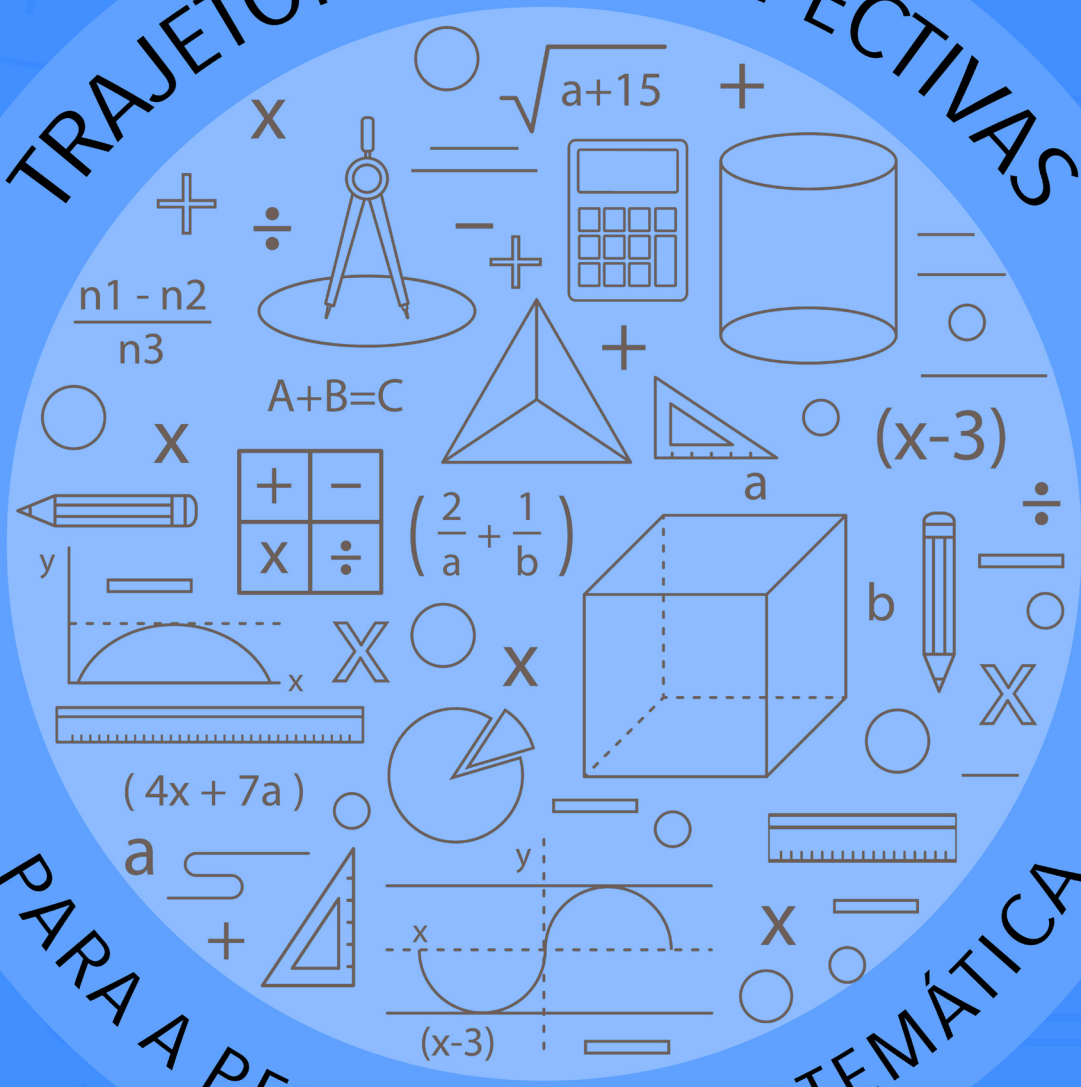
ANIELE DOMINGAS PIMENTEL SILVA
(Organizadora)

