Iranian Journal of Animal Science
Vol 51, No 4, Winter 2021 (313-336)
DOI: 10.22059/ijas.2021.296109.653768

علوم دامی ایران
دورهٔ ۵۱، شمارهٔ ۴، زمستان ۱۳۹۹ (ص ۳۳۶-۳۱۳)

بررسی مقایسه‌ای روش‌های مختلف آماری در ارزیابی ژنومی با استفاده از کدهای R

زهرا اکبری[۱]، آرش اردلان[۲*]، مصطفی قادری زفره‌ای[۳*]، فرجاد رفیعی[٤] و میثاق مریدی[٥]
۱ و ۲. دانش‌آموخته کارشناسی ارشد و استادیار آمار ریاضی، دانشکده علوم پایه، دانشگاه یاسوج
۳. دانشیار بیوانفورماتیک و ژنتیک، دانشکده کشاورزی، دانشگاه یاسوج
٤. استادیار، گروه بیوتکنولوژی کشاورزی، دانشکده علوم کشاورزی، دانشگاه گیلان
۵. دانش‌آموخته دکتری ژنتیک و اصلاح نژاد دام، دانشکده علوم کشاورزی، دانشگاه گیلان
(تاریخ دریافت: ۱۳۹۸/۱۱/۱٤- تاریخ پذیرش: ۱۳۹۹/۱۱/۱۸)

چکیده

انتخاب ژنومی از بزرگ‌ترین پیشرفت‌های حوزه به‌نژادی حیوانات و گیاهان در اوایل قرن بیست و یکم میلادی محسوب می‌گردد. روال ارزیابی ژنومی، که روی انتخاب به کمک نشانگر بنا نهاده شد، متکی به پیش‌فرض وجود عدم تعادل پیوستگی بین نشانگرهای تک نوکلئوتیدی (SNP) متراکم در سطح ژنوم و جایگاه‌های کنترل کننده صفات کمی (QTL) است. از نظر ارزیابی ژنتیکی، انتخاب ژنومی، بسیاری از مدل‌های رایج را تحت تأثیر قرار داده و منجر به ایجاد مدل‌های آماری-ژنتیکی جدیدی شده است که هر یک فرضیه‌های مختلفی را کنکاش می‌کنند. گرچه این مدل‌ها را می‌توان بر اساس معیارهای مختلفی گروه بندی کرد، اما با در نظر گرفتن توزیع صفات مورد بررسی، می‌توان آنها را در دو گروه فراسنجه‌ای و نافراسنجه‌ای تقسیم‌بندی نمود. در این پژوهش صحت ارزش‌های ارثی ژنومی با استفاده از روش‌های آماری مختلف فراسنجه‌ای و نافراسنجه‌ای مورد بررسی قرار گرفته است. روش‌های فراسنجه‌ای مورد استفاده عبارت از رگرسیون ستیغی، رگرسیون لاسو، روش الاستیک-نت، مدل‌های مختلط و روش‌های بیزی شامل رگرسیون ستیغی بیزی، لاسو بیزی، بیز A، بیز B، بیز C و بیز D هستند. روش‌های نافراسنجه‌ای شامل ارگرسیون هسته‌ای، فضای هیلبرت با هسته بازآفرین و ماشین بردار پشتیبان رگرسیونی می باشند. تمامی این روش‌ها روی یک مجموعه داده واقعی شامل اطلاعات ژنومی و فنوتیپی مربوط به ۲۳۰۰ حیوان، با استفاده از کُدهای R اجرا شدند. برای انتخاب مدل مناسب، از معیارهای صحت (همبستگی ارزش ارثی واقعی و برآورد شده) و میانگین مربعات خطا (MSE) استفاده شد. نتایج نشان داد که کارایی پیش‌بینی روش های فراسنجه‌ای نسبت به روش های نافراسنجه‌ای بالاتر است. در میان مدل‌های ارزیابی ژنومی مورد استفاده به‌طور نسبی نشان داده شد که روش بیز B نسبت به سایر مدل‌ها، از صحت و عملکرد بهتری برخوردار است و این با نتایج سایر پژوهشگران همخوانی نداشت. این تضاد احتمالاً به دلیل ساختار داده‌های مورد استفاده بوده است. یکی از اهداف این پژوهش ارایه مدل‌های آماری ارزیابی ژنومی همراه با کُدهای اجرایی آنها در محیط R بوده است، لذا کُدهای یاد شده در این مقاله می‌توانند در یادگیری مدل‌های ارزیابی ژنتیکی مورد بحث کمک شایانی به‌کاربران بکنند.

واژه‌های کلیدی: انتخاب ژنومی، روش‌های فراسنجه‌ای و نافراسنجه‌ای، صحت ارزیابی ژنومی، محیط R.

Comparative study of statistical methods for genomic evaluation using R codes

Zahra Akbari[1], Arash Ardalan[2*], Mostafa Ghaderi-Zefrehei[3*], Farjad Rafeie[4] and Misagh Moridi[5]
1, 2. M. Sc. Graduate and Assistant Professor of Mathematical Statistics, Faculty of Sciences, Yasouj University, Yasouj, Iran
3. Associate Professor of Bioinformatics and Genetics, Faculty of Agriculture, Yasouj University, Yasouj, Iran
4. Assistant Professor, Department of Agricultural Biotechnology, Faculty of Agricultural Sciences, University of Guilan, Rasht, Iran
5. Ph.D. in Animal Breeding and Genetics, University of Guilan, Rasht, Iran
(Received: Feb. 3, 2020 - Accepted: Feb. 6, 2021)

ABSTRACT

Genomic selection is one of the greatest advances in the field of animal and plant breeding in the early twentieth century. This genomic evaluation procedure, which was based on marker-assisted selection, relies on the assumption that there is linkage disequilibrium between dense single nucleotide markers (SNPs) at the genome level and quantitative trait control (QTL) sites. In terms of genetic evaluation, genomic selection influenced many common models and led to the development of new statistical genetic models, each of which explored different hypotheses. Although these models can be grouped according to different criteria, but by considering the distribution of the studied traits, they can be divided into: parametric and non-parametric groupes. In this study, the accuracy of genomic breeding values was investigated using various parametric and non-parametric statistical methods. Parametric methods were ridge regression, Lasso regression, Elastic net method, mixed models, Bayesian methods including Bayesian regression, Lasso Bayes, Bayes A, Bays B, Bays C and Bayes D. Non-parametric methods were kernel regression, reproducing kernel Hilbert spaces regression and regression support vector machine. All of these methods were performed on a real data set including genomic and phenotypic information of 2300 animals using R codes. To select the appropriate model, the criteria of accuracy (correlation of actual and estimated breeding values) and mean squared error (MSE) were used. The results showed that the predictive efficiency of parametric methods was higher than non-parametric-methods. Among the genomic evaluation models, it was shown that Bayes B was relatively more accurate and efficient than other models, however, this results did not agree with the results of other researchers, which may have been due to the data structure used in this study. Since one of the objectives of this study was to provide statistical models of genomic evaluation along with their executive codes in R environment, so the codes mentioned in this article could help the users to learn the genetic evaluation models discussed in this study.

Keywords: Accuracy of genomic evaluation, genomic selection, parametric and non-parametric methods, R environment.

* Corresponding author E-mail: a.ardalan@yu.ac.ir; mghaderi@yu.ac.ir

مقدمه

تاریخچه اصلاح نژاد دام، دوره طولانی را با اتکا به انتخاب فنوتیپی تجربه کرده و پیشرفت ژنتیکی کُندی را شاهد بوده است. در دهه ۱۹۵۰ روش بهترین برآورد نااُریب خطی (BLUP) توسعه داده شد که امکان برآورد ارزش ارثی تمامی افراد در شجره را فراهم کرده و از آن به‌طور گسترده‌ای در اصلاح نژاد دام و گیاه گردید (Henderson, 1975; Robinson, 1991; Jonas & de Koning, 2015). در دهه‌های اخیر همزمان با پیشرفت علم ژنتیک مولکولی، تحقیقات زیادی در زمینه شناسایی ژن‌ها و یا قطعات کروموزومی مؤثر بر صفات مهم اقتصادی در حیوانات اهلی انجام شده است. پیشرفت‌های ایجادشده در فناوری‌های پُربروندا زیستی به‌طور چشمگیری هزینه تعیین ژنوتیپ را کاهش داده است. بنابراین این امکان فراهم شده است که بتوان انتخاب بر پایه نشانگر را در یک مقیاس ژنومی به‌کار گرفت. این روش را اصطلاحاً انتخاب ژنومی (Genomic Selection) یا GS می‌نامند. امروزه GS به‌طور گسترده‌ای در اصلاح نژاد گیاهان و حیوانات مورد استفاده قرار می‌گیرد که این امر به‌طور چشمگیری باعث افزایش صحت انتخاب و همچنین تسریع در روند اصلاح نژاد شده است (Hayes *et al*., 2009a; Elshire *et al*., 2011; Bhat *et al*., 2016; Meuwissen *et al*., 2016; Crossa *et al*., 2017; Weller *et al*., 2017). انتخاب ژنومی به چند شکل اجرا می‌شود. ولی امروزه در بسیاری از گونه‌ها از انتخاب ژنومی تک مرحله‌ای استفاده می‌شود. انتخاب ژنومی در واقع نوعی انتخاب به کمک نشانگر است که در دو مرحله صورت می‌گیرد. در مرحله اول، با استفاده از مدل‌های آماری، اثر تمام نشانگرهای موجود در جمعیت مرجع که دارای اطلاعات فنوتیپی هستند، به‌طور همزمان برآورد گردیده و در مرحله دوم، آثار نشانگری برآورد شده برای پیش‌بینی ارزش ارثی ژنومی (GEBV) حیواناتی که تنها دارای اطلاعات ژنوتیپی بوده و اطلاعات فنوتیپی آنها در دسترس نیست، مورد استفاده قرار می‌گیرند (Goddard & Hayes, 2007). روش‌های آماری نقش مهمی در GS دارند و در یک نگاه خاص می‌توان آنها را به روش‌های مبتنی بر روابط خویشاوندی (Relationship-based methods) و روش‌های مبتنی بر اثر نشانگر (Marker effect-based methods) طبقه‌بندی کرد. در مقایسه با روش سنتی BLUP، روش GBLUP که از ماتریس خویشاوندی مبتنی بر داده‌های ژنومی استفاده می‌کند، ضرایب خویشاوندی دقیق‌تری را در بین افراد فراهم نموده و لذا دقت تخمین را افزایش می‌دهد (VanRaden, 2008). به منظور استفاده حداکثری از اطلاعات شجره و اطلاعات ژنومی، روش BLUP تک مرحله‌ای توسعه یافته و در اصلاح دام مورد استفاده گسترده قرار گرفته است (Aguilar *et al*., 2010; Christensen & Lund, 2010; Christensen *et al*., 2012; Misztal *et al*., 2013; Zhang *et al*., 2016). پس از آن اثر غالبیت در قالب یک مدل اثر چندگانه تصادفی برای استخراج واریانس اثر ژنتیکی غیرافزایشی توسعه داده شد (Aliloo *et al*., 2017؛ Varona *et al*., 2018). در ادامه اطلاعات مربوط به سطوح اُمیکس نظیر حاشیه‌نگاری ژن و QTL برای وزن‌دهی نشانگرها در ساخت ماتریس‌های روابط خویشاوندی ژنومی، که در GBLUP به‌کار گرفته می‌شوند، نیز مورد استفاده قرار گرفتند (Wimmer *et al*., 2013; An *et al*., 2017; Gao *et al*., 2017; Fikere *et al*., 2018; Schrag *et al*., 2018). اگرچه مجموعه‌ای از روش‌های GBLUP تصحیح شده برای افزایش دقت پیش‌بینی توسعه یافته‌اند، ولی روش اصلی GBLUP که از نظر محاسباتی کارآمدتر است و نیازهای پیش‌بینی صفات هدف پرورش با معماری ژنتیکی پیچیده را برآورده می‌کند، هنوز از محبوب‌ترین روش‌های مورد استفاده می‌باشد.

بر خلاف GBLUP، روش‌های مبتنی بر اثر نشانگر شامل دو مرحله اساسی است: ۱- برآورد اثر نشانگرهای ژنتیکی، ۲- تجمیع اثر نشانگرها برای به‌دست‌آوردن ارزش ارثی تخمین زده شده (EBV) هر فرد. در این موارد معمولاً از روش رگرسیون ستیغی (Ridge یا RRBLUP) و روش الفبای بیزی استفاده می‌شود (Whittaker *et al*., 1999; Endelman, 2011). در RRBLUP فرض می‌شود که همه نشانگرها واریانس ژنتیکی یکسانی دارند. ثابت شده است که این

روش معادل روش GBLUP است (Goddard, 2009; Hayes *et al*., 2009b). در روش BayesA فرض می‌شود که واریانس همه اثر نشانگر از توزیع معکوس - مربع پیروی می‌کند (Meuwissen *et al*., 2001). ولی در روش BayesB فرض می‌شود که واریانس اثر نشانگر از یک توزیع معکوس- مربع پیروی کرده (Meuwissen *et al*., 2001) و تنها نسبت π از نشانگرها در تحلیل ساختار ژنتیکی صفت هدف به‌کار گرفته می‌شوند. روش‌هایی که پس از آن توسعه یافته‌اند (نظیر BayesC، BayesCπ و BayesLASSO)، بهینه‌سازی بهتری از π را ایجاد کرده و توزیع‌های متمایزی را برای واریانس اثر نشانگر اختصاصی در نظر گرفته‌اند (Yi & Xu, 2008; Habier *et al*., 2011). Zhou *et al*. (2013) فرضیه هر دو روش GBLUP و بیزی را ترکیب کرده و فرضیه جدیدی را پیشنهاد کردند که در آن فرض می‌شود که همه نشانگرهای ژنتیکی دارای اثر بوده و نسبتی از اثر نشانگرها را می توان با GBLUP به‌دست آورد. ولی نسبت دیگری نشانگرها دارای اثر افرایشی بوده و واریانس این آثار افزایشی از توزیع نرمال پیروی می‌کند. این روش BSLMM (Bayesian Sparse Linear Mixed Model) نامیده شد. فرضیه انعطاف‌پذیر به‌کار رفته در BSLMM سبب شد که دقت پیش‌بینی بالاتری نسبت به BayesCπ، BayesLASSO و BVSR به‌دست آید (Zhou *et al*., 2013).

