبیه سازی

در این بخش با استفاده از شبیه سازی به مقایسه‌ی برخی از روش‌های انتخاب مدل می‌پردازیم. داده‌های مورد نیاز برای انجام شبیه‌سازی‌ها را از مدل زیر تولید کردیم:

$$\boldsymbol{y} = \boldsymbol{X}\beta_\circ + \sigma\boldsymbol{\epsilon}, \qquad (۱۴)$$

به طوریکه $\beta_\circ = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ و عناصر ماتریس $\boldsymbol{X}$ و بردار $\boldsymbol{\epsilon}$ از توزیع نرمال استاندارد تولید شده‌اند. ضمن اینکه برای حادتر کردن شرایط، همبستگی بین متغیرهای توضیحی $\boldsymbol{X}_i$ و $\boldsymbol{X}_j$ را برابر $0/5^{|i-j|}$ قرار داده‌ایم. این شرایط را تیب‌شیرانی [۱۹] و فن و لی [۸] برای انجام شبیه سازی‌های خود به کار بردند. جهت مقایسه‌ی نتایج از سه شاخص Corr، Incorr و MRME[۲۰] استفاده نمودیم. Corr متوسط تعداد ضرایبی است که به درستی برابر صفر قرار می‌گیرند. از آنجایی که در بردار $\beta_\circ$، ۵ ضریب واقعاً برابر صفر هستند، لذا هر چه این شاخص به ۵ نزدیکتر باشد بهتر خواهد بود. به همین طریق Incorr متوسط تعداد ضرایبی است که به غلط برابر صفر قرار می‌گیرند. نزدیکی این شاخص به صفر، به منزله‌ی خوب بودن روش متناظر با آن از نظر شاخص Incorr می‌باشد. شاخص MRME نیز عبارتست از میانه‌ی مقادیر زیر که به ازای هر مجموعه داده‌ی تولید شده، محاسبه می‌گردند:

$$RME = \frac{(\hat{\beta}_{SCAD} - \beta_\circ)'\boldsymbol{\Sigma}(\hat{\beta}_{SCAD} - \beta_\circ)}{(\hat{\beta}_{OLS} - \beta_\circ)'\boldsymbol{\Sigma}(\hat{\beta}_{OLS} - \beta_\circ)},$$

به طوریکه $\boldsymbol{\Sigma}$ ماتریس کوواریانس متغیرهای توضیحی است. $\hat{\beta}_{SCAD}$ نیز با استفاده از رابطه‌ی (۱۰) بدست

می‌آید. RME در حقیقت نسبت خطای مدل برآوردگر SCAD به خطای مدل برآوردگر OLS است. علی رغم اینکه عملکرد توابع تاوان در درجه‌ی اول به پارامتر نظم وابسته است اما تا کنون هیچ گونه مطالعه‌ای بر اساس شبیه‌سازی در این زمینه انجام نشده است. به همین دلیل، در ابتدا جهت بررسی تاثیر مقدار پارامتر نظم بر عملکرد توابع تاوان SCAD و LASSO، ۵۰۰ مجموعه‌ی داده از مدل (۱۴) به ازای $\sigma = 3$ تولید کرده و به ازای هر یک از آنها سه مقدار متفاوت برای پارامتر نظم در نظر گرفته شد: i: $\lambda \in (1, 1/5)$، ii: $\lambda \in (0/5, 1)$ و iii: $\lambda \in (0, 0/5)$ و در هر یک از این حالات با استفاده از توابع تاوان LASSO و SCAD متغیرها انتخاب شدند. نتایج حاصل در جدول ۱ آمده است.

همانطور که در جدول مذکور مشاهده می‌شود، با بزرگ شدن مقدار λ، متناظر با افزایش شاخص Corr، شاخص Incorr نیز افزایش می‌یابد. این نتیجه به معنای آن است که با بزرگ شدن λ نه تنها ضرایب واقعاً صفر $(\beta_3, \beta_4, \beta_6, \beta_7, \beta_8)$، برابر صفر قرار می‌گیرند، بلکه ضرایب واقعاً مخالف صفر $(\beta_1, \beta_2, \beta_5)$ نیز به غلط برابر صفر قرار می‌گیرند و این سبب کم برآوردی می‌گردد.

