مقایسه روش های برآوردیابی و انتخاب متغیر در مدل های رگرسیونی با استفاده از داده‌های شبیه سازی

علی احمدی [۱]، هوشنگ طالبی [۲]

چکیده:

در این مقاله ضمن معرفی روش های جدیدی که در زمینه برآوردیابی و انتخاب متغیر در مدل های رگرسیونی مطرح شده‌اند، براساس یک بررسی مبتنی بر شبیه سازی از مدل های مختلف، به بررسی عملکرد این روش ها و همچنین مقایسه آنها با روش های معمول مثل روش انتخاب پیشرو و روش رگرسیون مرزی خواهیم پرداخت.

واژه‌های کلیدی: شبیه‌سازی، برآوردیابی، انتخاب متغیر، گارت نامنفی، لاسو، لارس، الاستیک نت، اسکار.

۱ مقدمه

انتخاب متغیر و برآورد کردن ضرایب در مدل رگرسیونی اساسی ترین بخش در مدل سازی است. روش برآوردیابی حداقل مربعات، روش انتخاب متغیر به صورت پیشرو و یا حتی روش رگرسیون مرزی [۳] در مواجه با داده هایی که از ویژگی های متفاوتی برخوردار باشند، عملکرد قابل اطمینانی از خود نشان نمی دهند. عدم پایداری، دقت پیشبینی کم و انتخاب نادرست متغیرها نمونه هایی از آسیب های مدل در هنگام استفاده از این روش ها هستند. بعلاوه این آسیب ها زمانی که همبستگی بین پیشبین ها زیاد باشد تشدید نیز می شوند. روش های انقباضی [۴] بعنوان راهکاری در جهت کاهش این آسیب ها بخصوص زمانی که همبستگی بین پیشبین ها زیاد باشد مورد توجه قرار گرفته است. جیمز و استین[۴] نشان دادند، برای تابع زیان مربع خطا، روش های انقباضی کارایی برآوردگر را افزایش می دهند. در این روش ها ضرایب رگرسیونی را با اعمال محدودیت روی دامنه تغییرات آنها برآورد می کنند. وجود چنین محدودیتی واریانس برآوردگر را کاهش داده ولی همراه با ایجاد اریبی برای برآوردگرخواهد بود(دادوستد واریانس—اریبی [۵])، بطوریکه می توان امیدوار بود در نهایت میانگین مربعات خطا کاهش یافته باشد. مدل خطی زیر را در نظر بگیرید:

$$y = X\beta + \epsilon \qquad (۱)$$

که در آن $y_{(n*1)}$ بردار مشاهدات مرکزی شده متغیر وابسته، $X_{(n*P)}$ ماتریس مقادیر استاندارد شده متغیرهای پیشبین، $\beta_{(P*1)}$ بردار ضرایب رگرسیونی و $\epsilon_{(n*1)}$ بردار

[۱] کارشناس ارشد آمار
[۲] عضو هیات علمی دانشگاه اصفهان
[۳] Ridge Regression
[۴] Shrinkage Methods
[۵] Bias-Variance Tradeoff

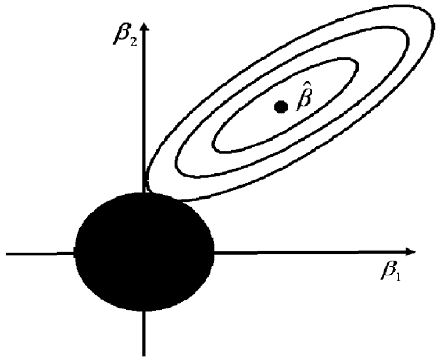
متغیرهای تصادفی خطاست. در این مدل ستون های ماتریس X را x_j و عناصر آن را با x_{ij} نمایش می دهیم.
در دو دهه اخیر روش های انقباضی مختلفی برای برآورد ضرایب رگرسیونی ارائه شده است. در بخش دوم برخی از این روش ها و ویژگی های آنها را مرور خواهیم کرد. برای مقایسه این روش هادر شرایط مختلف، در بخش سوم مدل های متنوعی را شبیه سازی کرده و ضرایب رگرسیونی را با استفاده از روش های ذکر شده در بخش دوم برآورد می کنیم. در بخش چهارم نتایج بدست آمده را تحلیل و درباره آسیب ها و مزایای این روش ها در شرایط مختلف بحث می کنیم.

۲ روش های انقباضی

روش های انقباضی از جمله روش های جدید در امر برآوردیابی و انتخاب متغیر در مدل های رگرسیونی هستند. روش رگرسیون مرزی یکی از اولین روش های انقباضی است. این روش ضرایب رگرسیونی را بصورت زیر برآورد می کند:

$$\widehat{\beta}_{Ridge} = \underset{\beta}{argmin} \parallel y - X\beta \parallel^2 \quad s.t \ \sum_{j=1}^{P} \beta_j^2 \leq t \quad (2)$$

که در آن $t \geq 0$ پارامتر کنترل [۶] بوده و میزان انقباض تحمیل شده به ضرایب را کنترل می کند. ایراد این روش استفاده از یک تابع تاوان [۷] توان دوم از ضرایب رگرسیونی است که مانع از صفر برآورد شدن ضرایب و در نتیجه عدم حذف متغیر از مدل می شود ([۵]).
در شکل (۱) ناحیه دایره ای، ناحیه تاوان ($\beta_1^2 + \beta_2^2 \leq t$) و نواحی بیضوی، مجموع توان دوم مانده ها با مرکزیت برآورد