Moser *et al*. (2015) روش BayesR را پیشنهاد کردند که در آن کلیه نشانگرها به چهار گروه طبقه‌بندی شده و فرض می‌شود که واریانس آثار نشانگر در گروه‌های مختلف از توزیع نرمال با میزان واریانس‌های متفاوت پیروی می‌کند. انبوهی از آزمایش‌های شبیه‌سازی شده و واقعی نشان داده‌اند که BayesR همیشه نسبت به سایر روش‌های مبتنی بر اثر نشانگر از دقت پیش‌بینی بالاتر یا لااقل مشابهی برخوردار است (Zeng & Zhou, 2017; Hayes *et al*., 2019). علاوه بر روش آماری، دقت پیش‌بینی ارزش ارثی ممکن است تحت تأثیر تراکم نشانگر، فراوانی آللی جزیی (MAF)، وراثت‌پذیری و معماری ژنتیکی صفت هدف قرار گیرد. با این حال افزایش تعداد نشانگرها به ویژه برای روش‌های الفبای بیزی، به‌طور جدی بار محاسبات را افزایش می‌دهد. زیرا تعداد فراسنجه‌های ناشناخته که باید تخمین زده شوند، بسیار افزایش می‌یابند.

بررسی اثر روش‌های آماری و تراکم نشانگر بر تعادل در دقت پیش‌بینی و بازده محاسباتی برای برنامه‌های پرورشی عملی ضروری و دارای اهمیت است. روش‌های آماری زیادی با رویکردهای متفاوتی در حوزه GS توسعه داده شده و به کار رفته است. شاید بتوان گفت که مهم‌ترین وجه مشترک این روش‌ها بالا بردن دقت انتخاب ژنومی برای صفات با ماهیت و معماری ژنتیکی متفاوت بوده است. از یک نگاه فراسنجه‌محور، روش‌های آماری که به‌طور معمول در GS به‌کار رفته است را می‌توان به دو گروه روش‌های آماری فراسنجه‌ای و نافراسنجه‌ای طبقه-بندی نمود. هدف از پژوهش حاضر معرفی روش‌های یاد شده فوق و اجرای آنها توسط کُدهای R بوده است. در این راستا، دقت عملکرد آنها در برآورد ارزش ارثی روی داده‌های ژنومی موجود، با توجه به دو معیار همبستگی پیرسون ارزش‌های ارثی واقعی و ارزش ارثی پیش‌بینی شده و همچنین خطای پیش‌بینی بررسی می‌شود. با در نظر گرفتن معماری ژنتیکی صفات مختلف، اثر متقابل محیط و ژن و در دسترس قرار گرفتن روزافزون داده‌های اُمیکس، انتظار می‌رود که مدل‌های جدیدی در حوزه GS توسعه داده شوند. بنابراین به‌طور منطقی کسانی می‌تواند در مسیر توسعه مدل‌های جدید قدم بردارند که با ماهیت ریاضی- آماری و اجرای مدل‌های معمول فعلی آشنا باشند. از این دیدگاه، پژوهش حاضر می‌تواند به علاقمندان حوزه GS یاری رساند.

مواد و روش‌ها

در این پژوهش، برخی روش‌های فراسنجه‌ای و نافراسنجه‌ای مورد استفاده در انتخاب ژنومی مورد بررسی قرار گرفته‌اند. روش‌های فراسنجه‌ای به دو گروه روش‌های انقباضی فراوانی‌گرا (Frequentist shrinkage) و روش‌های انقباضی بیزی (Bayesian

shrinkage methods) تقسیم می‌شوند. روش‌های انقباضی فراوانی‌گرا شامل رگرسیون ستیغی (Ridge regression)، رگرسیون لاسو (Lasso regression)، روش الاستیک‌نت (Elastic net method) و مدل‌های مختلط (Mixed models) هستند. همچنین، روش‌های انقباضی بیزی نیز شامل رگرسیون ستیغی بیزی (Bayesian ridge regression)، لاسو بیزی (Bayesian lasso) و الفبای بیزی (Bayesian alphabet) می‌باشند. الفبای بیزی مطرح شده در این پژوهش شامل روش‌های بیز A (Bayes A)، بیز B (Bayes B)، بیز C (Bayes C) و بیز D (Bayes D) هستند. روش‌های نافراسنجه‌ای نیز شامل: رگرسیون هسته‌ای روی نشانگرهای SNP (Kernel regression on SNP markers)، فضای هیلبرت با هسته بازآفرین (Reproducing kernel hilbert space regression) و رگرسیون ماشین بردار پشتیبان (Support vector machine regression) می‌باشند.

روش‌های فراسنجه‌ای در انتخاب ژنومی

در انتخاب ژنومی، هدف اصلی پیش‌بینی ارزش ارثی افراد توسط مدل‌سازی ارتباط بین ژنوتیپ و فنوتیپ آنها است. یکی از ساده‌ترین مدل‌ها که می‌توان در این رابطه استفاده کرد، مدل رگرسیون خطی ساده است.که به‌صورت زیر نوشته می‌شود:

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + e_i$$

$(j = 1,...,p \text{ و } i = 1,...,n)$

در این رابطه، y_i ، ارزش فنوتیپی فرد i اُم، μ، میانگین کُل، x_{ij}، یک درایه از ماتریس طرح مربوط به نشانگر j اُم و فرد i اُم، β_j، اثر ثابت نشانگر j اُم و e_i، اثر تصادفی باقیمانده است. معمولاً باقیمانده e دارای توزیع نرمال با میانگین صفر و واریانس σ_e^2 است. برآورد رگرسیون خطی ساده حاصل از روش حداقل مربعات (Least squares) به‌صورت $\hat{\beta} = (X^T X)^{-1} X^T y$ می‌باشد. در یک تحلیل رگرسیونی هنگامی‌که ماتریس طرح دارای ابعاد بالایی باشد، به‌عبارت دیگر، تعداد فراسنجه‌های مدل رگرسیونی بیشتر از اندازه نمونه باشد، متغیرهای کمکی (ستون‌های ماتریس $\boldsymbol{X}$) اَبَر هم‌خط می‌شوند. به‌این ترتیب، زیر فضای تشکیل شده توسط متغیرهای کمکی هم‌خط ممکن است پُر رتبه نباشد. هنگامی‌که زیر فضا پُر رتبه نباشد، نمی‌توان سهم هر متغیر کمکی را به‌طور جداگانه در رابطه با تغییرات متغیر پاسخ به‌طور قطعی بیان کرد. بنابراین این عدم قطعیت در برازش مدل رگرسیونی روی داده‌ها اغلب با یک خطای برآورد فراسنجه‌های رگرسیونی مربوط به متغیرهای کمکی هم‌خط، منعکس می‌گردد. برای برطرف کردن مشکلات رگرسیون خطی ساده از روش‌های انقباضی که دارای یک فراسنجه جریمه λ هستند استفاده می‌شود. این فراسنجه جریمه (Penalty) بین نکویی برازش مدل و پیچیدگی آن یعنی تعداد متغیرهای پیش‌بینی‌کننده موجود در مدل یک موازنه برقرار می‌کند. Ogutu *et al.* (2012) کاربرد روش‌های انقباضی در انتخاب ژنومی را مطرح کردند.

روش‌های انقباضی فراوانی‌گرا

در روش‌های انقباضی، برآورد حاصل از رگرسیون خطی ساده به‌علاوه یک تابع جریمه می‌شود. برآوردها با مینیمم‌کردن عبارت Q در معادله زیر به‌دست می‌آیند.

$$Q = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda J(\beta)$$

در این رابطه، $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$، مجموع مربعات باقیمانده‌ها، λ فراسنجه جریمه و $J(\beta)$، یک تابع جریمه می‌باشد. تفاوت موجود بین روش‌های مختلف انقباضی تنها در تعیین همین تابع جریمه است. هر یک از این توابع جریمه منجر به یک رگرسیون جریمه‌دیده (Penalized regression) می‌شود. کُدنویسی این بخش در نرم‌افزار R جهت خوانش داده‌ها، پیش‌پردازش بر مبنای فراوانی آلل‌های جزیی و محاسبه‌ی صحت پیش‌بینی مدل‌های آماری برازش شده روی داده‌ها به کمک روش اعتبار سنجی متقابل ۵ دسته‌ای در ضمیمه (۱) ارائه شده است.

رگرسیون ستیغی

در رگرسیون خطی ساده روش برازش حداقل مربعات

فراسنجه‌های $\beta_0,\beta_1,\ldots,\beta_p$ را با استفاده از به حداقل‌رساندن معیار زیر برآورد می‌کند:

$$RSS=\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

در مواردی که بین متغیرهای پیش‌بینی کننده رابطه هم‌خطی وجود داشته باشد، برآوردهای حداقل مربعات (OLS) ضعیف عمل می‌کنند. از این رو، برای رسیدن به پیش‌بینی بهتر در مواجهه با هم‌خطی، روش رگرسیون ستیغی معرفی شده است (Hoerl & Kennard, 1970). در روش رگرسیون ستیغی از تابع جریمه زیر استفاده می‌شود:

$$J(\beta)=\sum_{j=1}^{p}\beta_j^2$$

با به حداقل‌رساندن عبارت Q، برآورد رگرسیون ستیغی به‌صورت زیر خواهد شد:

$$\hat{\beta}_R=(X^TX+\lambda I_p)^{-1}X^Ty$$

کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک رگرسیون ستیغی در ضمیمه (۲) ارائه شده است.

در رگرسیون ستیغی تابع جریمه که به‌صورت توان دوم ضرایب رگرسیونی است، مانع از صفر شدن برآورد این ضرایب گردیده و در نتیجه مانع از حذف متغیر می‌شود. بنابراین روش مزبور یک مدل قابل تفسیر ارائه نمی‌دهد. برای برطرف‌کردن این مشکل روش‌های جدید انقباضی که در آنها یافتن برآورد و انتخاب متغیر به‌صورت همزمان صورت می‌گیرد، ارائه شده‌اند. این روش‌ها شامل روش لاسو و الاستیک‌نت می‌باشند.

رگرسیون لاسو

Tibshirani (1996) یک روش جدید تحت عنوان لاسو معرفی کرد. در روش لاسو از تابع جریمه زیر استفاده می‌شود:

$$J(\beta)=\sum_{j=1}^{p}|\beta_j|$$

این روش برخی از ضرایب را منقبض کرده و سایر ضرایب را صفر می‌کند. از این رو، ضرایبی که تأثیر بیشتری بر متغیر پاسخ دارند در مدل نگه داشته شده و باعث بهبود تفسیر مدل می‌شوند. اما این روش نیز در مواردی که در بین پیش‌بینی کننده‌ها هم‌خطی بالایی وجود داشته باشد، نتایج ناسازگاری داشته و روش مناسبی محسوب نمی‌گردد. همچنین، زمانی‌که تعداد پیش‌بینی کننده‌ها بیشتر از تعداد مشاهدات باشد نمی‌تواند تعداد متغیرها را بیشتر از حجم نمونه انتخاب کند. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک لاسو در ضمیمه (۳) ذکر گردیده است.

روش الاستیک‌نت

روش الاستیک‌نت تعمیمی از روش لاسو است که در مقابل حداکثر همبستگی بین پیش‌بینی‌کننده‌ها توانمند است. در واقع این روش برای برطرف کردن مشکل روش لاسو در مواردی که پیش‌بینی‌کننده‌ها به‌شدت همبسته هستند ارائه گردید (Zou & Hastie, 2005). تابع $(1-\alpha)|\beta|_1+\alpha|\beta|^2$ را تابع جریمه الاستیک‌نت می‌نامند که ترکیبی محدب از جریمه‌های رگرسیون ستیغی و لاسو می‌باشد. در این رابطه، $\alpha=\lambda_1/(\lambda_1+\lambda_2)$ بوده و λ_1 و λ_2 به‌ترتیب فراسنجه‌های جریمه رگرسیون لاسو و ستیغی می‌باشند. زمانی‌که $\alpha=1$ باشد، الاستیک‌نت خام معادل با رگرسیون ستیغی شده و اگر $\alpha=0$ شود، معادل با رگرسیون لاسو می‌گردد. استفاده از این فرم تابع جریمه علاوه بر اینکه منجر به صفر شدن برآورد برخی از ضرایب بی‌اثر می‌گردد، سبب می‌شود که برآورد ضرایبی که اثر یکسانی روی متغیر پاسخ دارند و یا به شدت همبسته هستند نیز برابر گردند. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک الاستیک‌نت در ضمیمه (۴) ارائه شده است.