اینک روش‌های مختلف انتخاب متغیر مقایسه می‌شوند. هرچند که این قسمت از شبیه‌سازی قبلاً به طور محدود توسط فن و لی [8] انجام شده است ولی در اینجا، روش‌های مذکور به طور جامع‌تر و کامل‌تر با یکدیگر مقایسه می‌شوند. برای این منظور ۵۰۰ مجموعه‌ی داده از مدل (۱۴) به ازای $\sigma = 1, 3, 6$ و حجم نمونه‌ی $n = 60, 40, 20$ تولید شدند و برای هر مجموعه‌ی داده، روش‌های مختلف انتخاب متغیر اعمال شدند. نتایج حاصل در جدول ۲ آورده شده است.

ویژگی جالب توجه برآوردگر SCAD این است که با انتخاب مناسب پارامتر نظم و نیز افزایش حجم نمونه، با احتمالی که به سمت یک میل می‌کند، ضرایبی که واقعاً برابر صفر هستند از مدل خارج می‌شوند [8] البته برآوردگر مطرح شده توسط زو [21] نیز دارای این ویژگی می‌باشد. برای نشان دادن این ویژگی برآوردگر SCAD، ضرایبی که در β_0 برابر صفر بودند پیشاپیش از مدل خارج شده و سایر ضرایب به وسیله کمترین توان‌های دوم باقیمانده برآورد شده‌اند. نتایج مربوط به این برآوردگر در جدول ۲ با Oracle مشخص شده‌اند. همانطور که ملاحظه می‌شود با زیاد شدن حجم نمونه و نیز با کاهش واریانس خطا، عملکرد SCAD، AIC و BIC بهتر می‌شود؛ به طوریکه شاخص Corr افزایش می‌یابد و دو شاخص Incorr و MRME به تدریج کم می‌شوند. اما این نکته در مورد LASSO لزوماً برقرار نیست. به عنوان مثال شاخص MRME مربوط به این برآوردگر به ازای هر یک از مقادیر σ، با افزایش حجم نمونه افزایش می‌یابد. SCAD و BIC عملکرد مشابهی دارند به طوریکه به ازای هر یک از مقادیر σ، با زیاد شدن حجم نمونه نتایج مربوط به این دو، به نتایج بدست آمده از برآوردگر Oracle نزدیک می‌شوند. شاخص Corr مربوط به برآوردگر SCAD همواره بزرگتر از شاخص Corr مربوط به برآوردگر LASSO می‌باشد و این برتری با زیاد شدن n بارزتر می‌گردد. شاخص Corr مربوط به معیار AIC همواره کوچکتر از شاخص Corr مربوط به معیار BIC می‌باشد و این نشان می‌دهد که معیار AIC در

مقایسه با معیار BIC تمایل به بیش‌برآوردی دارد [۱۲].

جدول ۱. تاثیر پارامتر نظم بر انتخاب متغیر با استفاده از تابع تاوان.

		$\lambda \in (۰, ۰/۵)$		$\lambda \in (۰/۵, ۱)$		$\lambda \in (۱, ۱/۵)$	
		Corr	Incorr	Corr	Incorr	Corr	Incorr
$n = ۲۰$	SCAD	۳/۱۶۲	۰/۴۹	۳/۹۶۲	۰/۶۸۲	۴/۳۸۲	۰/۸۸۶
	LASSO	۲/۷۱۸	۰/۲۳	۳/۷۱	۰/۳۴۶	۴/۱۸۶	۰/۵۹۴
$n = ۴۰$	SCAD	۳/۸۸	۰/۱۸۴	۴/۴۲۶	۰/۳۷۲	۴/۵۸۸	۰/۵۲۸
	LASSO	۳/۳۵۲	۰/۰۵۴	۴	۰/۱۱	۴/۴۱۶	۰/۲۵۸
$n = ۶۰$	SCAD	۴/۱۳۲	۰/۰۸۸	۴/۵۸	۰/۲۰۶	۴/۶۷۲	۰/۳۷۶
	LASSO	۳/۶۰۸	۰/۰۱	۴/۱۵۸	۰/۰۲۶	۴/۵۲۸	۰/۱۱۲

جدول ۲. مقایسه‌ی روش‌های انتخاب متغیر در مدل‌های رگرسیون خطی.