شکل ۱. عملکرد روش رگرسیون مرزی در حالت $P = 2$

حداقل مربعات معمولی (OLS) در رابطه $\parallel y - X\beta \parallel^2 = n\widehat{\sigma}^2 + (\beta - \widehat{\beta}^{OLS})' X' X (\beta - \widehat{\beta}^{OLS})$ را نشان می دهد. این شکل به وضوح ناتوانی روش مرزی در صفر برآورد کردن ضرایب را نشان می دهد، زیرا برخورد دو ناحیه نمی تواند در نقطه ای باشد که در آن یکی از ضرایب صفر است. روش های جدید انقباضی میل دادن ضرایب به سمت صفر را به گونه ای انجام می دهند که بعضی از ضرایب، مربوط به متغیرهای بی اثر، دقیقا صفر برآورد شده و به این ترتیب متغیر مربوط به آن ضرایب از مدل خارج خواهد شد. بنابراین در این روش ها برآوردیابی و انتخاب متغیر تواما صورت می پذیرد. باید توجه داشت که در تمام روش های انقباضی ابتدا به ازاء مقادیر مختلف پارامتر کنترل برآورد ضرایب محاسبه شده و سپس با استفاده از معیارهای ارزیابی مدل، مثل $AIC, BIC, C_P, Cross - Validation, ...$، برآورد بهینه از میان مجموعه برآوردهای بدست آمده

[۶] Tuning Parameter
[۷] Penalty Function

انتخاب می شود. در ادامه این بخش روش های مختلف را به صورت مختصر مرور می کنیم.

۱.۲ روش گارت نامنفی

برآورد ضرایب رگرسیونی در روش گارت نامنفی [8][۲] به صورت زیر محاسبه می شوند:

$$\widehat{c} = (\widehat{c_1}, ..., \widehat{c_P})' = \underset{c}{argmin} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{P} c_j \widehat{\beta}_j x_{ij})^2 \quad (۳)$$

$$s.t \ \ \sum_{j=1}^{P} c_j \leq s \quad and \quad c_j \geq 0$$

که در آن $s \geq 0$ و $\widehat{\beta_j}$ برآورد حداقل مربعات β_j بوده که به عنوان برآورد اولیه استفاده شده است. در این روش برآورد نهایی به فرم $\widehat{\beta_j}^G = \widehat{c_j}\widehat{\beta_j}$ خواهد بود. ضرایب $\widehat{c_j}$ عامل های انقباض هستند که مقادیر کوچک یا صفر آنها باعث کوچک یا صفر شدن $\widehat{\beta_j}$ ها در مدل خواهد شد. این روش از نظر دقت پیشبینی و پایداری به خوبی روش مرزی بوده و در ضمن با صفر برآورد کردن بعضی از ضرایب عمل انتخاب متغیر را نیز انجام می دهد. ایراد اصلی این روش استفاده از برآورد حداقل مربعات بعنوان برآورد اولیه است. زیرا معایب برآوردگر حداقل مربعات ممکن است برآورد نهایی را نیز تحت تاثیر قرار دهد این عیب را معمولا می توان با استفاده از دیگر برآوردگرها به عنوان برآورد اولیه برطرف کرد.

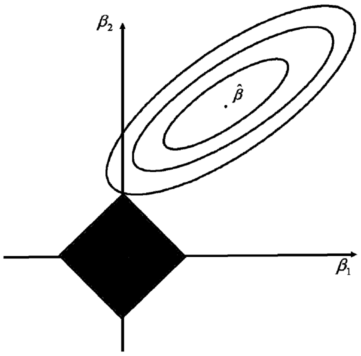
۲.۲ روش لاسو

روش لاسو[9][۵] اساسا مشابه روش رگرسیون مرزی است، با این تفاوت که بجای استفاده از تابع تاوان درجه دو از تابع تاوان مجموع قدرمطلق ضرایب استفاده می کند:

$$\widehat{\beta}_{Lasso} = \underset{\beta}{argmin} \parallel y - X\beta \parallel^2 \quad (۴)$$

$$s.t \ \ \sum_{j=1}^{P} |\beta_j| \leq t$$

این تغییر در نوع تابع تاوان امکان صفر برآورد شدن برخی از ضرایب را پدید می آورد.



شکل ۲. عملکرد روش لاسو در حالت $P = 2$

در شکل ۲ ناحیه لوزی شکل ناحیه تاوان ($|\beta_1| + |\beta_2| \leq t$) و نواحی بیضوی مجموع توان دوم مانده ها با مرکزیت برآورد حداقل مربعات را نشان می دهد. این شکل به وضوح توانایی روش لاسو در صفر برآورد کردن ضرایب را نشان می دهد، زیرا برخورد دو ناحیه می تواند در نقطه ای صورت گیرد که در آن یکی از ضرایب مقدار صفر را دارد. روش لاسو به لحاظ پایداری و دقت پیشبینی برآوردگرها عملکرد قابل قبولی را از خود نشان داده است. این روش همچنین بعنوان روشی پایه ای در ساخت روش های پیچیده تر نیز مورد استفاده قرار گرفته است.