مدل‌های مختلط

در روش بهترین پیش‌بینی ناأریب خطی (BLUP) اثر عوامل ثابت و ارزش‌های ارثی به‌طور همزمان برآورد می‌گردند (Henderson, 1949). اگر متغیرهای پاسخ، همبسته بوده و $\Sigma=\sigma^2V$ باشد، از مدل‌های مختلط برای بیان رابطه بین متغیرهای مستقل و متغیرهای پاسخ استفاده می‌شود. شکل کُلی یک مدل مختلط به‌صورت $y=X\beta+Zu+e$ است. در این رابطه، **y**، یک

بردار $n\times1$ متشکل از مقادیر متغیر پاسخ، β، بردار اثر عوامل ثابت، X، ماتریس طرح آثار ثابت، u، بردار آثار تصادفی، Z، ماتریس ضرایب پیونددهنده مشاهدات (مقادیر متغیر پاسخ) به بردار آثار تصادفی و e بردار آثار تصادفی باقیمانده‌ها می‌باشند. امید ریاضی و واریانس آثار تصادفی عبارت از:

$$\mathbf{Cov}\begin{pmatrix}\mathbf{u}\\ \mathbf{e}\end{pmatrix}=\begin{pmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{pmatrix} \text{ و } \mathbf{E}\begin{pmatrix}\mathbf{u}\\ \mathbf{e}\end{pmatrix}=\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}$$

در حالت کلی معادله فوق مدل مختلط نامیده شده و می‌توان آن را برای برآورد فراسنجه **β** به‌صورت $\mathbf{y}=\mathbf{X\beta}+\mathbf{e}^*$ و با مفروضات $\mathbf{e}^*=\mathbf{Zm}+\mathbf{e}$ و $\mathbf{Cov}(\mathbf{e}^*)=\mathbf{V}=\mathbf{ZGZ}^{\mathbf{T}}+\mathbf{R}$ نوشت. روش‌های زیادی برای به‌دست آوردن برآورد اثر عوامل ثابت و پیش‌بینی آثار تصادفی وجود دارد. یک روش استفاده از فرضیات توزیعی $m\sim N(\mathbf{0},\mathbf{G})$ و $y|m\sim N(\mathbf{X\beta}+\mathbf{Zm},\mathbf{R})$ است. حداکثر درستنمایی $(\mathbf{y},\mathbf{m})$ تحت **β** و **m** نامعلوم به‌صورت زیر برآورد می‌گردد:

$$L(\mathbf{y},\mathbf{m})=f(\mathbf{y}|\mathbf{m})f(\mathbf{m})=\frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{R}|^{\frac{1}{2}}}\frac{1}{|\mathbf{R}|^{\frac{1}{2}}}\left[\frac{-1}{2}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zm})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zm})\right]\frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{G}|^{\frac{1}{2}}}\left[\frac{-1}{2}\mathbf{m}^{\mathrm{T}}\mathbf{G}^{-1}\mathbf{m}\right]$$

این معادله منجر به معیار زیر می‌شود:

$$(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zm})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zm})+\mathbf{m}^{\mathrm{T}}\mathbf{G}^{-1}\mathbf{m}$$

این معیار بهترین پیش‌بینی ناأریب خطی (BLUP) از (β, m) است که شامل حداقل مربعات تعمیم یافته همراه با یک جریمه می‌باشد. بنابراین بهترین برآورد ناأریب خطی $\hat{\boldsymbol{\beta}}$ و بهترین پیش‌بینی ناأریب خطی $\hat{\mathbf{m}}$ به‌صورت $\hat{\boldsymbol{\beta}}=(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$ و $\hat{\mathbf{m}}=\mathbf{GZ}^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$ خواهند بود. در اینجا، $\mathbf{R}=\mathbf{Var}(\mathbf{e})$ و $\mathbf{G}=\mathbf{Var}(\mathbf{m})$ است. فرض می‌شود که R و G ماتریس‌های کوواریانس مشخصی هستند. در حالت کلی، مقدار کوواریانس β و m نامعلوم است. برای برآورد فراسنجه‌ها در ماتریس کوواریانس از روش‌های حداکثر درستنمایی (LM) و حداکثر درستنمایی محدود شده (REML) استفاده می‌شود. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک BLUP در ضمیمه (۵) ذکر گردیده است.

روش‌های انقباضی بیزی

رگرسیون ستیغی بیزی

Pérez *et al.* (2010) کاربرد رگرسیون ستیغی بیزی در انتخاب ژنومی را مطرح کردند. در رگرسیون ستیغی بیزی، توزیع پیشنهادی برای ضرایب رگرسیونی (اثر نشانگر)، یک توزیع پیشین نرمال است که در آن یک واریانس مشترک برای تمام ضرایب رگرسیون در نظر گرفته می‌شود:

$$P\left(\beta_R\middle|\sigma^2_{\beta_R}\right)=\prod_{j=1}^{P}N\left(\beta_{R_j}\middle|\sigma^2_{\beta_R}\right)$$

این پیشین برآوردهایی را ارائه می‌دهد که معادل بیزی از برآوردهای به‌دست آمده از رگرسیون ستیغی است. فراسنجه واریانس $\sigma^2_{\beta_R}$ به‌صورت مقداری نامشخص رفتار می‌کند که دارای چگالی پیشین معکوس مربع کای مقیاس‌دار به‌صورت زیر است:

$$P\left(\sigma^2_{\beta_R}\right)=\chi^{-2}\left(\sigma^2_{\beta_R}\middle|df_{\beta_R},S_{\beta_R}\right)$$

در اینجا، df_{β_R}، درجه آزادی و S_{β_R}، فراسنجه مقیاس است. در رگرسیون ستیغی بیزی نیز مانند رگرسیون ستیغی همه آثار به میزان مشابهی منقبض می‌شوند. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک رگرسیون ستیغی بیزی در ضمیمه (۶) ارائه شده است.

لاسو بیزی

در روش لاسو، برآوردها را می‌توان از طریق به‌حداقل رساندن Q به‌صورت زیر به‌دست آورد:

$$Q=\sum_{i=1}^{n}\left(y_i-X_i^T\beta\right)^2+\lambda\sum_{j=1}^{P}\left|\beta_j\right|$$

این رابطه را می‌توان به‌عنوان نمای (Mode) توزیع پسین یک مدل بیزی با تابع درستنمایی نرمال و چگالی پیشین نمایی دوگانه برای ضرایب رگرسیونی β در نظر گرفت. تابع چگالی نمایی دوگانه را می‌توان به‌عنوان ترکیبی از تابع چگالی نرمال مقیاس‌دار که فراسنجه واریانس آن دارای توزیع نمایی است، به‌دست آورد. بنابراین می‌توان آنرا به‌صورت زیر نوشت:

$$\beta_j\sim DE(\beta_j|\lambda)=\frac{\lambda}{2}e^{-\lambda|\beta_j|}=\int_0^{\infty}\left[\frac{exp\left(-\left(\beta_j^2/2\sigma_j^2\right)\right)}{\sqrt{2\pi\sigma_j^2}}\right]\left[-\frac{\lambda^2}{2}exp\left(-\frac{\lambda^2}{2}\sigma_j^2\right)\right]d\sigma_j^2$$

در این رابطه، β_j، اثر نشانگر نامعلوم j اُم بوده و

σ_j^2، فراسنجه واریانس مرتبط با β_j است. با استفاده از این رابطه، مدل سلسله مراتبی لاسو بیزی (BL) به‌صورت زیر پیشنهاد گردید:

$$P(y|\beta,\sigma_{\varepsilon}^2)=\prod_{i=1}^{n}N(y_i|X_i^T\beta,\sigma^2)$$

$$P(\beta,\sigma_{\varepsilon}^2,\tau^2,\lambda^2)=P(\beta|\sigma^2,\tau^2)P(\sigma^2)P(\tau^2|\lambda)P(\lambda^2)$$

$$=\left[\prod_{i=1}^{p}N(\beta_j|0,\tau_j^2\sigma_{\varepsilon}^2)\right]\chi^{-2}(\sigma^2|d.f,S)\times\left[\prod_{j=1}^{p}exp(\tau^2|\lambda)\right]G(\lambda^2|\alpha_1,\alpha_2)$$

در این رابطه، $N\left(y_i|X_i^T\beta,\sigma_{\varepsilon}^2\right)$ و $N\left(\beta_j|0,\tau_j^2\sigma_{\varepsilon}^2\right)$، به‌ترتیب، چگالی‌های نرمال با میانگین $X_i^T\beta$ و 0 و با واریانس‌های σ_{ε}^2 و $\tau_j^2\sigma_{\varepsilon}^2$ هستند. همچنین $\chi^{-2}\left(\sigma_{\varepsilon}^2|d.f,S\right)$، چگالی معکوس کای مربع با درجه‌آزادی df و فراسنجه مقیاس S، $\exp(\tau_j^2|\lambda)=(\lambda^2/2)\exp(-(\lambda^2/2)\tau_j^2)$، تابع چگالی نمایی و $G\left(\lambda^2|\alpha_1,\alpha_2\right)$، توزیع گاما با فراسنجه شکل α_1 و فراسنجه میزان α_2 است. de los Campos *et al.* (2009) کاربرد لاسو بیزی در انتخاب ژنومی را تشریح نمودند. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک لاسو بیزی در ضمیمه ۷ ذکر گردیده است.

الفبای بیزی

اگر این فرض پذیرفته شود که واریانس نشانگرها در جایگاه‌های ژنی مختلف، متفاوت است بنابراین لازم است که $\sigma_{\beta_j}^2$ برآورد گردد. روش‌های بیزی تحت عنوان بیز A و بیز B برای برآورد همزمان اثر نشانگرها و واریانس آنها پیشنهاد شده‌اند (Meuwissen *et al.*, 2001). در بیز A داده‌ها در دو سطح مدل‌بندی می‌شوند. اولین مدل در سطح داده‌ها به‌صورت $y=\mu 1_n+X_j\beta_j+e$ است. در این مدل، توزیع پیشین مربوط به اثر نشانگر در جایگاه jاُم به‌صورت $\beta_j \sim N(0,\sigma_{\beta_j}^2)$ است. دومین مدل در سطح واریانس اثر جایگاه‌های نشانگر می‌باشد. توزیع پیشین واریانس اثر جایگاه‌های نشانگر به‌صورت $\sigma_{\beta_j}^2 \sim \chi^{-2}(\nu,S)$ بوده که در آن S، فراسنجه مقیاس و ν، درجه آزادی است. در نهایت، توزیع پیشین حاشیه‌ای اثر نشانگر $P(\beta_j|\nu,S^2)$ در بیز A دارای یک هسته از چگالی توزیع $t(0,\nu,S^2)$ با میانگین صفر و فراسنجه مقیاس S^2 و درجه آزادی ν می‌باشد. مزیت استفاده از یک توزیع معکوس مربع کای برای واریانس جایگاه‌های نشانگر این است که دارای یک توزیع پیشین مزدوج است. بنابراین توزیع پسین آن نیز به‌صورت یک توزیع معکوس مربع کای با فراسنجه مقیاس $S+\beta_j^T\beta_j$ و درجه آزادی $\nu+n_j$ به‌صورت زیر تعریف می‌گردد:

$$p\left(\sigma_{\beta_j}^2|\beta_j\right)=\chi^{-2}\left(\nu+n_j,S+\beta_j^T\beta_j\right)$$

در این رابطه، n_j، تعداد اثرات هاپلوتیپ جایگاه نشانگر jاُم است. از این توزیع پسین به‌طور مستقیم برای برآورد $\sigma_{\beta_j}^2$ نمی‌توان استفاده کرد. زیرا شرط آثار β_j نامعلوم است. Meuwissen *et al.* (2001) از نمونه‌گیری گیبس برای برآورد آثار و واریانس‌ها استفاده کردند. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک بیز A در ضمیمه (۸) ارائه شده است.

در روش بیز B فرض بر این است که تمام نشانگرها به تنوع ژنتیکی کمک نمی‌کنند. در واقع، توزیع واریانس ژنتیکی در میان جایگاه‌های نشانگر به گونه‌ای است که تعداد زیادی از جایگاه‌های نشانگر بدون واریانس ژنتیکی بوده و تعداد نسبتاً اندکی از جایگاه‌های نشانگر دارای واریانس ژنتیکی هستند. بنابراین چگالی پیشین روش بیز A این موضوع را برنمی‌گرداند. زیرا در این روش، احتمال $\sigma_{\beta_j}^2=0$ بی‌نهایت کوچک است. Meuwissen *et al.* (2001) به این موضوع در روش بیز B پرداختند. در این روش از یک چگالی پیشین آمیخته که دارای یک چگالی بالای π در $\sigma_{\beta_j}^2=0$ و یک توزیع معکوس مربع کای برای $\sigma_{\beta_j}^2>0$ است، استفاده می‌شود. این توزیع پیشین به‌صورت زیر است:

$$\beta_j\left|\sigma_{\beta_j}^2\right.=\begin{cases}C & \sigma_{\beta_j}^2=0\\ N\left(0.\sigma_{\beta_j}^2\right) & \sigma_{\beta_j}^2>0\end{cases}$$

در نهایت، توزیع پیشین حاشیه‌ای اثر نشانگر یک توزیع آمیخته t به‌صورت زیر خواهد شد:

$$p\left(\beta_j|\pi\right)=\begin{cases}0 & \pi\\ t\left(0.\nu.S^2\right) & 1-\pi\end{cases}$$

نمونه‌گیر گیبس مورد استفاده در روش بیز A نمی‌تواند در روش بیز B استفاده شود. زیرا کل فضای نمونه را در بر نمی‌گیرد. این مشکل توسط نمونه‌گیری همزمان $\sigma^2_{\beta_j}$ و β_j با استفاده از الگوریتم متروپلیس-هستینگ (Metropolis–Hastings) برطرف خواهد شد. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک بیز B در ضمیمه ۹ ذکر گردیده است. مشکل روش‌های بیز A و بیز B در این است که آثار SNP ها حول صفر قرار دارند. این موضوع ناشی از آن است که آثار SNP با احتمال π برابر با صفر و با احتمال $1-\pi$ دارای توزیع $N\left(0,\sigma^2_{g_i}\right)$ هستند. $\sigma^2_{g_i}$ دارای توزیع پیشین معکوس مربع کای با درجه آزادی پایین و فراسنجه مقیاس S^2_a است. همچنین توزیع پسین آن نیز تنها یک درجه آزادی بیشتر از توزیع پیشین دارد. این موضوع با مفهوم بیزی ناسازگار است. بدین ترتیب انقباض اثر SNPها به شدت به S^2_a بستگی دارد. برای رفع این مشکل، Gianola *et al.* (2009) روش‌های بیز C و بیز D را پیشنهاد کردند. در روش بیز C برای تمام SNPها یک واریانس مشترک در نظر گرفته می‌شود. در روش بیز D نیز فراسنجه مقیاس توزیع $\chi^{-2}(\nu,S^2_a)$ به عنوان یک مقدار نامعلوم که خود دارای یک توزیع پیشین می‌باشد، در نظر گرفته می‌شود. یکی دیگر از نقاط ضعف بیز A و بیز B این است که مقدار احتمال π (احتمال اینکه اثر SNPها صفر باشد)، مشخص است. در بیز A، $\pi=0$ است. بنابراین تمام SNPها دارای اثر غیرصفر هستند. ولی در بیز B، $\pi>0$ است. بنابراین تعدادی از SNPها دارای اثر صفر می‌باشند. انقباض اثر SNP توسط فراسنجه π تحت تأثیر قرار می‌گیرد. در نتیجه برای تفسیر آن باید مقدارش نامعلوم باشد. برای رفع این مشکل Habier *et al.* (2011) دو روش بیز C_π و بیز D_π را که در آنها فراسنجه احتمال π به‌صورت مقداری مجهول است، پیشنهاد دادند. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک بیز C و بیز D در ضمیمه (۱۰) ارائه شده است.