		$\sigma = ۶$			$\sigma = ۳$			$\sigma = ۱$		
		Corr	Incorr	MRME×۱۰۰	Corr	Incorr	MRME×۱۰۰	Corr	Incorr	MRME×۱۰۰
$n = ۲۰$	SCAD	۲/۹۹۲	۱/۱۲۸	۸۲/۳۰	۳/۶۶۶	۰/۶۹۴	۸۴/۸۲	۴/۰۹۴	۰/۰۱۲	۶۱/۸۳
	LASSO	۲/۸۸۴	۰/۷۹	۴۸/۴۴	۳/۱۶	۰/۳۴۴	۶۶/۴۵	۳/۰۱	۰/۰۰۲	۷۱/۹۱
	AIC	۳/۳۲۶	۱/۳۳۲	۸۷/۴۲	۳/۳۳۲	۰/۶۲۶	۹۱/۸۸	۳/۴۹	۰/۰۰۶	۸۰/۳۳
	BIC	۳/۹	۱/۶۱	۷۷/۳۷	۳/۹۳	۰/۸۱۸	۸۸/۱۹	۳/۹۸۴	۰/۰۱	۶۹/۵۲
	Oracle	۵	۰	۲۴/۶۹	۵	۰	۲۶/۱۸	۵	۰	۲۴/۵۶
$n = ۴۰$	SCAD	۳/۶۷۸	۰/۸۱۸	۸۲/۲۲	۴/۱۵	۰/۲۸۲	۷۳/۵۷	۴/۷۳۶	۰	۳۹/۶۹
	LASSO	۳/۴۰۲	۰/۵۳۸	۶۱/۵۳	۳/۶۱۲	۰/۰۷۸	۶۸/۲۶	۳/۵۸۶	۰	۷۱/۲۷
	AIC	۳/۸۱۸	۰/۹۹	۹۱/۷۰	۳/۸۳۲	۰/۲۰۴	۸۰/۳۹	۳/۹۵۶	۰	۷۴/۴۱
	BIC	۴/۴۷	۱/۳۵۶	۸۹/۵۹	۴/۴۹۸	۰/۳۵۲	۶۹/۵۹	۴/۶۲۸	۰	۴۷/۷۹
	Oracle	۵	۰	۳۲/۷۸	۵	۰	۳۰/۳۵	۵	۰	۲۸/۸۳
$n = ۶۰$	SCAD	۳/۸۹۴	۰/۵۷۴	۸۳/۲۲	۴/۳۸۸	۰/۱۶۴	۶۳/۳۸	۴/۸۲	۰	۳۸/۵۹
	LASSO	۳/۶	۰/۳۰۶	۶۵/۵۹	۳/۷۰۲	۰/۰۱۸	۷۲/۸۵	۳/۸۴	۰	۷۲/۸۹
	AIC	۳/۹۲۸	۰/۶۹۶	۸۹/۰۴	۳/۹۸۲	۰/۰۷۴	۷۴/۴۶	۴/۰۶۶	۰	۷۳/۱۳
	BIC	۴/۵۸۴	۱/۱۰۱	۹۱/۵۵	۴/۶۷۸	۰/۱۸۴	۵۲/۹۳	۴/۶۹	۰	۴۷/۰۸
	Oracle	۵	۰	۳۲/۹۳	۵	۰	۳۱/۸۹	۵	۰	۳۱/۷۹

۷ مثال کاربردی

در اینجا با یک مجموعه داده‌ی واقعی، نحوه‌ی بکارگیری روش‌های مختلف انتخاب متغیر را در عمل نشان می‌دهیم. داده‌های این مثال از ۱۸۰ دانش آموز دختر دوره‌ی راهنمایی شهر سمنان در سال تحصیلی ۱۳۸۷–۱۳۸۶ جمع‌آوری شده‌اند [۱]. هدف از این مطالعه بررسی تاثیر سبک‌های فرزند پروری و انواع عزت نفس بر عملکرد تحصیلی دانش آموزان می‌باشد. برای این منظور دو پرسشنامه‌ی عزت نفس آلیس پوپ و سبک‌های فرزند پروری دیانابامریند به ترتیب توسط دانش آموزان و والدین آنها تکمیل شد و با مقیاس بندی‌های به کار رفته معلوم گردید که هر دانش آموز به چه میزان از انواع عزت نفس ($X_۱$: عمومی، $X_۲$: تحصیلی، $X_۳$: خانوادگی، $X_۴$: اجتماعی و $X_۵$: جسمانی) و والدین آنها به چه میزان از انواع سبک‌های فرزند پروری ($X_۶$: آسان گیرانه، $X_۷$: مستبدانه و $X_۸$: مقتدرانه) برخوردار بودند. به منظور بررسی تاثیر متغیرهای توضیحی فوق بر عملکرد تحصیلی دانش آموزان، معدل نیمسال تحصیلی آنها به عنوان متغیر پاسخ در نظر گرفته شد.