TRAJETÓRIAS E PERSPECTIVAS PARA A PESQUISA EM MATEMÁTICA

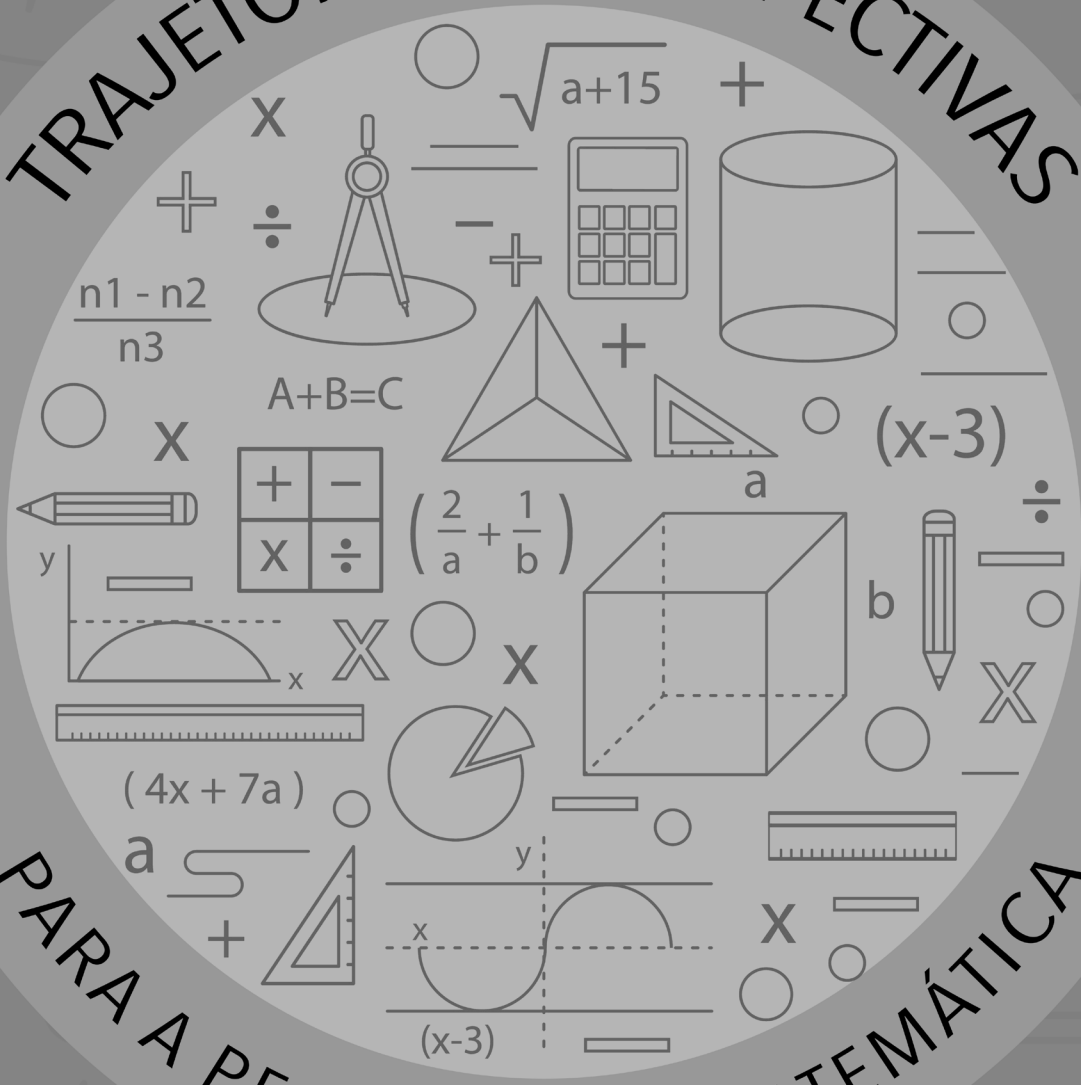


2

Atena Editora
Ano 2023

ANIELE DOMINGAS PIMENTEL SILVA
(Organizadora)

TRAJETÓRIAS E PERSPECTIVAS PARA A PESQUISA EM MATEMÁTICA



2

Atena Editora
Ano 2023

CIÊNCIAS EXATAS E DA TERRA E ENGENHARIAS

Editora chefe
Profª Drª Antonella Carvalho de Oliveira
Editora executiva
Natalia Oliveira
Assistente editorial
Flávia Roberta Barão
Bibliotecária
Janaina Ramos
Projeto gráfico
Bruno Oliveira
Camila Alves de Cremo
Luiza Alves Batista
Imagens da capa
iStock
Edição de arte
Luiza Alves Batista

O conteúdo dos artigos e seus dados em sua forma, correção e confiabilidade são de responsabilidade exclusiva dos autores, inclusive não representam necessariamente a posição oficial da Atena Editora. Permitido o *download* da obra e o compartilhamento desde que sejam atribuídos créditos aos autores, mas sem a possibilidade de alterá-la de nenhuma forma ou utilizá-la para fins comerciais.