روش‌های نافراسنجه‌ای در انتخاب ژنومی

در مواردی که شکل رابطه بین متغیر پاسخ و مجموعه پیش‌بینی کننده‌ها نامعلوم است، از روش‌هایی موسوم به روش‌های نافراسنجه‌ای استفاده می‌شود. زمانی که در بین ژن‌ها یا نشانگرها، تعاملات ژنی برقرار باشد، معمولاً از این روش‌ها برای پیش‌بینی ارزش‌های ارثی استفاده می‌شود.

رگرسیون هسته‌ای روی نشانگرهای SNP

یکی از روش‌های پرکاربرد در زمینه‌ی برآورد چگالی، روشی پیشنهادی توسط Silverman (1986) بر مبنای رگرسیون هسته‌ای می‌باشد. در اینجا هدف برآورد چگالی نامعلوم با استفاده از یک منحنی هموار است. رگرسیون هسته‌ای، شایع‌ترین روش مورد استفاده در برآوردهای نافراسنجه‌ای است. Gianola *et al.* (2006) تابع رگرسیونی برای انتخاب ژنومی را به‌صورت $y_i=g(x_i)+e_i$ پیشنهاد نمودند ($i=1,2,\ldots,n$). در این رابطه، y_i، مقدار فنوتیپ اندازه‌گیری شده (نظیر تولید شیر یا وزن بدن) برای فرد i اُم، X_i، بردار $p\times 1$ متشکل از نشانگرهای SNP در i اُمین مشاهده و $g(.)$ توابع نامعلومی که ژنوتیپ‌ها را به فنوتیپ‌ها مرتبط می‌کنند، است. $g(\mathrm{x}_i)=\mathrm{E}(\mathrm{y}_i\,|\,\mathrm{x}_i)$ تابع امید ریاضی شرطی بوده که بیانگر میانگین ارزش فنوتیپی تعداد نامتناهی از افراد که تمامی آنها دارای ژنوتیپ x_i هستند، می‌باشد. $e_i \sim (0,\sigma^2)$ یک باقیمانده تصادفی مستقل از x_i بوده که دارای واریانس σ^2 است. تابع امید ریاضی شرطی را می‌توان به‌صورت زیر نوشت:

$$g(x)=\frac{\int yp(x,y)dy}{p(x)}$$

برای برآورد $p(x)$ می‌توان از برآوردکننده هسته‌ای نیمه فراسنجه‌ای که توسط Silverman (1986) پیشنهاد شده است، استفاده کرد. این برآورد کننده به‌صورت زیر است:

$$\hat{p}(x)=\frac{1}{nh^p}\sum_{i=1}^{n}K\left(\frac{x_i-x}{h}\right)$$

$$\left(\int_{-\infty}^{\infty}\hat{p}(x)dx=1\right)$$

در اینجا، $\mathrm{K}((\mathrm{x_i}-\mathrm{X})/\mathrm{h})$، یک تابع هسته و h، پهنای باند یا فراسنجه هموارسازی است. همچنین x_i، ژنوتیپ SNP بعدی مشاهده شده برای فرد i اُم در نمونه است. بنابراین $\hat{p}(x)$ برآورد چگالی جمعیت یا

فراوانی‌ها است. به‌طور مشابه، می‌توان چگالی توام ژنوتیپ و فنوتیپ را در نقطه (x, y) به‌صورت زیر به‌دست آورد:

$$\hat{p}(x,y)=\frac{1}{nh^{p+1}}\sum_{i=1}^{n}K\left(\frac{y_i-y}{h}\right)K\left(\frac{x_i-x}{h}\right)$$

در این رابطه نیز $K((y_i-y)/h)$ یک تابع هسته است. علاوه بر این، y_i مقدار نمونه‌ی مشاهده شده از متغیر y در فرد i اُم است. Nadaraya (1964) و Watson (1964) با کمک معادله فوق نشان دادند که تابع امید ریاضی شرطی را می‌توان به‌صورت زیر نوشت:

$$\hat{E}(y|x)=\hat{g}(x)=\frac{\int y\hat{p}(x,y)dy}{\hat{p}(x_n)}=\frac{\frac{1}{nh^{p+1}}\sum_{i=1}^{n}K\left(\frac{y_i-y}{h}\right)K\left(\frac{x_i-x}{h}\right)}{\frac{1}{nh^{p}}\sum_{i=1}^{n}K\left(\frac{x_i-x}{h}\right)}=\sum_{i=1}^{n}y_i w_i(x)$$

که در آن:

$$w_i(x)=\frac{K\left(\frac{x_i-x}{h}\right)}{\sum_{i=1}^{n}K\left(\frac{x_i-x}{h}\right)}$$

این ترکیب خطی را برآوردکننده نادارایا-واتسن (Nadarya-Watson estimator) از تابع رگرسیونی می‌نامند. همان‌طور که در این ترکیب خطی دیده می‌شود، مقدار برازش‌شده در مختصات x یک میانگین وزنی از تمام نقاط ارائه می‌دهد. این برآوردکننده را یک برآوردکننده موضعی می‌نامند. زیرا مشاهدات با مختصات x_i نزدیک‌تر به نقطه کانونی x در محاسبه مقدار برازش‌شده $\hat{E}(y|x)$ وزن بیشتری می‌گیرند. چندین روش برای انتخاب فراسنجه هموارسازی h وجود دارد. بدین منظور، Gianola *et al.* (2006) از روش اعتبارسنجی حذف تکی (Leave-one-out) استفاده کردند. در این روش، ابتدا مشاهده‌ی i اُم (x_i, y_i) حذف شده و مدل روی n-1 مشاهده باقیمانده برازش می‌شود. سپس با استفاده از اطلاعات نشانگر، $\hat{g}_{i,-i}(x_i|h)$ پیش‌بینی می‌شود. در اینجا، $-i$ بیانگر تمامی داده‌ها به غیر از داده i اُم است. این فرآیند برای تمامی داده‌ها تکرار می‌شود. معیار اعتبارسنجی متقابل را می‌توان به‌صورت زیر به‌دست آورد:

$$CV(h)=\frac{\sum_{i=1}^{n}\left[y_i-\hat{g}_{i,-i}(x_i|h)\right]^2}{n}$$

برآورد اعتبارسنجی متقابل h عبارت است از مقداری از h که $CV(h)$ را به کمترین مقدار خود برساند. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک رگرسیون هسته‌ای روی نشانگرهای SNP در ضمیمه (۱۱) ذکر گردیده است.

رگرسیون فضای هیلبرت با هسته بازآفرین

de los Campos *et al.* (2009, 2010) روش رگرسیون فضای هیلبرت با هسته بازآفرین را برای ارزیابی ژنتیکی و پیش‌بینی ارزش‌های ژنتیکی برای صفات کمی مورد استفاده قرار دادند. برآوردهای حاصل از روش رگرسیون فضای هیلبرت با هسته بازآفرین (RKHS) را می‌توان به‌عنوان راه حلی برای یک مسئله بهینه سازی جریمه شده از توابع ارزش واقعی یا به عنوان نمای پسین در کلاس خاصی از مدل‌های بیزی در نظر گرفت. در رگرسیون RKHS مجموعه‌ای از توابع یا فضاها تعریف می‌شوند و رگرسیون توسط انتخاب یک هسته بازآفرین انجام می‌گیرد. هسته بازآفرین می‌تواند هر تابع معین مثبت نگاشت شده از جفت نقاط موجود در فضای ورودی به سمت فضای اعداد حقیقی به‌صورت $K(x_i,x_{i'}):\{(x_i,x_{i'})\rightarrow\Re\}$ باشد. در RKHS تابع رگرسیونی ترکیبی خطی از توابع اُریب ارائه شده توسط هسته بازآفرین $K(x_i,x_{i'})$ به‌صورت زیر است:

$$g(x_i)=\sum_{i'}K(x_i,x_{i'})\alpha_{i'}$$

مربع نرم این تابع توسط $\|g\|^2=\sum_i\sum_{i'}K(x_i,x_{i'})\alpha_{i'}$ داده می‌شود. اگر نمایش برداری $g=K\alpha$ در نظر گرفته شود، $\|g\|^2=\alpha^T K\alpha$ خواهد شد. در اینجا، $g=\{g_i\}$ ، $K=\{K_{ii'}=K(x_i,x_{i'})\}$ و $\alpha=\{\alpha_i\}$ است. به‌طورکلی، برآوردها در RKHS از طریق مجموع مربعات باقیمانده‌های جریمه‌شده $\hat{\alpha}=\underset{\alpha}{\arg\min}\left\{(y-K\alpha)^T(y-K\alpha)+\lambda\alpha^T K\alpha\right\}$ به‌دست می‌آید (عرض از مبدأ و آثار غیر واقعی برای سهولت در نمادگذاری حذف شده‌اند). در این رابطه، $(y-K\alpha)^T(y-K\alpha)$، مجموع مربعات باقیمانده‌ها، λ، فراسنجه تنظیم و $\alpha^T K\alpha$، یک جریمه روی پیچیدگی مدل است. برآوردهای حاصل از مسئله بهینه‌سازی فوق را می‌توان به‌صورت زیر نمایش داد:

$$\hat{\alpha}=\left[K^T K+\lambda K\right]^{-1}K^T y$$

پیش‌بینی‌ها نیز عبارتند از:

$$K\hat{\alpha} = K\left[K^T K + \lambda K\right]^{-1} K^T y = \left[I + \lambda K^{-1}\right]^{-1} y$$

در روش رگرسیون RKHS مشخصات مدل توسط دو عنصر اصلی تعریف می‌شود: ۱- انتخاب هسته بازآفرین که توسط تابعی اُریب و ضرب داخلی برای تعریف فضای هیلبرت تعریف می‌گردد و ۲- فراسنجه تنظیم λ که در رگرسیون ستیغی نشان‌دهنده فراسنجه انقباضی است. در حالت بیزی، روش رگرسیون جریمه‌هیده RKHS را می‌توان به‌صورت نمای پسین از بردار ضرایب رگرسیونی مدل بیزی زیر در نظر گرفت:

$$\begin{cases} y = K\alpha + \varepsilon \\ \begin{pmatrix} \varepsilon \\ \alpha \end{pmatrix} \middle| \sigma_\varepsilon^2, \sigma_\alpha^2 \sim N\left[0, \begin{pmatrix} I\sigma_\varepsilon^2 & 0 \\ 0 & K^{-1}\sigma_\alpha^2 \end{pmatrix}\right] \end{cases}$$

در این پژوهش روش رگرسیون فضای هیلبرت با هسته بازآفرین با استفاده از یک هسته گوسی که توسط مربع فاصله اقلیدسی بین ژنوتیپ‌ها به‌صورت زیر حاصل می‌آید:

$$K\left(x_i, x_{i'}\right) = exp\left\{-h \times \frac{\sum_{k=1}^{p}\left(x_{ik} - x_{i'k}\right)^2}{p}\right\}$$

روی مجموعه داده شبیه‌سازی شده برازش یافت. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک رگرسیون فضای هیلبرت با هسته بازآفرین در ضمیمه ۱۲ ارائه شده است.

رگرسیون ماشین بردار پشتیبان

Cortes & Vapnik (1995) نسخه‌ای از ماشین بردار پشتیبان را پیشنهاد دادند که به جای طبقه‌بندی، رگرسیون را اجرا می‌نمود. این نسخه به رگرسیون بردار پشتیبان (SVR) شهرت یافت. این روش برای انتخاب ژنومی در اصلاح نباتات نیز به‌کار برده شده است (Maenhout *et al.*, 2007; Long *et al.*, 2011). یکی از ویژگی‌های خوب رگرسیون ماشین بردار پشتیبان در برنامه‌های کاربردی اصلاح نباتات این است که رابطه بین ژنوتیپ نشانگرها و فنوتیپ‌ها را می‌توان با یک نگاشت خطی یا غیرخطی مدل‌سازی کرده و نمونه‌ها را از یک فضای پیش‌بینی در یک فضای ویژگی چند بعدی استخراج نمود. نگاشت $f(x): \mathbb{R}^p \to \mathbb{R}$ با توجه به مجموعه داده آموزشی $(x_1, y_1)\ldots\ldots(x_i, y_i)\ldots\ldots(x_n, y_n), x_i \in \mathbb{R}^p, y_i \in \mathbb{R}$ در نظر گرفته می‌شود. در اینجا، x_i یک بردار با p عضو شامل ارزش‌های ژنوتیپی p نشانگر SNP برای شخص i اُم و y_i ارزش فنوتیپی شخص i اُم است. فرض می‌شود که $f(x)$ تابعی خطی به‌صورت زیر است:

$$f(x) = b + < w, x >$$

در این تابع، b، یک فراسنجه ثابت و w برداری از وزن‌های نامشخص یا به‌عبارت دیگر ضرایب رگرسیونی است. ثابت b بیانگر حداکثر خطایی است که در هنگام برآورد وزن w رُخ می‌دهد. یادگیری $f(x)$ با به حداقل‌رساندن عبارت زیر به‌دست می‌آید:

$$\lambda\sum_{i=1}^{n} L\left(y_i - f\left(x_i\right)\right) + \frac{1}{2}w^2$$

در این رابطه، $L(.)$، تابع زیان، $\|w\|^2 = w^T w$، پیچیدگی مدل و λ یک فراسنجه تنظیم مثبت است که بین پیچیدگی مدل و پراکندگی آن یک موازنه برقرار می‌کند. افزایش λ منجر به جریمه بیشتر روی خطا می‌شود.