[8] Nonnegative Garrote
[9] (Least Absolute Shrinkage and Selection Operator)Lasso

۳.۲ روش لارس

روش لارس [10] [۳] با رویکردی شبیه روش انتخاب پیشرو نسبت به ورود متغیرها به مدل عمل می کند. بدین ترتیب که در هر گام یک متغیر پیشبین به صورت زیر وارد مدل می شود:

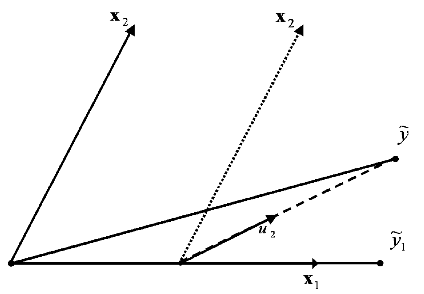
۱− ابتدا همه ضرایب صفر در نظر گرفته می شوند، سپس متغیری انتخاب می شود که بیشترین همبستگی را با متغیر وابسته داشته باشد، آنرا x_1 می نامیم.

۲− سپس بزرگترین گام ممکن در جهت این متغیر(با تغییر ضریب مربوط به آن)، تا جایی که همبستگی متغیری دیگر مثل x_2 با مانده‌ها برابر با همبستگی متغیرx_1 با مانده‌ها شود برداشته می شود(این برخلاف روش پیشرو است که در آن تا جایی پیش می رویم که همبستگی متغیرx_1 با مانده ها صفر شود).

۳− در مرحله بعد در مسیر متساوی الزاویه بین x_1 و x_2 پیش می رویم، تا جایی که همبستگی متغیری دیگر مثل x_3 با مانده برابر با همبستگی دو متغیر قبلی با مانده ها شده و بتواند وارد مدل شود(توجه کنید که با حرکت در مسیر متساوی الزاویه، همبستگی دو متغیر x_1 و x_2 با مانده ها به یک میزان کاهش می یابد).

به همین ترتیب در هر مرحله در مسیر متساوی الزاویه بین متغیرهای داخل مدل حرکت کرده تا اینکه همبستگی متغیری دیگر با مانده برابر با همبستگی متغیرهای داخل مدل با مانده شده و وارد مدل شود. این روند تا ورود همه متغیرها به مدل و صفر شدن ماکزیمم قدرمطلق همبستگی ها ادامه می یابد. مراحل الگوریتم لارس را در حالت $P=2$ با کمک شکل ۳ تشریح می کنیم:

در شکل ۳ $X=(x_1,x_2)$ و $\widetilde{y}_2$ برآورد حداقل مربعات بردار y است. بنابراین $X'(y-\widehat{\mu})=X'(\widetilde{y}_2-\widehat{\mu})$.



شکل ۳. عملکرد روش لارس در حالت $p=2$

الگوریتم لارس با $\widehat{\mu}_0=0$ آغاز می شود. با توجه به شکل ۳ $\widetilde{y}_2-\widehat{\mu}_0$ زاویه کمتری با متغیرx_1 نسبت به x_2 می سازد، یعنی همبستگی x_1 با y بیشتر است، بنابراین x_1 وارد مدل می شود. سپس $\widehat{\mu}_0$ در جهت x_1 تصحیح می شود یعنی در اولین گام خواهیم داشت: $\widehat{\mu}_1=\widehat{\mu}_0+\widehat{\gamma}_1 x_1$ که در آن $\widehat{\gamma}_1$ طوری انتخاب می شود که به ازاء آن: $\widetilde{y}_2-\widehat{\mu}_1$ نیمساز زاویه بین $x_1-\widehat{\mu}_1$ و x_2 باشد. اگر u_2 بردار یکه ای باشد که در طول نیمساز قرار گیرد آنگاه برآور بعدی لارس در گام دوم به فرم $\widehat{\mu}_2=\widehat{\mu}_1+\widehat{\gamma}_2 u_2$ خواهد بود. در این رابطه $\widehat{\gamma}_2$ طوری انتخاب می شود که همبستگی x_1 و x_2 با مانده ها صفر شود، یعنی $\widetilde{y}_2=\widehat{\mu}_2$.

برآوردهای بدست آمده در روش لارس تا حدود زیادی به روش لاسو نزدیک است و برخلاف روش لاسو، در روش لارس برآورد ضرایب نیازی به استفاده از الگوریتم های پیچیده ریاضی ندارد.

[10] (Least Angle Regression)LARS

۴.۲ روش الاستیک نت

برآورد اولیه در روش الاستیک نت[11] [۷] به صورت زیر محاسبه می شود:

$$\widehat{\beta}_{NEN} = \underset{\beta}{argmin}(\| y - X\beta \|^{۲} + \lambda_{۲} \| \beta \|^{۲} + \lambda_{۱} \| \beta \|_{۱}) \quad (۵)$$