Todos os manuscritos foram previamente submetidos à avaliação cega pelos pares, membros do Conselho Editorial desta Editora, tendo sido aprovados para a publicação com base em critérios de neutralidade e imparcialidade acadêmica.

A Atena Editora é comprometida em garantir a integridade editorial em todas as etapas do processo de publicação, evitando plágio, dados ou resultados fraudulentos e impedindo que interesses financeiros comprometam os padrões éticos da publicação. Situações suspeitas de má conduta científica serão investigadas sob o mais alto padrão de rigor acadêmico e ético.

CIÊNCIAS EXATAS E DA TERRA E ENGENHARIAS

Trajetórias e perspectivas para a pesquisa em matemática 2

Diagramação: Camila Alves de Cremo
Correção: Yaiddy Paola Martinez
Indexação: Amanda Kelly da Costa Veiga
Revisão: Os autores
Organizadora: Aniele Domingas Pimentel Silva

Dados Internacionais de Catalogação na Publicação (CIP)

T768 Trajetórias e perspectivas para a pesquisa em matemática 2 / Organizadora Aniele Domingas Pimentel Silva. – Ponta Grossa - PR: Atena, 202

Formato: PDF
Requisitos de sistema: Adobe Acrobat Reader
Modo de acesso: World Wide Web
Inclui bibliografia
ISBN 978-65-258-1050-8
DOI: https://doi.org/10.22533/at.ed.508231502

1. Matemática. I. Silva, Aniele Domingas Pimentel (Organizadora). II. Título.

CDD 510

Elaborado por Bibliotecária Janaina Ramos – CRB-8/9166

Atena Editora
Ponta Grossa – Paraná – Brasil
Telefone: +55 (42) 3323-5493
www.atenaeditora.com.br
contato@atenaeditora.com.br

DECLARAÇÃO DOS AUTORES

Os autores desta obra: 1. Atestam não possuir qualquer interesse comercial que constitua um conflito de interesses em relação ao artigo científico publicado; 2. Declaram que participaram ativamente da construção dos respectivos manuscritos, preferencialmente na: a) Concepção do estudo, e/ou aquisição de dados, e/ou análise e interpretação de dados; b) Elaboração do artigo ou revisão com vistas a tornar o material intelectualmente relevante; c) Aprovação final do manuscrito para submissão.; 3. Certificam que os artigos científicos publicados estão completamente isentos de dados e/ou resultados fraudulentos; 4. Confirmam a citação e a referência correta de todos os dados e de interpretações de dados de outras pesquisas; 5. Reconhecem terem informado todas as fontes de financiamento recebidas para a consecução da pesquisa; 6. Autorizam a edição da obra, que incluem os registros de ficha catalográfica, ISBN, DOI e demais indexadores, projeto visual e criação de capa, diagramação de miolo, assim como lançamento e divulgação da mesma conforme critérios da Atena Editora.

DECLARAÇÃO DA EDITORA

A Atena Editora declara, para os devidos fins de direito, que: 1. A presente publicação constitui apenas transferência temporária dos direitos autorais, direito sobre a publicação, inclusive não constitui responsabilidade solidária na criação dos manuscritos publicados, nos termos previstos na Lei sobre direitos autorais (Lei 9610/98), no art. 184 do Código Penal e no art. 927 do Código Civil; 2. Autoriza e incentiva os autores a assinarem contratos com repositórios institucionais, com fins exclusivos de divulgação da obra, desde que com o devido reconhecimento de autoria e edição e sem qualquer finalidade comercial; 3. Todos os e-book são *open access, desta forma* não os comercializa em seu site, sites parceiros, plataformas de *e-commerce,* ou qualquer outro meio virtual ou físico, portanto, está isenta de repasses de direitos autorais aos autores; 4. Todos os membros do conselho editorial são doutores e vinculados a instituições de ensino superior públicas, conforme recomendação da CAPES para obtenção do Qualis livro; 5. Não cede, comercializa ou autoriza a utilização dos nomes e e-mails dos autores, bem como nenhum outro dado dos mesmos, para qualquer finalidade que não o escopo da divulgação desta obra.