توابع زیان بسیاری نظیر زیان مربع، زیان قدر مطلق و زیان ε غیرحساس وجود دارند که در رگرسیون ماشین بردار پشتیبان مورد استفاده قرار می‌گیرند. در اینجا از تابع زیان ε غیرحساس استفاده می‌شود. در رگرسیون ماشین بردار پشتیبان ε غیرحساس $\varepsilon-$SVR هدف یافتن یک تابع $f(x)$ است که بیشترین اختلاف ε از مقدار هدف y برای هر نمونه‌ی آموزشی $0 \le i \le n$ را داشته باشد. در $\varepsilon-$SVR، L یک تابع زیان غیرحساس برای ε است که به‌صورت زیر معرفی می‌شود:

$$L\left(y - f\left(x\right)\right) = \begin{cases} 0 & \left|y - f\left(x\right)\right| < \varepsilon \\ \left|y - f\left(x\right)\right| - \varepsilon & 0.w \end{cases}$$

در این رابطه، ε، تعداد بردارهای پشتیبان مورد استفاده در تابع رگرسیونی است. طبق تعریف مطرح‌شده توسط Vapnik (1995)، یک بردار پشتیبان، بردار x_i است که در معادله $y_i(wx_i + b) = 1$ صدق می‌کند. با افزایش ε بردارهای پشتیبان کمتری در برازش مورد استفاده قرار می‌گیرند. تابع زیان ε غیرحساس، خطاهای رگرسیونی که اندازه‌ای کمتر از ε دارند را نادیده می‌گیرد. همچنین زمانی که خطا

بزرگ‌تر از εاست، زیان به‌صورت $|y-f(x)|-\varepsilon$ می‌شود. تابع زیان εغیرحساس نیازمند یک نمایش توانمندتری برای محاسبه مقدار خطا در داده‌ها است. بدین منظور می‌توان یک هزینه بیشتر یا یک عدم اطمینان بیشتر را با معرفی متغیرهای کمکی نامنفی محدود شده (ξ^* و ξ) به تابع زیان اضافه کرد. $\xi > 0$ و $\xi^* = 0$ برای نقاط داده با خطاهای مثبت و $\xi^* > 0$ و $\xi = 0$ برای نقاط داده با خطاهای منفی تعریف می‌شود. با در نظر گرفتن این دو فرض، مسئله بهینه‌سازی می‌تواند به‌صورت زیر فرمول‌بندی شود:

$$\min_{w,b,\xi,\xi^*}\left(\lambda\sum_{i=1}^{n}\left(\xi+\xi^*\right)\right)+\frac{1}{2}w^2$$

اگر قیدهای $\xi^*\geq 0, \xi\geq 0$، $y_i \leq f(x_i)+\varepsilon+\xi_i$ و $y_i \geq f(x_i)-\varepsilon-\xi_i^*$ (در تمامی موارد $i=1,\ldots,n$) در نظر گرفته شوند و اگر $\hat{w}$ و $\hat{b}$ حداقل کننده‌های مسئله بهینه‌سازی فوق باشند، حل مسئله بهینه‌سازی به‌صورت زیر خواهد شد:

$$\hat{w}=\sum_{i=1}^{n}\left(\hat{\alpha}_i-\hat{\alpha}_i^*\right)x_i$$

$$\hat{f}(x)=\sum_{i=1}^{n}\left(\hat{\alpha}_i-\hat{\alpha}_i^*\right)<x_i,x_j>+\hat{b}$$

برآورد $f(x)$ یک ترکیب خطی از ضرب داخلی $<x_i,x_j>$ است. بنابراین این تابع در داده‌های ورودی x خطی است. ضرایب α_i یا α_i^* تنها زمانی می‌توانند غیرصفر باشند که $|y-f(x)|\geq\varepsilon$ شود. بنابراین تمام نمونه‌های آموزشی در فرمول $f(x)$ مورد استفاده قرار نمی‌گیرند. با توجه به ماهیت قیدهای مطرح شده معمولاً فقط یک زیر مجموعه از مقادیر $(\hat{\alpha}_i-\hat{\alpha}_i^*)$ غیرصفر هستند و مقادیر داده مرتبط با آنها بردارهای پشتیبان نامیده می‌شوند. یک ویژگی رگرسیون بردار پشتیبان (SVR) این است که فرمول‌بندی و حل آن تنها به ضرب داخلی بستگی دارد.

تابع هسته مقدار ضرب داخلی بین دو بردار موجود در فضای ویژگی که می‌تواند ابعاد نامتناهی داشته باشد را برمی‌گرداند و بنابراین می‌تواند به جای ضرب داخلی در حل رگرسیون بردار پشتیبان استفاده شود. ویژگی هسته برای ماشین‌های بردار پشتیبان منحصر به فرد نیست. بنابراین از هسته‌های گوناگونی می‌توان استفاده کرد. به‌طور کلی یک هسته را می‌توان به‌صورت $k(x,z)=\phi(x)^T\phi(z)$ بیان کرد. در این رابطه، x و z، دو بردار در فضای اصلی و $\phi(x)$ و $\phi(z)$، بردارهایی در فضای ویژگی هستند. در این مطالعه از تابع هسته‌ای براساس شعاع گوسی (RBF گوسی) استفاده شده است. شکل کلی این هسته به‌صورت زیر است:

$$k(x,z)=\exp(-\sigma\|x-z\|^2)$$

در این رابطه، σ، فراسنجه پهنای باند است. یک فراسنجه خاص هسته است که اجازه می‌دهد تا یک تابع خطی در یک فضای ویژگی بی‌نهایت بزرگ پیدا شود. با توجه به انتخاب تابع هسته، مدل رگرسیون بردار پشتیبان (SVR) غیرخطی حاصل، ترکیبی خطی از تابع هسته‌ای به‌صورت زیر خواهد شد:

$$\hat{f}(x)=\sum_{i=1}^{n}\hat{\alpha}_i k(x_i,x_j)+\hat{b}$$

کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک رگرسیون ماشین بردار پشتیبان در ضمیمه (۱۳) ارائه شده است.

ساختار داده‌ها و تحلیل‌های آماری

در این پژوهش، تمامی روش‌های آماری مطرح‌شده روی یک مجموعه داده واقعی توسط نرم‌افزار R برازش داده شد. این مجموعه داده، شامل ۲۳۰۰ موش بود. فنوتیپ مورد نظر استاندارد نرمال شد یعنی صفتی پیوسته با میانگین صفر و انحراف معیار یک و دارای توزیع نرمال در محاسبات استفاده شد. فایل نشانگرهای مولکولی شامل ۱۷۰۰ نشانگر که در ۳۰ کروموزوم توزیع شده بودند. در نهایت، هر حیوان دارای ۵۱۰۰۰ نشانگر SNP بود. همچنین این مجموعه داده دربردارنده اطلاعات مربوط به ۲۳۰۰ QTL بوده است. اثر QTL در واقع ارزش‌های ارثی واقعی بودند که برای مطابقت دادن با مقادیر ارزش‌های ارثی برآورد شده جهت تعیین صحت پیش‌بینی مدل برازش شده روی داده‌ها مورد استفاده قرار گرفتند. داده‌های ژنوتیپی قبل از استفاده در تحلیل‌ها بایستی تحت کنترل کیفیت

قرار گرفتند. برای این مجموعه داده، نشانگرهایی که دارای فراوانی آللی کمتر از ۰/۰۵ بودند (MAF<0.05) از مجموعه آنالیزها خارج شدند، پس از کنترل کیفیت داده‌ها، تعداد نشانگرها از ۵۱۰۰۰ نشانگر به ۳۶۳۲۲ نشانگر کاهش یافتند. مدل‌های بیزی توسط الگوریتم MCMC با ۵۰۰۰ تکرار زنجیره مارکف و تعیین ۱۰۰۰ تکرار اولیه زنجیره به‌عنوان تکرارهای سوخته (burn-in) اجرا شدند. در روش‌های انقباضی فراوانی‌گرا فراسنجه جریمه λ توسط روش اعتبارسنجی متقابل ۵ دسته‌ای محاسبه شد. همچنین در روش‌های نافراسنجه‌ای مقدار فراسنجه هموارساز برابر با ۰/۵ قرار داده شد. روش‌های رگرسیونی فراسنجه‌ای برازش داده شده روی داده‌ها عبارت بودند از رگرسیون ستیغی، رگرسیون لاسو، روش الاستیک‌نت، روش GBLUP، رگرسیون ستیغی بیزی، لاسو بیزی، بیز A، بیز B، بیز C و بیز D. روش‌های رگرسیونی نافراسنجه نیز شامل روش‌های رگرسیون هسته‌ای با استفاده از برآوردگر نادارایا-واتسن، رگرسیون فضای هیلبرت با هسته باز آفرین گوسی در حالت بیزی و رگرسیون ماشین بردار پشتیبان با در نظر گرفتن هسته RBF گوسی بودند.

برای بررسی صحت پیش‌بینی مدل برازش شده روی داده‌ها از معیارهای میانگین مربعات خطای پیش‌بینی (MSE) و همبستگی پیرسون (ρ) بین مقادیر آثار QTL شبیه‌سازی شده به‌عنوان مقادیر ارزش‌های ارثی واقعی (TBV) و مقادیر ارزش‌های ارثی ژنومی پیش‌بینی شده (GEBV) استفاده گردید. تعداد متغیرهای مستقل یا SNP های موجود در مدل نیز به عنوان یک فاکتور جریمه در برازش مدل، در نظر گرفته شد. هر مدل آماری که دارای بیشترین صحت پیش‌بینی بود، برای ارزیابی ژنومی و پیش‌بینی ارزش‌های ارثی ژنومی به‌عنوان مدل برتر معرفی گردید. تمامی محاسبات در محیط R انجام شد و بسته‌های نرم‌افزاری مورد استفاده به منظور پیاده‌سازی روش‌های آماری مطرح شده روی داده‌ها عبارت از بسته نرم‌افزاری snpReady برای پیش‌پردازش داده‌های ژنومی به‌منظور آماده‌سازی آنها جهت استفاده در تحلیل‌ها (Granato *et al.*, 2018)؛ بسته نرم‌افزاری glmnet برای برازش روش‌های رگرسیون ستیغی، رگرسیون لاسو و روش الاستیک‌نت (Simon *et al.*, 2011)؛ بسته نرم‌افزاری rrBLUP برای برازش مدل‌های مختلط روی داده‌ها (Endelman, 2011)؛ بسته نرم‌افزاری BGLR برای برازش مدل‌های رگرسیون ستیغی بیزی، لاسو بیزی، بیز A، بیز B، بیز C، بیز D و روش رگرسیون فضای هیلبرت با هسته بازآفرین (Pérez & de los Campos, 2014)؛ بسته نرم‌افزاری kernlab برای برازش مدل بردار پشتیبان رگرسیونی روی داده‌ها (Karatzoglou *et al.*, 2004)؛ کدنویسی روش رگرسیون هسته‌ای در محیط R جهت برآورد ارزش‌های ارثی با کمک برآوردکننده نادارایا-واتسن، بودند.

نتایج و بحث

صحت حاصل از برازش مدل‌های آماری روی داده‌ها به‌طور خلاصه در جدول ۱ بیان شده است. همان‌طور که در این جدول نیز مشاهده می‌شود، از بین روش‌های انقباضی فراوانی‌گرای برازش شده روی داده‌ها، روش‌های لاسو و الاستیک‌نت به دلیل وجود قابلیت انتخاب متغیر در آنها، عملکرد بهتری نسبت به روش‌های رگرسیون ستیغی و GBLUP داشتند. زیرا این روش‌ها با انتخاب تعداد متغیرهای پیش‌بینی‌کننده (SNP های) کمتر، صحت پیش‌بینی بالاتری نسبت به روش‌های رگرسیون ستیغی و GBLUP دارند. از بین روش‌های رگرسیون ستیغی و GBLUP، روش GBLUP با شرکت دادن تعداد SNPهای بیشتر در مدل دارای همبستگی پیرسون بیشتر و خطای پیش‌بینی کمتر نسبت به روش رگرسیون ستیغی بوده و بنابراین دارای عملکرد پیش‌بینی بهتری می‌باشد. همچنین از میان روش‌های لاسو و الاستیک‌نت نیز، عملکرد روش لاسو بهتر است. زیرا با شرکت‌دادن تعداد متغیرهای پیش‌بینی‌کننده بیشتر نسبت به روش الاستیک‌نت، همبستگی بالاتر و خطای پیش‌بینی کمتری دارد.

جدول ۱. نتایج حاصل از مدل‌های آماری برازش‌یافته روی مجموعه داده شبیه‌سازی‌شده

Table 1. The results of fitted models on simulated data set

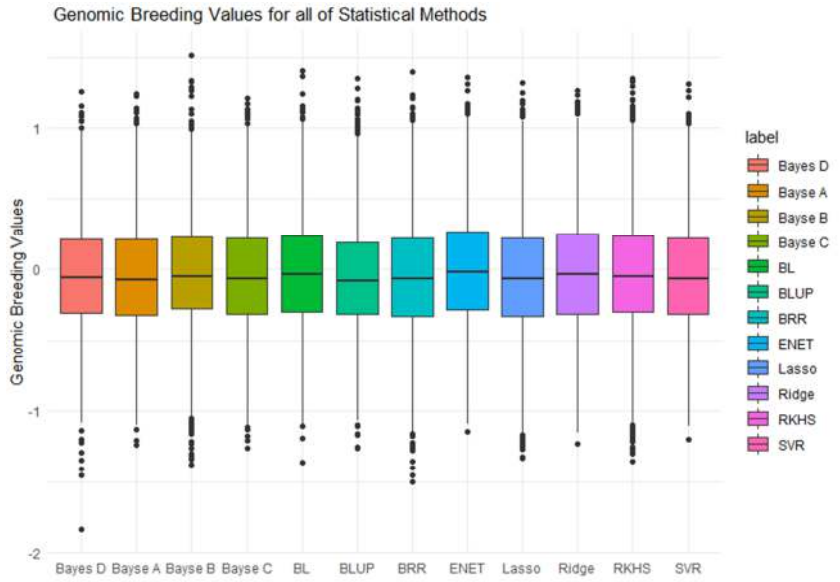
Models	Number of variables	Cor (GEBV, TBV)	MSE (GEBV)	Max GEBV	Selected animals
Ridge regression	36302	0.65	0.13	0.678	1914
Lasso regression	152	0.72	0.1	0.698	166
Elastic net method	134	0.71	0.14	0.448	1321
GBLUP	36322	0.66	0.11	0.956	1024
Bayesian lasso	292	0.66	0.11	0.975	1915
Bayesian ridge regression	285	0.66	0.24	1.357	1024
Bayes A	330	0.71	0.10	1.096	2263
Bayes B	280	0.73	0.21	1.416	2263
Bayes C	309	0.67	0.24	0.690	1024
Bayes D	260	0.72	0.69	0.249	2263
RKHS	673	0.59	0.16	1.410	1024
SVR	1674	0.58	0.18	1.336	1915

از میان روش‌های فراسنجه‌ای فراوانی‌گرای بیزی، روش بیز B با شرکت دادن تعداد SNPهای نسبتاً کمتر در مدل، دارای همبستگی پیرسون بیشتر و به‌طور تقریبی خطای پیش‌بینی کمتر نسبت به روش‌های دیگر است. بنابراین عملکرد پیش‌بینی روش بیز B مناسب‌تر از سایر روش‌های فراسنجه‌ای فراوانی‌گرای بیزی است.