که در آن $\lambda_{۱}$ و $\lambda_{۲}$ مقادیری نامنفی و $\| \beta \|_{۱} = \sum_{j=۱}^{P} |\beta_j|$ است. با در نظر گرفتن $\alpha = \lambda_{۲}/(\lambda_{۱} + \lambda_{۲})$ می توان برآورد الاستیک نت اولیه را به فرم تاوانی زیر نوشت:

$$\widehat{\beta}_{NEN} = \underset{\beta}{argmin} \| y - X\beta \|^{۲} \quad (۶)$$

$$s.t \quad \alpha \| \beta \|^{۲} + (۱ - \alpha) \| \beta \|_{۱} \leq t \quad and \quad t \geq ۰$$

ناحیه تاوان در این روش ترکیبی اکیدا محدب از نواحی تاوان در روش های لاسو و مرزی است. این فرم از تابع تاوان علاوه بر قابلیت صفر برآورد کردن برخی از ضرایب، توانایی برابر برآورد کردن ضرایب متغیرهایی که اثر یکسان روی متغیر پاسخ دارند و یا به شدت همبسته هستند را نیز دارد. به این ترتیب از این روش می توان برای گروهبندی پیشبین ها به خصوص زمانی که تعداد آنها زیاد باشد استفاده کرد. برآورد اصلاح شده الاستیک نت که دارای دقت پیشبینی بالاتری نسبت به برآورد اولیه است بصورت $\widehat{\beta}_{EN} = (۱ + \lambda_{۲})\widehat{\beta}_{NEN}$ تعریف می شود که در آن $\widehat{\beta}_{EN}$ و $\widehat{\beta}_{NEN}$ به ترتیب برآوردگرالاستیک نت و برآوردگر الاستیک نت اولیه (خام)[12] هستند. روش الاستیک نت در اکثر مواقع از روش های گارت نامنفی، لاسو، لارس و مرزی دقت پیشبینی بالاتری دارد.

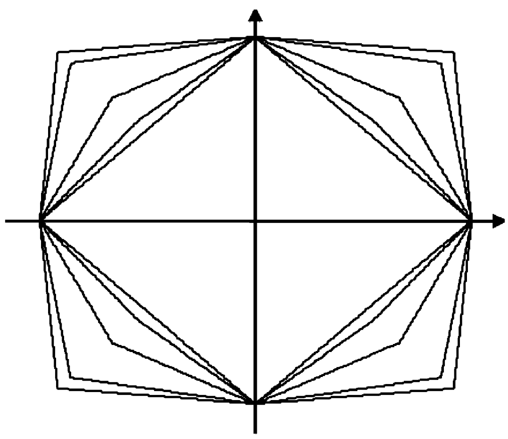
۵.۲ روش اسکار

روش اسکار[13] [۱] برآورد ضرایب رگرسیونی را بصورت زیر انجام می دهد:

$$\widehat{\beta}_{OSCAR} = \underset{\beta}{argmin} \| y - X\beta \|^{۲} \quad (۷)$$

$$s.t \ \sum_{j=۱}^{P} |\beta_j| + c\sum_{j<k} max(|\beta_j|, |\beta_k|) \leq t$$

تابع تاوان به کار رفته در این روش به لحاظ هندسی هشت گوشه است و این روش همانند روش الاستیک نت، توانایی برابر برآورد کردن پیشبین های به شدت همبسته و یا آنهایی که اثری یکسان روی متغیر پاسخ دارند را دارد. یعنی علاوه بر برآوردیابی و انتخاب متغیر قابلیت گروه بندی آنها نیز در این روش وجود دارد.



شکل ۴. ناحیه تاوان در روش اسکار در حالت $p = ۲$

شکل ۴ نواحی تاوان را به ازاء مقادیر مختلف c نشان می دهد. همانطور که در اشکال ۱ و ۲ هم نشان داده شد، در اینجا در صورتی که همبستگی بین دو پیشبین زیاد باشد، برخورد نواحی بیضوی با ناحیه تاوان می تواند در گوشه ای از ناحیه تاوان رخ دهد که در آن، ضرایب با مقداری

[11] Elastic Net
[12] Naive Elastic Net
[13] (Octagonal Shrinkage and Clustering Algorithm for Regression)OSCAR

برابر برآورد شوند. روش اسکار دارای دقت پیشبینی بالا و خاصیت گروهی قوی می باشد.

۳ شبیه سازی

در این بخش با استفاده از شبیه سازی به ارزیابی عملکرد و مقایسه روش های پیش گفته می پردازیم. از آنجایی که روش های انقباضی برای غلبه بر معایب روش های قدیمی، بخصوص زمانی که همبستگی بین پیشبین ها زیاد باشد، طراحی شده اند مدل های شبیه سازی شده را بیشتر براساس ساختارهای مختلف ماتریس کواریانس X مورد مطالعه قرار می دهیم. ابتدا ویژگی های داده های شبیه سازی شده را بیان و سپس نتایج کاربرد روش ها را ارائه می دهیم.

شبیه سازی اول: ۱۰۰ مجموعه داده از مدل ۱ تولید می کنیم که در آن $n=30$، $P=8$، $\epsilon\sim N_n(0,3^2I)$، $cov(x_i,x_j)=0.7^{|i-j|}$ و $\beta=(3,2,1.5,0,0,0,0,0)'$. در این مدل، متغیرهای موثر با همدیگر همبستگی نسبتا بالایی دارند. متغیرهای بی اثر نیز با همدیگر تا حدودی همبسته بوده ضمن اینکه متغیر موثر سوم و متغیر بی اثر چهارم نیز با همدیگر همبستگی نسبتا بالایی دارند.