علاوه بر متغیرهای $X_۱$، ... و $X_۸$، اثرات متقابل $X_۶X_۷$، $X_۶X_۸$ و $X_۷X_۸$ نیز به عنوان متغیرهای توضیحی در فرآیند انتخاب متغیر وارد مدل شدند. نتایج انتخاب

متغیر در جدول ۳ آمده است. اعداد داخل پرانتز برآورد انحراف استاندارد ضرایب را نشان می‌دهد. خروجی مربوط به دو معیار AIC و BIC با استفاده از روش PROC GLMSELECT و خروجی مربوط به تابع تاوان SCAD با بکارگیری روش PROC SCADLS (هر دو از نرم افزار SAS) بدست آمده‌اند. خروجی مربوط به OLS نیز با استفاده از روش PROC REG حاصل شده است.

همانطور که در جدول ۳ مشاهده می‌شود، معیار AIC مدل پیچیده‌تری را نسبت به معیار BIC انتخاب کرده است که این نکته با ویژگی تمایل AIC به بیش‌برآوردی، همخوانی دارد. همچنین نتیجه‌ی حاصل از SCAD و BIC یکسان می‌باشد اما به هر حال همانطور که بریمن [۳] نشان داده است، معیارهای اطلاع در مقابل تغییرات ناچیز داده‌ها بی‌ثبات هستند. برای مشاهده‌ی موردی در این رابطه می‌توان، به ونگ و همکاران [۲۰]، صفحه‌ی ۵۶۴، مراجعه نمود.

با توجه به اینکه هدف از انتخاب متغیر، تعیین متغیرهایی است که در مدل نهایی حضور پیدا می‌کنند و از طرفی عرض از مبدا عملاً به عنوان یک متغیر توضیحی در نظر گرفته نمی‌شود، دو روش PROC SCADLS و PROC GLMSELECT ابتدا متغیرهای توضیحی را استاندارد کرده و پس از انتخاب متغیر، ضرایب را به مقیاس اصلی برمی‌گردانند.

جدول ۳. برآورد ضرایب و انحراف استاندارد آنها در داده‌های عزت نفس و سبک‌های فرزند پروری.

	OLS	AIC	BIC	SCAD
Intercept	۱۴/۸۹۵ (۱/۴۷۵)	۱۴/۶۷۷ (۰/۷۷۶)	۱۳/۴۳۰ (۰/۲۴۷)	۱۳/۴۳۰ (۰/۲۴۶)
X_1	۰/۰۲۷ (۰/۰۳۴)	۰ (–)	۰ (–)	۰ (–)
X_2	۰/۰۹۲ (۰/۰۲۹)	۰/۱۱۳ (۰/۰۲۱)	۰/۱۲۱ (۰/۰۲۰)	۰/۱۲۱ (۰/۰۲۰)
X_3	۰/۰۹۶ (۰/۰۲۷)	۰/۱۱۴ (۰/۰۲۱)	۰/۱۲۱ (۰/۰۲۱)	۰/۱۲۱ (۰/۰۲۱)
X_4	۰/۰۲۰ (۰/۰۳۱)	۰ (–)	۰ (–)	۰ (–)
X_5	۰/۰۰۳ (۰/۰۲۴)	۰ (–)	۰ (–)	۰ (–)
X_6	−۰/۰۵۳ (۰/۰۷۶)	−۰/۰۵۹ (۰/۰۳۲)	۰ (–)	۰ (–)
X_7	−۰/۰۶۹ (۰/۰۴۳)	−۰/۰۵۹ (۰/۰۳۲)	۰ (–)	۰ (–)
X_8	۰/۰۲۹ (۰/۰۴۹)	۰/۰۵۱ (۰/۰۱۲)	۰/۰۵۴ (۰/۰۰۹)	۰/۰۵۴ (۰/۰۰۹)
X_6X_7	۰/۰۰۳ (۰/۰۰۲)	۰/۰۰۴ (۰/۰۰۲)	۰ (–)	۰ (–)
X_6X_8	۰/۰۰۰ (۰/۰۰۲)	۰ (–)	۰ (–)	۰ (–)
X_7X_8	۰/۰۰۱ (۰/۰۰۱)	۰ (–)	۰ (–)	۰ (–)