از بین روش‌های نافراسنجه نیز روش رگرسیون فضای هیلبرت با هسته بازآفرین گوسی (RKHS)، با شرکت دادن تعداد SNPهای کمتر در مدل، دارای همبستگی پیرسون بیشتر و خطای پیش‌بینی کمتر نسبت به روش رگرسیون ماشین بردار پشتیبان می‌باشد. روش رگرسیون هسته‌ای نیز چون دارای انحراف استانداردی برابر با صفر است، برای برازش روی این داده‌ها مورد استفاده قرار نمی‌گیرد. از میان تمامی روش‌ها، روش بیز B و لاسو دارای عملکرد پیش‌بینی بالاتر نسبت به سایر روش‌ها هستند. از بین این دو روش نیز، روش بیز B با شرکت دادن تعداد SNP های بیشتر نسبت به روش لاسو، دارای همبستگی پیرسون بالاتر است.

همان‌گونه که در جدول ۱ ملاحظه می‌شود، روش‌های بیز B، بیز D و لاسو دارای بیشترین همبستگی بین مقادیر ارزش‌های ارثی واقعی (TBV) و مقادیر ارزش‌های ارثی ژنومی پیش‌بینی شده (GEBV) نسبت به سایر روش‌ها می‌باشند. همچنین روش‌های لاسو و بیز A در مقایسه با سایر روش‌ها، کمترین میانگین مربعات خطای پیش‌بینی را دارند. نمودار جعبه-ای در شکل ۱ نیز نشان‌دهنده ارزش‌های ارثی ژنومی می‌باشد. در یک نتیجه‌گیری کلی با توجه به نتایج ارائه‌شده در جدول ۱، به‌طور تقریبی روش بیز B دارای عملکرد پیش‌بینی بالاتری نسبت به روش‌های دیگر بوده است. با برازش مدل رگرسیونی بیز B که در بردارنده ۲۸۰ نشانگر SNP است، تعدای از حیوانات دارای بالاترین ارزش ارثی شدند که در روش‌های دیگر تعداد اندکی از این حیوانات وجود داشتند. هدف اصلاح نژاد تحت تأثیر نوع و ماهیت صفت تعیین شده و در برخی از صفات نظیر تعداد روزهای باز و میزان چربی لاشه، هر قدر ارزش ارثی به‌دست آمده پایین‌تر باشد، بهتر است. بنابراین این‌که کدام حیوان انتخاب شده و والد نسل بعد گردد، تابع ماهیت صفت است. انتظار می‌رود که با انجام انتخاب، میانگین ارزش ژنتیکی صفت در نسل بعد نسبت به نسل والدین تغییر کند. البته مساله انتخاب مساله مهمی در اصلاح نژاد بوده و عواقب افزایش همخونی ناشی از انتخاب نامناسب در جمعیت ممکن است منجر به از دست رفتن تنوع ژنتیکی و اُفت ناشی از همخونی گردد. در پژوهش‌های انتخاب ژنومی، تعداد متغیرهای مستقل بسیار بیشتر از اندازه نمونه است. بنابراین برآورد اثر نشانگرها با روش‌های معمول رگرسیونی امکان‌پذیر نمی‌باشد. از این‌رو، از روش‌های برآورد انقباضی فراوانی‌گرا و برآورد انقباضی بیزی و همچنین مدل‌های مختلط برای برآورد ارزش‌های ارثی استفاده می‌گردد.

از میان تمامی روش‌ها، روش‌های انقباضی توأم با انتخاب متغیر، شایستگی بیشتری در برآورد صحت انتخاب ژنومی نشان می‌دهند. شکل ۱ نشان می‌دهد که به‌طور میانگین، ارزش‌های ارثی برآورد شده برای تمام رویه‌ها یکسان است (این موضوع را از خط سیاه افقی در داخل جعبه‌ها می‌توان دریافت).



شکل ۱. نمودار جعبه‌ای مربوط به مقادیر ارزش‌های ارثی ژنومی برآورد شده برای تمامی حیوانات
Figure 1. Box chart of genomic estimated breeding values (GEBVs) for all animals

اما میزان تنوع در ارزش‌های ارثی برآورد شده متفاوت است. به عنوان مثال، این تفاوت را در بیز D می‌توان مشاهده نمود. روش‌های لاسو، لاسو بیزی، الاستیک‌نت، بیز B، بیز C و بیز D از جمله روش‌های تاوانی توأم با انتخاب متغیر هستند. Abdollahi-Arpanahi *et al.* (2013) به‌منظور مقایسه روش‌های گوناگون آماری در پیش‌بینی ارزش‌های ارثی ژنومی برای صفاتی با معماری ژنتیکی متفاوت از نظر توزیع آثار ژنی و نیز تعداد متفاوت جایگاه‌های صفت کمی (QTLها) از داده‌های حاصل از شبیه‌سازی استفاده کردند. به این منظور، ژنومی حاوی ۵۰۰ نشانگر تک نوکلئوتیدی (SNP) دو آللی روی کروموزومی به‌طول ۱۰۰ سانتی‌مورگان شبیه‌سازی شده و نه صفت با معماری ژنتیکی متفاوت ایجاد گردید.

به منظور پیش‌بینی ارزش‌های ارثی ژنومی از شش روش بهترین پیش‌بینی ناریب خطی ژنومی (GBLUP)، رگرسیون ستیغی (RRBLUP)، بیز A، بیز B، بیز C_π و بیز L استفاده شد. نتایج این پژوهش نشان داد که در مجموع روش‌های بیزی و GBLUP از نظر صحت ارزش‌های ارثی پیش‌بینی شده در مقایسه با روش RRBLUP عملکرد بهتری دارند.

همچنین هنگامی‌که معماری صفات بررسی شده از مدل تعداد زیاد جایگاه ژنی پیروی نکند، معمولاً روش‌های بیزی بر روش‌های GBLUP و RRBLUP ارجحیت دارند. Mohammadi *et al.* (2015) برای محاسبه دقت پیش‌بینی ژنومی برخی روش‌های فراسنجه‌ای و نافراسنجه‌ای از داده‌های ۳۴۵ گاو هلشتاین استفاده کردند. در این پژوهش از دو روش فراسنجه‌ای بهترین پیش‌بینی ناأریب خطی ژنومی (GBLUP) و بیز B و دو روش نافراسنجه‌ای فضای هیلبرت با هسته باز آفرین (RKHS) و شبکه‌های عصبی (NN) برای برآورد اثر نشانگری و پیش‌بینی دقت ارزش‌های ارثی ژنومی استفاده گردید. در مقایسه با تمامی روش‌ها دقت بیز B بیشتر و همچنین میانگین مربعات خطای پیش‌بینی روش GBLUP نیز کمتر برآورد شد. در نهایت، مدل رگرسیون بیز B برای انتخاب ژنومی در این جمعیت مطلوب بود. Teimurian *et al.* (2016) روش‌های آماری در انتخاب ژنومی را با کمک داده های شبیه سازی شده مربوط به گاوهای هلشتاین بررسی کردند. در این مطالعه مدل GBLUP و چهار مدل بیز A ، B و $C\pi$ و لاسو در حالت‌های مختلفی از جمعیت مرجع، تعداد جایگاه‌های صفات کمی و تراکم نشانگری مقایسه شدند. نتایج حاصل نشان دهنده ارجحیت روش‌های بیزی نسبت به روش GBLUP بود. در این مطالعه، برای انتخاب ژنومی در

جمعیت گاوهای هلشتاین ایران، به‌کارگیری مدل بیز B پیشنهاد شده است.

در مطالعه‌ای دیگر، Moradi *et al.* (2016) سه روش فراسنجه‌ای (GBLUP، بیز B و RKHS) و دو روش باز نمونه‌گیری (Bagging GBLUP و Random Forest) در پیش‌بینی ارزش‌های ارثی ژنومی برای صفاتی با ساختار ژنتیکی متفاوت را با هم مقایسه نمودند. در این مطالعه از داده‌های حاصل از شبیه‌سازی استفاده شده و برای مقایسه توانایی پیش‌بینی روش‌های آماری از معیار همبستگی بین ارزش‌های ارثی پیش‌بینی‌شده و واقعی و همچنین رگرسیون ارزش ارثی واقعی بر پیش‌بینی شده استفاده شد. به‌طور کلی نتایج بیانگر برتری عملکرد دو روش GBLUP و بیز B نسبت به دیگر روش‌ها بودند. Hosseini-Vardanjani *et al.* (2018) عملکرد انتخاب ژنومی را برای صفات تولید شیر، درصد چربی و درصد پروتئین روی ۲۱۱ گاو نژاد نجدی در گله‌های ایستگاهی و اقماری با استفاده از مدل‌های آماری مختلف ارزیابی کردند. در این پژوهش از چهار مدل پیش‌بینی نااُریب خطی ژنومی مقیاس شده با فراوانی آللی مشاهده شده (GBLUP)، فراوانی آللی ۰/۵ (G05BLUP)، بیز A و بیز B برای ارزیابی قابلیت پیش‌بینی استفاد شد. همچنین از دو معیار صحت ارزش‌های ژنومی ارثی و میزان اُریبی پیش‌بینی شده برای مقایسه بین روش‌های آماری استفاده گردید. نتایج حاصل نشان داد که روش بیز B دارای بیشترین صحت برآورد ارزش‌های ارثی برای صفات تولید شیر و درصد چربی بوده است. این نتیجه گیری در تقابل با پژوهش‌های دیگر است. برای درصد پروتئین روش‌های بیزی و GBLUPها دارای صحت مشابهی بودند. در بین تمام روش‌ها، روش بیز A و در بین صفات مختلف، تولید پروتئین کمترین اُریبی را در پیش‌بینی داشتند. از نظر مفهومی و بنیادی، مدل‌های نافراسنجه‌ای به‌صورت $y_i=f(x_i) + e_i$ ارایه می‌شوند. به عبارت دیگر، توزیع صفت توسط تابعی به اضافه خطا ارایه می‌شود. اگر بتوان $f(x_i)$ را به‌صورت به‌صورت مضربی از فراسنجه و متغیر مستقل نشان داد (مانند یک مدل رگرسیون ساده نرمال) آن‌گاه شکل نافراسنجه‌ای به فراسنجه‌ای تبدیل می‌شود. در بیشتر موارد پژوهشگر بر اساس روش‌های دیداری و آزمون‌های آماری، متوجه می‌شود که آیا توزیع صفت فراسنجه‌گرا است یا خیر؟ اگر این موضوع برای پژوهش‌گر روشن نشود، ممکن است که از روش‌های نافراسنجه‌ای استفاده کند. در آن صورت پیداکردن $f(x_i)$ با توجه به فرض ژنتیکی مورد نظر (نظیر بررسی آثار افزایشی، غالبیت و اپیستازی) کار سختی خواهد بود. شاید به این دلیل باشد که معمولا در مدل‌سازی روش‌های نافراسنجه‌ای کمتر از روش‌های فراسنجه‌ای به‌کار گرفته می‌شوند.

نتیجه گیری

هدف از پژوهش حاضر، بررسی ژنومی صفتی خاص یا توسعه مدل آماری ژنومی جدیدی نبوده است. همانطور که گفته شد این پژوهش بیش از هر چیز مبتنی بر گردآوری مدل‌های آماری رایج ژنومی مختلف و ارایه آنها همراه با کُدهای R بوده است. در اغلب موارد در پژوهش‌های انتخاب ژنومی، تعداد نشانگرها بسیار بیشتر از حجم نمونه است. بنابراین، برآورد اثر نشانگرها با روش‌های رگرسیونی ساده متداول امکان‌پذیر نمی‌باشد. از این رو، از روش‌های برآورد انقباضی برای برآورد ارزش‌های ارثی استفاده می‌گردد. از میان تمامی روش‌های انقباضی، روش‌های انقباضی توأم با انتخاب متغیر، شایستگی بیشتری در برآورد صحت انتخاب ژنومی نشان می‌دهند که به‌طور خاص، در این مطالعه روش بیز B دارای شایستگی پیش‌بینی ارزش‌های ارثی ژنومی بیشتر نسبت به سایر روش‌های مورد بررسی بود. این نتیجه در تضاد با نتایج به‌دست آمده از روش‌های الفبای بیزی در برخی از پژوهش‌های دیگران است. یکی دلایل این تضاد می‌تواند ساختار احتمالی عدم تعادل پیوستگی نشانگرها با SNP ها باشد. لذا پیشنهاد می‌شود که این مدل‌ها روی داده‌های شبیه‌سازی شده اعمال گردیده و نتایج آنها گزارش شود.