شبیه سازی دوم: شبیه سازی مدل دوم مشابه مدل اول است با این تفاوت که در آن $\beta=(3,0,0,1.5,0,0,0,2)'$. در نتیجه در این مدل ، متغیرهای موثر همبستگی ناچیزی بایکدیگر داشته ولی با متغیرهای بی اثر همبستگی نسبتا بالایی دارند.

شبیه سازی سوم: ۱۰۰ مجموعه داده از مدل ۱ تولید می کنیم که در آن $n=100$، $\epsilon\sim N_n(0,5^2I)$، $cov(x_i,x_j)=0.5\ ,\ i\neq j$ و $Var(x_i)=1$ و $\beta=(\underbrace{0,\dots,0}_{10},\underbrace{2,\dots,2}_{10},\underbrace{0,\dots,0}_{10},\underbrace{2,\dots,2}_{10})'$. در این مدل گروه هایی از متغیرهای موثر و غیرموثر وجود دارد که همبستگی بین آنها متوسط و یکسان است.

شبیه سازی چهارم: ۱۰۰ مجموعه داده از مدل ۱ تولید می کنیم که در آن $n=40$، $\epsilon\sim t_n(0,I,5)$، $cov(x_i,x_j)=0.5^{|i-j|}$ و $\beta=(3,1.5,0,0,0,2,0,0)'$. این مدل، داده هایی را نشان می دهد که در آنها ضمن وجود مشاهدات پرت، متغیرهای موثر اول و دوم با همدیگر همبستگی متوسط و با متغیر موثر ششم همبستگی کمی دارند.

شبیه سازی پنجم: ۱۰۰ مجموعه داده از مدل ۱ تولید می کنیم که در آن $\epsilon_i=\rho\epsilon_{i-1}+u_i$، $\epsilon_0=0$، $u_i\sim N_n(0,0.5)$، $\beta=(3,1.5,0,0,0,2,0,0)'$، $\rho=0.4$، $n=40$ و $cov(x_i,x_j)=0.5^{|i-j|}$ این مدل داده هایی را نشان می دهد که در آنها ϵ_i ها دارای خودهمبستگی هستند.

شبیه سازی ششم: ۱۰۰ مجموعه داده از مدل ۱ تولید می کنیم که در آن $n=30$، $\epsilon\sim N_n(0,3^2I)$ و $\beta=(3,1.5,2,0,0,0,2,0)'$. در این مدل ساختار ماتریس کواریانس X به فرم زیر است:

جدول ۱. ساختار ماتریس کواریانس X

1	0.9	0.85	0.7	0.4	0.4	0.4	0.4
0.9	1	0.9	0.8	0.4	0.4	0.4	0.4
0.85	0.9	1	0.8	0.4	0.4	0.4	0.4
0.7	0.8	0.8	1	0.4	0.4	0.4	0.4
0.4	0.4	0.4	0.4	1	0.4	0.4	0.4
0.4	0.4	0.4	0.4	0.4	1	0.4	0.4
0.4	0.4	0.4	0.4	0.4	0.4	1	0.8
0.4	0.4	0.4	0.4	0.4	0.4	0.8	1

داده های شبیه سازی شده در این مدل مشابه داده هایی هستند که در آنها گروه هایی از متغیرهای به شدت همبسته وجود دارند. در این مدل متغیرهای پیشبین اول تا سوم موثر بوده و با متغیر بی اثر چهارم تشکیل یک گروه به شدت همبسته را می دهند. متغیر هفتم نیز موثر بوده و با متغیر بی اثر هشتم تشکیل یک گروه به شدت همبسته را می دهند.

برای داده های تولید شده در این شبیه سازی ها به ترتیب از روش های برآوردیابی حداقل مربعات معمولی، انتخاب پیشرو، روش مرزی، گارت نامنفی، روش لاسو، لارس، الاستیک نت و اسکار ضرایب رگرسیونی را برآورد و همچنین روش گارت نامنفی با برآوردهای اولیه مرزی، لارس و الاستیک نت را تکرار می کنیم. بعلاوه از معیار$ME = (\widehat{\beta} - \beta)' V(\widehat{\beta} - \beta)$، که در آن V ماتریس کواریانس X است، برای انتخاب برآورد بهینه در بین مجموعه برآوردهایی که در هر تکرار برای هر روش محاسبه می شوند استفاده شده است (برای جزئیات بیشتر در مورد معیار ME به [۵] مراجعه کنید). نتایج حاصل از شبیه سازی ها در جداول ۱، ۲ و ۳ ارائه شده اند (در این جداول عبارت «د.ا.م.د» نشان دهنده درصد انتخاب مدل درست و عبارت «م.ت.ض.غ.ص» نشان دهنده میانه تعداد ضرایب غیر صفر برآورد شده توسط هر روش در تکرارهای مختلف هر مدل است).