۸ بحث و نتیجه‌گیری

با توجه به اهمیت روزافزون مبحث انتخاب متغیر در علوم مختلف، همواره روش‌های جدیدی برای این منظور ارایه می‌گردد. در این بین استفاده از تابع تاوان، بسیار مورد توجه قرار گرفته است. در این مقاله علاوه بر معرفی اجمالی برخی از توابع تاوان و بسته‌های نرم‌افزاری مرتبط با آنها، تابع تاوان SCAD را، به دلیل ویژگی‌های بهینه‌ای که در مقایسه با سایر توابع تاوان دارد، به طور خاص مورد بررسی قرار دادیم. بر اساس نتایج بدست آمده از مطالعات شبیه‌سازی، تابع تاوان SCAD در مقایسه با LASSO از برتری قابل توجه‌ای برخوردار است (جدول ۲ را ببینید) همچنین بر اساس نتایج مندرج در جدول ۱، عملکرد توابع تاوان تحت تاثیر مقدار پارامتر نظم قرار

دارد؛ بنابراین انتخاب مناسب پارامتر نظم در درجه‌ی اول اهمیت، برای استفاده از هر تابع تاوانی قرار دارد.

قدردانی

نویسندگان مقاله از جناب آقای مهدی حسین‌پوری، به دلیل فراهم نمودن امکان استفاده از داده‌های مربوط به بخش مثال کاربردی، کمال تشکر را دارند.

مراجع

[۱] قاسمی، ح. (۱۳۸۷)، بررسی رابطه‌ی چندگانه‌ی سبک‌های فرزند پروری و انواع عزت نفس با عملکرد تحصیلی دانش آموزان دختر دوره‌ی راهنمایی شهر سمنان در سال تحصیلی ۸۷–۱۳۸۶، پایان نامه کارشناسی، دانشگاه سمنان.

[2] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *Proceeding 2nd Inter Symposium on Information Theory*, 267-281, Budapest.

[3] Breiman, L. (1996). Heuristics of Instability and Stabilization in Model Selection, *the Annals of Statistics*,**24**, 6, 2350-2383.

[4] Burnham, K.P. and Anderson, D.R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*, Springer.

[5] Craven, P. and Wahba, G. (1979). Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation, *Numerische Mathematika*, **31**, 377-403.

[6] Dziak, J. J., Lemmon, D. R., and Li, R. (2008). PROC SCADLS User's Guide (version 1.0.5 beta). State College, PA: The Methodology Center, Pennsylvania State University.

[7] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, *Annals of statistics*, **32**, 407-451.

[8] Fan, J. and Li, R. (2001). Variable Selection Via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association*, **96**, 1348-1360.

[9] Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery, *Proceedings of the International Congress of Mathematicians*, (M. Sanzsole, J. Soria, J.L. Varona, J. Verdera, eds.) Vol. III, 595-622.

[10] Frank, I.E. and Friedman, J.H. (1993). A Statistical View of Some Chemometrics Regression Tools, *Technometrics*, **35**, 109-148.

[11] Fu, W.J. (1998). Penalized Regression: The Bridge Versus the LASSO, *Journal of Computational and Graphical Statistics*, **7**, 397-416.

[12] George, E. I. (2000). The Variable Selection Problem, *Journal of the American Statistical Association*, **95**, 452, 1304-1308.

[13] Hoerl, A.E. and Kennard, R. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, **12**, 55-67.

[14] Hunter, D.R. and Li. R. (2005). Variable Selection Using MM Algorithms, *The Annals of Statistics*, **33**, 1617-1642.

[15] Leng, C., Lin, Y. and Wahba, G. (2006). A Note on the lasso and Related Procedures in Model Selection, *Statistica Sinica*, **16**, 1273-1284.

[16] Montgomery, D.C and Peck, E.A (1992).*Introduction to Linaer Regression Analysis*, 2nd Ed., Wiley: New York.

[17] Rencher, A.C. and Schaalje, G.B. (2008). *Linear Models in Statistics*, 2nd ed, John Wiley and Sons.

[18] Schwarz, G. (1978), Estimating the Dimension of a Model, *Annals of Statistics*, **6**, 461-464.

[19] Tibshirani, R.J. (1996). Regression Shrinkage and Selection Via the LASSO, *Journal of the Royal Statistical society* Ser B, **58**, 267-288.

[20] Wang, H., Li. R. and Tsai, C. L. (2007). Tuning Parameters Selectors for Smoothly Clipped Absolute Deviation Method, *Biometrika Trust*, **94**, 553-568.

[21] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association*, **101**, (476), 1418-1429.

[22] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B*, **67**, 2, 301-320.