REFERENCES

1. Abdollahi-Arpanahi, R., Pakdel, A., Nejati-Javaremi, A. & Moradi Shahrbabak, M. (2013). Comparison of genomic evaluation methods in complex traits with different genetic architecture. *Journal of Animal Production,* 15(1), 65-77. (In Farsi)
2. Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S. & Lawlor, T. J. (2010). Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score1. *Journal of Dairy Science,* 93, 743-752.
3. Aliloo, H., Pryce, J. E., González-Recio, O., Cocks, B. G., Goddard, M. E. & Hayes, B. J. (2017). Including nonadditive genetic effects in mating programs to maximize dairy farm profitability. *Journal of Dairy Science,* 100, 1203-1222.
4. An, N.-R., Lee, S.-S., Park, J.-E., Chai, H.-H., Cho, Y.-M. & Lim, D. (2017). Current status of genomic prediction using Multi-omics data in livestock. *Journal of Biomedical and Translational Research,* 18, 151-156.
5. Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., Tiagi, A., Mushtaq, M., Jain, N., Singh, P. K., Singh, G. P. & Prabhu, K. V. (2016). Genomic selection in the Era of next generation sequencing for complex traits in plant breeding. *Frontiers in Genetics,* 7(221), 1-11.
6. Christensen, O.F. & Lund, M.S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution,* 42(2), 1-8.
7. Christensen, O. F., Madsen, P., Nielsen, B., Ostersen, T. & Su, G. (2012). Single-step methods for genomic evaluation in pigs. *Animal*, 6, 1565-1571.
8. Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning,* 20, 273-297.
9. Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de Los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J. & Varshney, R. K. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science,* 22, 961-975.
10. de los Campos, G., Gianola, D. & Rosa, G.J.M. (2009). Reproducing Kernel Hilbert Spaces Regression: a General Framework for Genetic Evaluation. *Journal of Animal Science,* 87(6), 1883.
11. de los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A. & Crossa, J. (2010). Semi-parametric Genomic-enabled Prediction of Genetic Values Using Reproducing Kernel Hilbert Spaces Methods. *Genetics Research,* 92(04), 295-308.
12. de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Crossa, J., Legarra, ,.AManfredi, E., Weigel, K. & Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigrees. *Genetics,* 182(1),: 375-385.
13. Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S. & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE,* 6, e19379.
14. Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R Package rrBLUP. *Plant Genome,* 4, 250-255.
15. Fikere, M., Barbulescu, D. M., Malmberg, M. M., Shi, F., Koh, J. C. O., Slater, A. T., MacLeod, I. M., Bowman, P. J., Salisbury, P. A., Spangenberg, G. C., Cogan, N. O. I. & Daetwyler, H. D. (2018). Genomic prediction using prior quantitative trait loci information reveals a large reservoir of underutilised blackleg resistance in diverse canola (*Brassica napus* L.) lines. *Plant Genome,* 11(2), 1-16.
16. Gao, N., Martini, J. W. R., Zhang, Z., Yuan, X., Zhang, H., Simianer, H., *et al*. (2017). Incorporating gene annotation into genomic prediction of complex phenotypes. *Genetics,* 207, 489-501.
17. Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E. & Fernando., R. (2009). Additive genetic variability and the bayesian alphabet. *Genetics,* 183, 347-363.
18. Gianola, D., Fernando, R. L. & Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics,* 173, 1761-1776.
19. Goddard, M. E. & Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics*, 124(6), 323-330.
20. Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, 136, 245-257.
21. Granato, I. S. C., Galli, G., de Oliveira Couto, E. G., Souza, M. B., Mendonça, L. F. & Fritsche-Neto, R. (2018). snpReady: a tool to assist breeders in genomic analysis. *Molecular Breeding*, 38, 102.
22. Habier, D., Fernando, R. L., Kizilkaya, K. & Garrick., D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics,* 12, 186.
23. Hayes, B. J., Bowman, P. J., Chamberlain, A. J. & Goddard, M. E. (2009a). Invited review: genomic selection in dairy cattle: progress and challenges. *Journal of Dairy Science,* 92, 433-443.

24. Hayes, B. J., Visscher, P. M. & Goddard, M. E. (2009b). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research,* 91, 47-60.
25. Hayes, B. J., Corbet, N. J., Allen, J. M., Laing, A. R., Fordyce, G., Lyons, R., McGowan, M. R. & Burns, B. M. (2019). Towards multi-breed genomic evaluations for female fertility of tropical beef cattle. *Journal of Animal Science,* 97(1), 55-62.
26. Henderson, C.R. (1949). Estimates of changes in herd environment. *Journal of Dairy Science,* 32, 706.
27. Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423-447.
28. Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, 12, 55-67.
29. Hosseini-Vardanjani, S. M., Shariati, M. M., Moradi Shahrebabak, H. & Tahmoorespur, M. (2018) The accuracy of genomic predictions for milk related traits in Najdi cattle breed. *Animal Science Journal (Pajouhesh & Sazandegi),* 122, 93-104. (In Farsi)
30. Jonas, E. & de Koning, D.-J. (2015). Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. *Frontiers in Genetics,* 6(49), 1-8.
31. Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, 11(9), 1-20.
32. Long, N., Gianola, D., Rosa, G.J.M. & Weige, K.A. (2011). Application of support vector regression to genome-assisted prediction of quantitative traits. *Theoretical and Applied Genetics,* 123, 1065-1074.
33. Maenhout, S., De Baets, B., Haesaert, G. & Van Bockstaele, E. (2007). Support vector machine regression for the prediction of maize hybrid performance. *Theoretical and Applied Genetics,* 115, 1003-1013.
34. Meuwissen, T. H. E., Hayes, B. J. & Goddard., M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics,* 157, 1819-1829.
35. Meuwissen, T., Hayes, B. & Goddard, M. (2016). Genomic selection: a paradigm shift in animal breeding. *Animal Frontiers,* 6, 6-14.
36. Misztal, I., Vitezica, Z. G., Legarra, A., Aguilar, I. & Swan, A. A. (2013). Unknown-parent groups in single-step genomic evaluation. *Journal of Animal Breeding and Genetics,* 130, 252-258.
37. Mohammadi, Y., Shariati, M. M., Zerehdaran, S., Razmkabir, M., Sayyadnejad, M.B. & Zandi, M.B. (2015). The accuracy of genomic breeding value for production trait in Iranian Holstein Dairy Cattle using parametric and non-parametric methods. *Journal of Animal Production,* 11(1), 1-11. (In Farsi)
38. Moradi, M., Abdollahi-Arpanahi, R., Hemmati, B. & Lavvaf, A. (2016). Comparison of parametric and resampling methods in genetic evaluation of quantitative traits with different genetic structure. *Journal of Animal Production,* 19(1), 1-12. (In Farsi)
39. Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R. & Visscher, P. M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genetics,* 11, e1004969.
40. Nadaraya, E. A. (1964) On Estimating Regression. *Theory of Probability and Application,* 9, 141-142.
41. Ogutu, J. O., Schulz-Streeck, T. & Piepho, H. P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, 6 (Suppl 2), S10. doi: 10.1186/1753-6561-6-S2-S10.
42. Pérez, P., de los Campos, G., Crossa, J. & Gianola, D. (2010). Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. *Plant Genome,* 3, 106-116.
43. Pérez, P. & de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics,* 198(2), 483-495.
44. Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. Statistical Science, 6(1), 48-51.
45. Schrag, T.A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S. & Melchinger, A.E. (2018). Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. *Genetics,* 208, 1373-1385.
46. Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.
47. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, 39(5), 1-13.
48. Teimurian, M., Shariati, M.M. & Aslaminejad, A.A. (2016). Comparison of Methods for the Implementation of Genomic Selection in Holstein. *Research on Animal Production,* 7(14), 198-203. (In Farsi)
49. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B,* 58, 267-288.

50. VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science,* 91, 4414-4423.
51. Vapnik, V. (1995). *The nature of statistical learning theory*. (2nd ed.). Springer.
52. Varona, L., Legarra, A., Toro, M.A. & Vitezica, Z.G. (2018). Non-additive effects in genomic selection. *Frontiers in Genetics,* 9(78), 1-12.
53. Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A,* 26(4), 359-372.
54. Weller, J. I., Ezra, E. & Ron, M. (2017). Invited review: a perspective on the future of genomic selection in dairy cattle. *Journal of Dairy Science*, 100, 8633-8644.
55. Wimmer, V., Lehermeier, C., Albrecht, T., Auinger, H.-J., Wang, Y. & Schön, C.-C. (2013). Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics,* 195, 573-587.
56. Whittaker, J. C., Thompson, R., and Denham, M. C. (1999). Marker-assisted selection using ridge regression. *Annals of Human Genetics,* 63, 366-366.
57. Yi, N. & Xu, S. (2008). Bayesian LASSO for quantitative trait loci mapping. *Genetics,* 179, 1045-1055.
58. Zeng, P. & Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nature Communications,* 8(456), 1-11.
59. Zhang, X., Lourenco, D., Aguilar, I., Legarra, A. & Misztal, I. (2016). Weighting strategies for single-step genomic BLUP: an iterative approach for accurate calculation of GEBV and GWAS. *Frontiers in Genetics,* 7(743), 1-14.
60. Zhou, X., Carbonetto, P. & Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics,* 9, e1003264.
61. Zou, H. & Hastie, T. (2005). Regularization and variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B,* 67(2), 301-320.

ضمائم

ضمیمه ۱. کُدنویسی در نرم‌افزار R جهت خوانش داده‌ها، پیش‌پردازش بر مبنای فراوانی آلل‌های جزیی و محاسبه صحت پیش‌بینی مدل‌های آماری برازش شده روی داده‌ها به کمک روش اعتبار سنجی متقابل ۵ دسته‌ای

```
# Reading Data
rm(list = ls())
geno = read.table("E:\\Dataset\\Marker Genotype.txt", header = T)  # Reading Genotype Data
geno = as.matrix(geno); gen = geno[, -1]
pheno =  read.table("E:\\Dataset\\Ped and Phenotype.txt", header = T)  # Reading Phenotype Data
ph = pheno$Phen

# Preprocessing with Minor Allel Frequancy
n = length(ph)
MA = colSums(gen)
MAF = MA / (2 * n)
w = which(MAF < 0.05)
pregen = gen[ ,-w]

#  K-fold Cross Validation
set.seed(100)
k = 5
folds = sample(1:k, length(ph), replace = T)
xtrain = pregen[folds!=k,]    # training data
ytrain = ph[folds!=k]          # training data
xtest = pregen[folds==k,]   # testing data
ytest = ph[folds==k]          # testing data
```

ضمیمه ۲. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک رگرسیون ستیغی

```
# Ridge Regression Method
library(glmnet)  # Package for Shirinkage Methods
set.seed(100)
grid = 10^seq(10, -2, length = 100)                  # we can compute model fits for a
```

```
# particular value of λ that is not one of the original grid values.
ridge.mod = glmnet(xtrain, ytrain, alpha = 0, lambda = grid)
plot(ridge.mod)
# 5-folds Cross Validation
ridge.cv.out.5 = cv.glmnet(xtrain, ytrain, alpha = 0, nfolds = 5, lambda = grid)
plot(ridge.cv.out.5)
bestlam.5.r = ridge.cv.out.5$lambda.min              # Best Value of λ
ridge.pred = predict(ridge.mod, s = bestlam.5.r, newx = xtest)
mseRCV = round(mean((ridge.pred - ytest)^2), 2) # MSE associated with best value of λ
corRCV = round(cor(ridge.pred, ytest), 2)          # Correlation Value
ridge.out = glmnet(pregen, ph, alpha = 0, lambda = grid)
plot(ridge.out)
ridge.coef.5 = predict(ridge.out, type = "coefficients", s = bestlam.5.r)
ridge.coefficients.5 = ridge.coef.5[ridge.coef.5 != 0]
sum(ridge.coef.5 == 0) # some of the coefficients exactly equal to zero
sum(ridge.coef.5 != 0)

# Breeding values with Ridge Regression Method
pred = ridge.coef.5[-1]
muR = ridge.coef.5[1]
GEBVR = pregen %*% pred
sort(GEBVR)
GEBVRmax = GEBVR[which.max(GEBVR)]; GEBVRmax
which(GEBVR == GEBVRmax)
BVR = predict(ridge.out, newx = pregen, s = bestlam.5.r)
BVRmax = BVR[which.max(BVR)]
BVRmax
which(BVR == BVRmax)
MSERidge = round(mean((GEBVR - QTL)^2), 2)   # MSE value
CorRidge = round(cor(GEBVR, QTL), 2)              # Correlation value
```

ضمیمه ۳. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک لاسو

```
# Lasso Regression Method
set.seed(100)
lasso.mod = glmnet(xtrain, ytrain, alpha = 1, lambda = grid)
plot(lasso.mod)
lasso.cv.out.5 = cv.glmnet(xtrain, ytrain, alpha = 1, nfolds = 5, lambda = grid)
plot(lasso.cv.out.5)
bestlam.5.l = lasso.cv.out.5$lambda.min
lasso.pred = predict(lasso.mod, s = bestlam.5.l, newx = xtest)
mseLcv = round(mean((lasso.pred - ytest)^2), 2)
corLCV = round(cor(lasso.pred, ytest), 2)
lasso.out = glmnet(pregen, ph, alpha = 1, lambda = grid)
plot(lasso.out)
lasso.coef.5 = predict(lasso.out, type = "coefficients", s = bestlam.5.l)
lasso.coefficients.5 = lasso.coef.5[lasso.coef.5 != 0]
sum(lasso.coef.5 != 0); sum(lasso.coef.5 == 0)

# Breeding values with Lasso Regression Method
predL = lasso.coef.5[-1]
muL = lasso.coef.5[1]
GEBVL = pregen %*% predL
sort(GEBVL)
GEBVLmax = GEBVL[which.max(GEBVL)]
GEBVLmax
which(GEBVL == GEBVLmax)
BVL = predict(lasso.out, newx = pregen, s = bestlam.5.l)
BVLmax = BVL[which.max(BVL)]; BVLmax
MSELasso = round(mean((GEBVL - QTL)^2), 2)   # MSE value
CorLasso = round(cor(GEBVL, QTL), 2)              # Correlation value
```

ضمیمه ۴. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک الاستیک‌نت

```
# Elastic Net Regression Method
set.seed(100)
ENET.mod = glmnet(xtrain, ytrain, alpha = 0.5, lambda = grid)
ENET.cv.out = cv.glmnet(xtrain, ytrain, alpha = 0.5, nfolds = 5, lambda = grid)
plot(ENET.cv.out)
bestlam = ENET.cv.out$lambda.min
ENET.pred = predict(ENET.mod, s = bestlam, newx = xtest)
mseENCV = round(mean((ENET.pred - ytest)^2), 2)
corENCV = round(cor(ENET.pred, ytest), 2)
ENET.out = glmnet(pregen, ph, alpha = 0.5, lambda = grid)
plot(ENET.out)
ENET.coef = predict(ENET.out, type = "coefficients", s = bestlam)
ENET.coefficients = ENET.coef[ENET.coef != 0]
sum(ENET.coef != 0); sum(ENET.coef == 0)

#Breeding values with Elastic Net Method
predEN = ENET.coef[-1]
muEN = ENET.coef[1]
GEBVEN = pregen %*% predEN
GEBVENmax = GEBVEN[which.max(GEBVEN)]
GEBVENmax
BVEN = predict(ENET.out, newx = pregen, s = bestlam)
which.max(BVEN)
BVENmax = BVEN[which.max(BVEN)]
BVENmax
MSEENET = round(mean((GEBVEN - QTL)^2), 2)      # MSE value
CorENET = round(cor(GEBVEN, QTL), 2)            # Correlation value
```