جدول ۱. نتایج مربوط به شبیه‌سازی‌های اول و دوم

روش برآوردیابی	مدل ۱			مدل ۲		
	میانه ME (انحراف معیار)	د.ا.م.د	م.ت.ض.غ.ص.	میانه ME (انحراف معیار)	د.ا.م.د	م.ت.ض.غ.ص.
$OLS\ Estimation$	۳.۳۴۰۹ (۰.۳۴۱۸)	۰	۸	۳.۲۹٦۸ (۰.۲۵۱۴)	۰	۸
$Forward\ Method$	۲.۰٦۱۸ (۰.۱۴۸۰)	۱۵	۳	۱.٦٦٦۹ (۰.۱۹۸۳)	۴۷	۳
$Ridge\ Regression$	۱.۵۷۰۱(۰.۱۹۷۲)	۰	۸	۱.۸۰۴۷ (۰.۰۹۹۳)	۰	۸
$N-Garrote(OSL)$	۱.۸۹۲۳ (۰.۱۵۴۳)	۸	۴	۱.۵۴۰۷ (۰.۱۸۵۵)	۱۱	۴
$Lasso$	۱.۳۵۳۸ (۰.۱۴۸۲)	۱۳	۵	۱.۴٦۷۰ (۰.۱۲۵۷)	۳	۵
$Lars$	۱.۴۲۴۷ (۰.۱۴٦۲)	۲۳	۴	۱.٦۰۹۲ (۰.۱۳۴٦)	۵	۵
$Elastic\ Net$	۰.۹۷۰۴ (۰.۱۱۸۷)	۴۴	۳	۱.۳۲٦۴ (۰.۱۲۳۸)	۹	۵
$Oscar$	۱.۰۴۴۷ (۰.۱۳۷۹)	۱۲	٦	۱.۵۵٦۹ (۰.۱۰۰٦)	۷	۴
$N-Garrote(Ridge)$	۱.۲۴۲۰ (۰.۱۲۷۲)	۱۸	۴	۱.۳۸۵۱ (۰.۱۲۷۹)	۱۴	۴
$N-Garrote(Lars)$	۱.۳۸۹۵ (۰.۱۴۲٦)	۲۹	۳	۱.۳۲۵۴ (۰.۱٦۰۵)	۱۹	۴
$N-Garrote(EN)$	۱.۱۳۷۹ (۰.۰۸۹۹)	۵۱	۳	۱.۲۹۴۵ (۰.۱۲۰۲)	۲۲	۴

جدول ۲. نتایج مربوط به شبیه‌سازی‌های سوم و چهارم

روش برآوردیابی	مدل ۳: میانه ME (انحراف معیار)	مدل ۳: د.ا.م.د	مدل ۳: م.ت.ض.غ.ص.	مدل ۴: میانه ME (انحراف معیار)	مدل ۴: د.ا.م.د	مدل ۴: م.ت.ض.غ.ص.
$OLS\ Estimation$	۱۵۲.۲۰ (۵.۷۱۸)	۰	۴۰	۰.۵۳۹۵ (۰.۰۲۶۱)	۰	۸
$Forward\ Method$	۸۳.۶۸ (۲.۳۱۹)	۰	۹	۰.۳۰۱۲ (۰.۰۲۷۳)	۶۰	۳
$Ridge\ Regression$	۲۰.۸۶ (۰.۹۹۵)	۰	۴۰	۰.۴۲۳۷ (۰.۰۲۶۶)	۰	۸
$N-Garrote(OSL)$	۷۰.۶۲ (۱.۹۰۵)	۰	۱۵	۰.۲۴۸۸ (۰.۰۲۹۹)	۲۵	۴
$Lasso$	۴۲.۶۷ (۱.۳۷۵)	۰	۲۰	۰.۳۲۲۵ (۰.۰۲۲۷)	۹	۶
$Lars$	۴۴.۵۹ (۱.۴۸۶)	۰	۲۱	۰.۳۴۱۳ (۰.۰۲۵۶)	۱۳	۵
$Elastic\ Net$	۴۱.۹۱ (۱.۴۷۲)	۰	۲۰	۰.۱۸۰۱ (۰.۰۲۳۳)	۳۷	۴
$Oscar$	۲۱.۷۷ (۱.۷۵۷)	۰	۳۱	۰.۲۹۴۸ (۰.۰۳۴۴)	۱۲	۶
$N-Garrote(Ridge)$	۵۱.۲۹ (۱.۷۷۳)	۰	۱۹	۰.۲۴۵۴ (۰.۳۰۱۰)	۲۴	۴
$N-Garrote(Lars)$	۴۵.۰۲ (۱.۵۶۹)	۰	۱۷	۰.۲۳۶۳ (۰.۰۳۳۹)	۳۵	۴
$N-Garrote(EN)$	۴۲.۰۷ (۱.۴۹)	۰	۱۷	۰.۲۲۳۹ (۰.۰۳۰۸)	۵۴	۳