APRESENTAÇÃO

A coleção “Trajetórias e perspectivas para a pesquisa em matemática 2” tem como foco criar espaços de discussão científica através dos diversificados trabalhos que a compõem. A coletânea abordará trabalhos, pesquisas com relatos de experiências e a matemática no campo interdisciplinar.

O objetivo principal é divulgar algumas pesquisas desenvolvidas por várias instituições de ensino superior do país, cujo eixo central dos trabalhos estão relacionados a metodologias de ensino, tendências em educação matemática e formação de professores. Nesse sentido, observa-se o avanço de pesquisas no campo da educação matemática, visando buscar maneiras que possam tornar a matemática mais atrativa e significativa aos alunos.

Os diversos temas discutidos nesse volume mostram que o conhecimento acadêmico é fundamental, propõe dialogo e reflexão para todos aqueles que tem interesse em conhecer e/ou melhorar sua prática pedagógica e ter um material disponível que permita o contato com essas pesquisas é extremamente relevante.

Deste modo a obra “Trajetórias e perspectivas para a pesquisa em matemática 2” apresenta resultados de pesquisas que foram satisfatórias e que podem aguçar a curiosidade e inspirar os leitores, por isso a importância de espaços como este de divulgação científica.

Aniele Domingas Pimentel Silva

SUMÁRIO

SUMÁRIO

CAPÍTULO 10

MÉTODOS PARA MAPEAMENTO DE QTL ATRAVÉS DE MARCADORES TIPO SNP: UMA COMPARAÇÃO

Data de aceite: 01/02/2023

Lara Midena João

Daiane Aparecida Zuanetti

RESUMO: O mapeamento de regiões no genoma associadas a traços quantitativos (QTLs) através de marcadores genéticos do tipo SNP tem sido um dos problemas centrais em Genética e Biologia Molecular e vários métodos de detecção e identificação de QTLs tem sido propostos na literatura. Neste trabalho, três diferentes metodologias foram aplicadas e comparadas nos dados GAW17 quanto ao seus desempenhos em identificar corretamente SNPs relevantes e reguladores de um traço quantitativo. São elas: regressão linear simples com e sem a correção de Bonferroni no nível de significância, LASSO e SPLS. A fim de comparar o desempenho dessas metodologias, utilizamos a sensibilidade e a especificidade como métricas e notamos que o LASSO e a regressão linear simples com nível de significância de 5% apresentam os melhores resultados, uma vez que equilibram valores relativamente altos de sensibilidade e especificidade. O LASSO, por sua vez, também identifica SNPs influentes mais raros. Para realizar esse estudo, utilizamos algoritmos que já estão implementados em pacotes estatísticos, tais como R, e que estão disponíveis para utilização.

1 | INTRODUÇÃO

Um dos problemas centrais na Genética e Biologia Molecular é a detecção e identificação de regiões no genoma associadas a traços quantitativos (fenótipos) de seres vivos. Essas regiões genômicas são geralmente conhecidas como regiões de traços quantitativos (do inglês *quantitative trait loci*, QTLs) e suas posições e efeitos sobre o fenótipo de interesse são estimados através de marcadores genéticos, mais comumente do tipo SNP (do inglês *single nucleotide polymorphism*), dos indivíduos.

Para identificar a(s) região(ões) genômica(s) causadora(s) ou promotora(s) (os QTLs) do fenótipo de interesse, milhares ou milhões de SNPs são genotipados em amostras compostas de centenas ou milhares de indivíduos. Os genótipos dos

SNPs são, então, vistos como covariáveis que podem afetar o fenótipo, considerado como a variável resposta. O fenótipo, quando se trata de uma variável contínua, é geralmente modelado com uma função linear dos efeitos aditivos e de dominância do genótipo dos SNPs e/ou suas interações de segunda, terceira ou maior ordem.

Muitos métodos tem sido propostos e estudados para identificar e selecionar os SNPs mais associados ao fenótipo de interesse. Os métodos mais usados são baseados na estimação do modelo de regressão linear simples entre o genótipo de cada SNP e o fenótipo em estudo.

Os SNPs mais significativos são escolhidos via teste de hipóteses baseado na significância do efeito do SNP no modelo estimado. As vantagens dessa abordagem simples são baixo tempo de processamento computacional, facilidade de uso e de interpretação dos resultados. No entanto, ela geralmente apresenta baixo poder, não permite que os efeitos dos SNPs sobre o fenótipo sejam conjuntamente estimados e também não considera a estrutura de associação existente entre eles (Yazdani, 2014, Feng et al., 2012, Oliveira, 2015).