ضمیمه ۵. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک BLUP

```
# BLUP Method
library(rrBLUP)  # Package for BLUP Method
impute = A.mat(pregen, max.missing = 0.5, impute.method = "mean", return.imputed = T)
XX_impute = impute$imputed
set.seed(100)
k = 5
folds = sample(1:k, nrow(XX_impute), replace = TRUE)
xtest = XX_impute[folds==5,]; ytest = ph[folds==5]
xtrain = XX_impute[folds!=5,]; ytrain = ph[folds!=5]
yield = ytrain
yield_answer = mixed.solve(yield, Z = xtrain, K = NULL, SE = F, return.Hinv = F)
YLD = yield_answer$u
sum(YLD != 0); sum(YLD == 0)
m = as.matrix(YLD)
pred_yield_valid = xtest %*% m
pred_yield = pred_yield_valid + rep(yield_answer$beta, length(pred_yield_valid))
yield_valid = ytest
YLD_accuracy = cor(pred_yield_valid, yield_valid, use = "complete")
corBLUPCV = round(YLD_accuracy, 3)          # Correlation value
rss = round(mean((pred_yield - ytest)^2), 3)      # MSE value

#  Breeding Values for RRBLUP
yield_answerT = mixed.solve(ph, Z = pregen, K = NULL, SE = F, return.Hinv = F)
YLDT = yield_answerT$u
mT = as.matrix(YLDT)
pred_yield_T = pregen %*% mT
GEBVblup = pred_yield_T
```

```
GEBVblupmax = GEBVblup[which.max(GEBVblup)]
bvBlup = pred_yield_T + rep(yield_answerT$beta, length(pred_yield_T))
bvBlupmax = bvBlup[which.max(bvBlup)]
bvBlupmax
MSEBLUP = round(mean((GEBVblup - QTL)^2), 2)      # MSE value
CorBLUP = round(cor(GEBVblup, QTL), 2)            # Correlation value
```

ضمیمه ۶. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک رگرسیون ستیغی بیزی

```
library(BGLR)   # Package for Bayesian Methods
set.seed(1)
samplesize = 479
whichNa = sample(1:length(ph), size = samplesize, replace = FALSE)
yNa = ph
yNa[whichNa] = NA
nIter = 6000 ; burnIn = 1000
# Bayesian Ridge Regression
ETA = list(MRK = list(X = pregen, model = "BRR"))
fmBRR = BGLR(y = yNa, ETA = ETA, nIter = nIter, burnIn = burnIn)
r_BRR = cor(ph[whichNa], fmBRR$yHat[whichNa])                    # Correlation value
MSEbrr = round(mean((ph[whichNa] - fmBRR$yHat[whichNa])^2), 2)   # MSE value

# Breeding Values with Bayesian Ridge Regression
ETA$MRK$model = "BRR"
fmBRRT = BGLR(y = ph, ETA = ETA, nIter = nIter, burnIn = burnIn, saveAt = "BRRT_ ")
bvBRR = pregen %*% fmBRRT$ETA$MRK$b
which.max(bvBRR)
bvBRRmax = bvBRR[which.max(bvBRR)]
 bvBRRmax
MSEBRR = round(mean((bvBRR - QTL)^2), 2)          # MSE value
CorBRR = round(cor(bvBRR, QTL), 2)                # Correlation value
```

ضمیمه ۷. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک لاسو بیزی

```
# Bayesian Lasso
ETA$MRK$model = "BL"
fmBL= BGLR(y = yNa, ETA = ETA, nIter = nIter, burnIn = burnIn, saveAt = "BL_ ")
r_BL = cor(ph[whichNa], fmBL$yHat[whichNa])                      # Correlation value
MSEbl = round(mean((ph[whichNa] - fmBL$yHat[whichNa])^2), 2)     # MSE value

# Breeding Values for Bayesian Lasso
ETA = list(MRK = list(X = pregen, model = "BL"))
fmBLT = BGLR(y = ph, ETA = ETA, nIter = nIter, burnIn = burnIn, saveAt = "BLT_ ")
bvBL = pregen %*% fmBLT$ETA$MRK$b
which.max(bvBL)
bvBLmax = bvBL[which.max(bvBL)]
bvBLmax
MSEBL = round(mean((bvBL - QTL)^2), 2)        # MSE value
CorBL = round(cor(bvBL, QTL), 2)              # Correlation value
```

ضمیمه ۸. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک بیز A

```
# Bayes A
ETA$MRK$model = "BayesA"
fmBA = BGLR(y = yNa, ETA = ETA, nIter = nIter, burnIn = burnIn, saveAt = "BA_ ")
r_BA = cor(ph[whichNa], fmBA$yHat[whichNa])                      # Correlation value
MSEba = round(mean((ph[whichNa] - fmBA$yHat[whichNa])^2), 2)     # MSE value
```

```
# Breeding Values with Bayes A
ETA$MRK$model = "BayesA"
fmBAT = BGLR(y = ph, ETA = ETA, nIter = nIter, burnIn = burnIn, saveAt = "BAT_ ")
bvBA = pregen %*% fmBAT$ETA$MRK$b
which.max(bvBA)
bvBAmax = bvBA[which.max(bvBA)]
bvBAmax
MSEBA = round(mean((bvBA - QTL)^2), 2)        # MSE value
CorBA = round(cor(bvBA, QTL), 2)                    # Correlation value
```

ضمیمه ۹. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک بیز B

```
# Bayes B
ETA$MRK$model = "BayesB"
fmBB = BGLR(y = yNa, ETA = ETA, nIter = nIter, burnIn = burnIn, saveAt = "B_ ")
r_BB = cor(ph[whichNa], fmBB$yHat[whichNa])                    # Correlation value
MSEbb = round(mean((ph[whichNa] - fmBB$yHat[whichNa])^2), 2)    # MSE value

# Breeding Values with Bayes B
ETA$MRK$model = "BayesB"
fmBBT = BGLR(y = ph, ETA = ETA, nIter = nIter, burnIn = burnIn, saveAt = "BBT_ ")
bvBB = pregen %*% fmBBT$ETA$MRK$b
which.max(bvBB)
bvBBmax = bvBB[which.max(bvBB)]
bvBBmax
MSEBB = round(mean((bvBB - QTL)^2), 2)         # MSE value
CorBB = round(cor(bvBB, QTL), 2)                  # Correlation value
```

ضمیمه ۱۰. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک بیز C و بیز D

```
# Bayes C
ETA$MRK$model = "BayesC"
fmBC = BGLR(y = yNa, ETA = ETA, nIter = nIter, burnIn = burnIn, saveAt = "BC_ ")
r_BC = cor(ph[whichNa], fmBC$yHat[whichNa])                  # Correlation value
MSEbc = round(mean((ph[whichNa] - fmBC$yHat[whichNa])^2), 2)    # MSE value

# Breeding Values with Bayes C
ETA$MRK$model = "BayesC"
fmBCT = BGLR(y = ph, ETA = ETA, nIter = nIter, burnIn = burnIn, saveAt = "BCT_ ")
bvBC = pregen %*% fmBCT$ETA$MRK$b
which.max(bvBC)
bvBCmax = bvBC[which.max(bvBC)]
bvBCmax
MSEBC = round(mean((bvBC - QTL)^2), 2)                  # MSE value
CorBC = round(cor(bvBC, QTL), 2)                        # Correlation value

# Bayes D
ETA$MRK$model = "BayesB"
S0 = rgamma(n = length(ph), 1, 1)
fmBD = BGLR(y = yNa, response_type = "gaussian", ETA = ETA, S0 = S0,
                                          nIter = nIter, burnIn = burnIn, saveAt = "BD_ ")
r_BD = cor(ph[whichNa], fmBD$yHat[whichNa])                    # Correlation value
MSEbd = round(mean((ph[whichNa] - fmBD$yHat[whichNa])^2), 2)    # MSE value

# Breeding Values with Bayes D
ETA$MRK$model = "BayesB"
S0 = rgamma(n = length(ph), 1, 1)
fmBDT = BGLR(y = ph, response_type = "gaussian", ETA = ETA, S0 = S0,
```

```
                                    nIter = nIter, burnIn = burnIn, saveAt = "BDT_ ")
bvBD = pregen %*% fmBDT$ETA$MRK$b
which.max(bvBD)
bvBDmax = bvBD[which.max(bvBD)]; bvBDmax
MSEBD = round(mean((bvBD - QTL)^2), 2)          # MSE value
CorBD = round(cor(bvBD, QTL), 2)                # Correlation value
```

ضمیمه ۱۱. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک رگرسیون هسته‌ای روی نشانگرهای SNP

```
# Kernel Regression and Nadaraya-Watson Estimator
xstar = colMeans(pregen)
p = ncol(pregen)
n = length(ph)
SXstar = numeric(0)
for(i in 1:n){
SXstar[i] = t(pregen[i,] - xstar) %*% (pregen[i,] - xstar)
}
SXstar
# calculating the optimal bandwidth value
set.seed(1)
gridh = seq(0.1, 1, length = length(ph))
cv = numeric(0); Ghat = numeric(0)
for(i in 1:length(gridh)){
myS    = sum(exp(-SXstar / (2 * gridh[i]^2)))
Wxstar = exp(-SXstar / (2 * gridh[i]^2)) / myS
for(j in 1:n){
Ghat[j] = t(Wxstar)[-j] %*% ph[-j]
}
cv[i] = mean((ph - Ghat)^2)
}
mcv = which.min(cv)
mcv
h = gridh[mcv]
h
# Nadaraya-Watson Estimator for predicting the phenotypic values
h = 0.5
SXstar = numeric(0)
for(i in 1:n){
SXstar[i] = t(pregen[i,] - xstar) %*% (pregen[i,] - xstar)
}
myS    = sum(exp(-SXstar / (2 * h^2)))
Wxstar = exp(-SXstar / (2 * h^2)) / myS
Gxstar = t(Wxstar) %*% ph; Gxstar
dxstar = matrix(NA, ncol = p , nrow = n, byrow = T)
for(i in 1:n){
dxstar[i,] = ((pregen[i,] - xstar) * Wxstar[i]) / (h^2)
}
rxstar = colSums(dxstar)
ghatxstar = ph %*% (dxstar - (Wxstar %*% t(rxstar)))
Eyx = numeric(0)
for(i in 1:n){
Eyx[i] = Gxstar + (ghatxstar %*% (pregen[i,] - xstar))
}
MSE = mean((ph - Eyx)^2); MSE       #MSE value
cor(myy, Eyx)                       # Correlation value
```

ضمیمه ۱۲. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک رگرسیون فضای هیلبرت با هسته بازآفرین

```
# Reproducing kernel Hilbert space Regression
```

```
p = ncol(pregen)
D = (as.matrix(dist(pregen, method = "euclidean"))^2) / p  # DISTANCE MATRIX
D = D / mean(D)
K = exp(-2 * D)       ## guasian kernel
h = 0.5
# GENERATES TESTING SET
set.seed(1)
samplesize = 479
whichNa = sample(1:length(ph), size = samplesize, replace = FALSE)
yNa = ph
yNa[whichNa] = NA
### FIT MODEL
ETA = list(list(K = K,df0 = 5,S0 = 0.5, model = "RKHS"))
fm = BGLR(y = yNa, ETA = ETA, nIter = 5000, burnIn = 1000, saveAt = "RKHS_ ")
MSE = mean((ph[whichNa] - fm$yHat[whichNa])^2)
VARE = fm$varE
VARU = fm$ETA[[1]]$varU
DIC = fm$fit$DIC
pD = fm$fit$pD
round(MSE, 2); round(VARE, 2); round(VARU, 2); round(DIC, 2); round(pD)
corRkhsCV = round(cor(ph[whichNa], fm$yHat[whichNa]), 2)

# Breeding Values with RKHS
fmRKHST = BGLR(y = ph, ETA = ETA, nIter = 5000, burnIn = 1000, saveAt = "RKHST_ ")
bvRKHS = fmRKHST$yHat - fmRKHST$mu
which.max(bvRKHS)
bvRKHSmax = bvRKHS[which.max(bvRKHS)]
 bvRKHSmax
MSERKHS = round(mean((bvRKHS - QTL)^2), 2)    # MSE value
CorRKHS = round(cor(bvRKHS, QTL), 2)                # Correlation value
```

ضمیمه ۱۳. کُدنویسی در نرم‌افزار R جهت برآورد ارزش‌های ارثی با کمک رگرسیون ماشین بردار پشتیبان

```
# Support Vector Regression Method
library(kernlab)  # Package for Support Vector Regression Method
set.seed(100)
k = 5
folds = sample(1:k, length(ph), replace = T)
xtrain = pregen[folds!=k,];
ytrain = ph[folds!=k]
xtest = pregen[folds==k,]
ytest = ph[folds==k]
## model fitting
regm = ksvm(xtrain, ytrain, epsilon = 0.1, kernel = "rbfdot",
                              type = "eps-svr", kpar = "automatic", C = 1, cross = 5)
z = fitted(regm); head(z)
b(regm)
nSV(regm)
# prediction
per = predict(regm, xtest, ytest, type = "response")
mseSVRCV = round(mean((per - ytest)^2), 3)      # MSE value
corSVRCV = round(cor(per, ytest), 3)                # Correlation value

#  Breeding Values with Support Vector Regression Method
regmT = ksvm(pregen, ph, epsilon = 0.1, kernel = "rbfdot",
                              type = "eps-svr", kpar = "automatic", C = 1, cross = 5)
bvSVR = predict(regmT, pregen, ph, type = "response")
which.max(bvSVR)
bvSVRmax = bvSVR[which.max(bvSVR)]
bvSVRmax
b0 = b(regmT)
GEBVSVR = bvSVR - b0
which.max(GEBVSVR)
GEBVSVRmax = GEBVSVR[which.max(GEBVSVR)]
GEBVSVRmax
MSESVR = round(mean((GEBVSVR - QTL)^2), 2)          # MSE value
CorSVR = round(cor(GEBVSVR, QTL), 2)                  # Correlation value
```