جدول ۳. نتایج مربوط به شبیه‌سازی‌های پنجم و ششم

روش برآوردیابی	مدل ۵: میانه ME (انحراف معیار)	مدل ۵: د.ا.م.د	مدل ۵: م.ت.ض.غ.ص.	مدل ۶: میانه ME (انحراف معیار)	مدل ۶: د.ا.م.د	مدل ۶: م.ت.ض.غ.ص.
$OLS\ Estimation$	۰.۳۰۱۷ (۰.۰۲۹۲)	۰	۸	۳.۷۵۱ (۰.۲۴۶)	۰	۸
$Forward\ Method$	۰.۲۳۳۱ (۰.۰۲۹۹)	۵۷	۳	۲.۴۰۳ (۰.۲۱۴)	۱	۴
$Ridge\ Regression$	۰.۲۳۳۸ (۰.۰۲۳۱)	۰	۸	۱.۶۸۶ (۰.۱۲۴)	۰	۸
$N-Garrote(OSL)$	۰.۲۰۷۵ (۰.۰۳۰۵)	۲۷	۴	۲.۲۷۹ (۰.۲۰۱)	۱	۵
$Lasso$	۰.۱۹۷۷ (۰.۰۲۵۲)	۲۰	۶	۱.۸۸۹ (۰.۱۷۰)	۵	۵
$Lars$	۰.۲۲۱۴ (۰.۰۲۴۰)	۲۲	۶	۱.۹۱۷ (۰.۱۷۸)	۶	۵
$Elastic\ Net$	۰.۱۱۱۷ (۰.۰۱۸۱)	۴۴	۴	۱.۳۲۶ (۰.۱۳۳)	۱۷	۵
$Oscar$	۰.۱۹۵۲ (۰.۰۳۱۸)	۹	۷	۱.۳۵۱ (۰.۱۱۴)	۱	۷
$N-Garrote(Ridge)$	۰.۲۰۴۲ (۰.۰۳۰۱)	۳۲	۴	۲.۰۵۳ (۰.۱۸۱)	۴	۵
$N-Garrote(Lars)$	۰.۲۰۴۶ (۰.۰۳۱۸)	۳۶	۴	۱.۸۱۵ (۰.۱۸۱)	۸	۵
$N-Garrote(EN)$	۰.۲۰۳۸ (۰.۰۳۰۰)	۵۱	۳	۱.۷۹۸ (۰.۱۸۲)	۱۷	۴

۴ تحلیل نتایج

در این بخش بصورت مختصر به تحلیل نتایج بدست آمده و ذکر معایب و محاسن هر یک از این روش ها خواهیم پرداخت. لازم به ذکر است که در این بررسی از

میانه مقادیر ME بدست آمده که حاصل از محاسبه این معیار برای ۱۰۰ مجموعه داده از هر مدل و در هر روش است، برای مقایسه خطای پیشبینی روش ها استفاده شده است. از درصد تعداد دفعاتی که هر روش مدل درست را انتخاب کرده و همچنین میانه تعداد ضرایب غیر صفر برآورد شده توسط هر روش در هر مدل برای مقایسه توانایی روش ها در انتخاب مدل صحیح استفاده شده است.

انتخاب متغیر در مدل معمولا باعث بالا رفتن دقت پیشبینی می شود. مدل هایی که در آنها برآورد OLS برای همه ضرایب متغیرها محاسبه شده و عملا انتخاب متغیری در آنها صورت نگرفته است دارای کمترین میزان دقت پیشبینی هستند. در مورد روش انتخاب پیشرو معروف است که این روش پرخور[۱۴] است، میانه تعداد ضرایب غیرصفر برآورد شده در جداول ۱ تا ۳ نیز گویای همین واقعیت است. این مسئله بخصوص در مدل ۳ مشهودتر است. مشاهده ضرایب بدست آمده نشان می دهد که این روش در انتخاب و تمایز بین متغیرهای موثر و غیرموثر همبسته توانایی کافی را ندارد. همچنین با زیاد شدن تعداد پیشبین ها و یا همبستگی بین آنها، از دقت پیشبینی و توانایی انتخاب مدل درست در این روش کاسته می شود.

مقایسه جداول ۱ تا ۳ برای روش رگرسیون مرزی نشان می دهد که این روش، معمولا نسبت به روش انتخاب پیشرو دقت پیشبینی بالاتری دارد. این روش بخصوص در مدل ۳، دارای بهترین دقت پیشبینی در بین سایر روش ها است. در این جداول مشاهده می شود که عیب اصلی این روش یعنی غیرصفر برآورد کردن ضرایب و در نتیجه عدم انتخاب متغیر در همه جا به چشم می خورد. می توان از این روش برای انتخاب برآورد اولیه در روش گارت نامنفی استفاده کرد.

روش گارت نامنفی با برآورد اولیه حداقل مربعات از نظر دقت پیشبینی از روش پیشرو بهتر است اما به دلیل استفاده از برآورد حداقل مربعات بعنوان برآورد اولیه از معایب این برآوردگر مستثنی نیست. البته همانطور که قبلا هم اشاره شد می توان از دیگر برآوردگرها بعنوان برآورد اولیه استفاده کرد که این تکنیک در اینجا مورد استفاده قرار گرفته است. نتایج نشان می دهد که استفاده از برآوردگرهایی که نسبت به برآورد حداقل مربعات بهتر هستند (مثل برآوردگر رگرسیون مرزی) بعنوان برآورد اولیه، باعث بهتر شدن نتایج در این روش می شود. اما باید توجه داشت که استفاده از روش گارت نامنفی بعد ازاستفاده از روش انقباضی دیگری، تضمینی برای بهتر کردن دقت پیشبینی نمی دهد. این مسپله ناشی از انقباض مضاعف[۱۵] است (یعنی اعمل اریبی بیش از حد به برآوردگر در حالی که باعث کاهش واریانس به اندازه کافی نشود).

بررسی نتایج بدست آمده نشان می دهد روش لاسو در مقایسه با روش گارت نامنفی (با برآورد اولیه حداقل مربعات) و روش رگرسیون مرزی در اکثر مواقع دقت پیشبینی بالاتری داشته و در مقایسه با روش پیشرو کمتر پرخور است. مشاهده برآوردهای بدست آمده برای

---

[۱۴] Greedy
[۱۵] Double Shrinkage

روش لاسو نشان می دهد که یکی از ایرادات این روش عدم ثبات کافی در انتخاب متغیرهای موثر، زمانی که داده ها شامل گروه هایی از متغیرهای پیشبین به شدت همبسته باشند، است.

نتایج بدست آمده در مورد روش لارس، همانطور که قبلا هم اشاره شد، بسیار به روش لاسو نزدیک است. یک چنین ارتباطی از نظر تئوری هم بین این دو روش وجود دارد(رجوع شود به [۱]).

روشی که از هر نظر نسبت به روش های دیگر برتری داشته روش الاستیک نت است. این روش بجز در مدل ۳، که مربوط به متغیرهای گروهی است، در سایر مدل ها از دیگر روش ها بهتر عمل کرده است. روش الاستیک نت هم از نظر دقت پیشبینی و هم از نظر انتخاب مدل صحیح از دیگر روش ها بهتر عمل کرده است. این روش از نظر تمایز بین متغیرهای موثر و غیرموثر همبسته هم بهتر از سایر روش ها عمل کرده است. روش اسکار نیز در اکثر مواقع از نظر دقت پیشبینی به خوبی روش الاستیک نت بوده و حتی در مدل ۳ از این روش نیز بهتر عمل کرده است. مشاهده برآوردهای بدست آمده در مدل های مختلف نشان می دهد که روش اسکار دارای اثر گروهی قویتری نسبت به روش الاستیک نت بوده است. البته تمایل این روش به برابر برآورد کردن متغیرهای به شدت همبسته باعث ضعف این روش در انتخاب مدل درست، وقتی همبستگی بین متغیر موثر و غیر موثر بالا باشد، شده است.

جمع بندی و ارائه چند پیشنهاد:

براساس نتایج شبیه سازی های انجام شده پیشنهادهایی را در مورد استفاده از روش های برآوردیابی و انتخاب متغیر ارائه می دهیم. باید توجه داشت که مشخص کردن ساختار ماتریس کواریانس متغیرهای پیشبین برای انتخاب درست یک روش برآوردیابی و انتخاب متغیر می تواند مفید باشد.

برای مدل هایی با تعداد متغیرهای متوسط یا کم(مثلا $P \leq ۱۰$) روش الاستیک نت هم در حالتی که همبستگی بین پیشبین ها زیاد باشد و هم در حالتی که این همبستگی چندان قابل توجه نباشد عملکرد قابل قبولی از خود نشان داده است. البته در حالتی که همبستگی بین پیشبین ها زیاد نباشد می توان از روش های دیگری مثل لارس و لاسو نیز استفاده کرد. بخصوص از روش لارس که هم از نظر دقت پیشبینی تقریبا مشابه روش لاسو بوده و هم از نظر شیوه محاسبات به سادگی روش پیشرو است. در این حالت اگر تعداد نمونه ها زیاد باشد می توان از روش گارت نامنفی نیز استفاده کرد(به دلیل خواص مجانبی آن[٦]). البته پیشنهاد می شود از برآورد مرزی بعنوان برآورد اولیه در روش گارت نامنفی استفاده شود. در این صورت هم از خواص خوب روش مرزی در دقت پیشبینی بهره برده ایم و هم خاصیت تنک بودن روش گارت و در ضمن تا حد امکان از انقباض مضاعف دوری جسته ایم. اما برای حالتی که همبستگی بین پیشبین ها زیاد باشد تنها استفاده از روش های الاستیک نت و اسکار توصیه می شود. در این حالت اگر تعداد متغیرهای پیشبین خیلی زیاد بوده و گروهبندی آنها مورد علاقه باشد می توان از روش اسکار استفاده کرد، زیرا این روش خاصیت گروه بندی قویتری نسبت به روش الاستیک نت دارد. در حالتی که تعداد متغیرهای پیشبین زیاد

باشد(مثلا $p \geq ۱۰$)، در دو حالت همبستگی زیاد یا کم استفاده شود.
پیشنهاد می شود تنها از روش الاستیک نت و یا اسکار

مراجع

[1] Bondell, H. Reich, B.(2008), Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR, *Biometrics*, 64, 115-123

[2] Breiman, L. (1995), Better subset regression using the nonnegative garrote, *Technometrics*, 37, 373-384.

[3] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R (2004). Least Angle Regression, *Ann.Statist*, 32, 407-499.

[4] James, W., C.Stein (1961), Estimation with quadratic loss,*Proceeding of the forth berkely symposium*, 1, 361-379.

[5] Tibshirani, R.(1996), Regression shrinkage and selection via the lasso,*J. Roy. Statist.Soc.Ser*, B.58, 267-288.

[6] Yuan, M., Lin, Y.(2007), On the nonnegative garrote estimator,*J. Roy. Statist.Soc.Ser*, B.69, 143-161.

[7] Zou, H., Hastie, T.(2005), Regularization and variable selection via the elastic net,*J.R Statist.Sco*, B.67, 301-320.