Métodos alternativos aos modelos de regressão linear simples são as metodologias que permitem analisar os SNPs conjuntamente, alguns deles identificados como métodos de aprendizado de máquina. Entre eles se destacam: florestas aleatórias (Breiman, 2001, Oliveira, 2015), algoritmos genéticos (Goldberg, 1989, Oliveira, 2015), LASSO (do inglês *least absolute shrinkage and selection operator*, Tibshirani, 1996) que pode ser visto como um modelo de regressão Bayesiano com função de verossimilhança normal e distribuição *a priori* exponencial dupla para cada coeficiente (Park e Casella, 2008), e métodos de regressão por mínimos quadrados parciais – SPLS (do inglês *sparse partial least squares*, Chun, 2010). Todos esses métodos nos permitem trabalhar com dados cujo número de covariáveis é muito superior ao número de indivíduos da amostra e nos quais existe covariáveis altamente correlacionadas. Essas características estão presentes nos dados de mapeamento de QTLs através de marcadores SNPs e precisam ser consideradas.

Nesse trabalho, portanto, estudamos e aplicamos três diferentes metodologias nos dados GAW (Genetic Analysis Workshop) 17 (Almasy et al., 2011) para detectar e identificar SNPs associados ao fenótipo de interesse, principalmente SNPs com variantes raras e baixa frequência do alelo menor. Também discutimos as vantagens e desvantagens de cada método e elencamos os algoritmos que já estavam implementados em pacotes estatísticos, tais como R, e disponíveis para utilização.

O texto está organizado como segue. Na Seção 2 descrevemos as três metodologias que são estudadas e aplicadas para a identificação de SNPs relevantes e como o desempenho delas é analisado e comparado. A Seção 3 apresenta o conjunto de dados GAW17 que é estudado e os resultados obtidos. Finalmente, a Seção 4 traz conclusões e discussões.

2 | METODOLOGIAS PARA SELEÇÃO DE SNPS INFLUENTES

Nessa seção, descrevemos os métodos que foram utilizados para detectar e identificar SNPs associados ao fenótipo de interesse Y. São eles: o valor-p do teste de significância associado ao efeito de cada SNP em um modelo de regressão linear simples, o LASSO e o SPLS. Além disso, a fim de comparar qual desses métodos é o melhor para selecionar os SNPs influentes, empregamos as seguintes métricas: especificidade (*E*) e sensibilidade (*S*) (Lee et al., 2011) que são descritas na Seção 2.4.

2.1 Regressão linear simples

A fim de identificar e selecionar os SNPs mais associados ao fenótipo de interesse, um método muito utilizado é baseado na estimação do modelo de regressão linear simples entre o genótipo de cada SNP e o fenótipo em estudo. Nesse modelo o genótipo de cada SNP é tido como variável preditora e o fenótipo é considerado como variável resposta.

Seja $Y = (Y_1, Y_2, ..., Y_n)$ um traço quantitativo a ser observado em n indivíduos. O fenótipo Y_i do i-ésimo indivíduo pode ser modelado pelo seguinte modelo de regressão linear simples:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{1}$$

em que β_0 é o valor esperado do fenótipo quando o genótipo do SNP vale zero, β_1 é o efeito aditivo que representa o acréscimo no valor esperado do fenótipo quando o genótipo do SNP vale 1, ou seja, quando o indivíduo é homozigoto dominante, x_i é o genótipo do SNP, considerado na específica análise, do i-ésimo indivíduo codificado como –1,0 ou 1 para *aa*, *Aa* ou *AA*, respectivamente, $i = 1,...,n$, $\varepsilon_i \sim$ Normal$(0,\sigma^2)$ é o erro aleatório e ε_i e $\varepsilon_{i'}$ são supostamente independentes para $i \neq i'$.

Os SNPs significativos são escolhidos via teste de hipóteses baseado na significância de cada SNP no modelo estimado. Esse teste de hipóteses tem como objetivo, além de verificar se a média do fenótipo dos indivíduos varia linearmente em função do genótipo do SNP, identificar quais marcadores genéticos são influentes. As hipóteses associadas ao teste são:

> H_0: $\beta_1 = 0$ (a média do fenótipo não varia linearmente em função do genótipo do SNP, ou seja, o efeito do SNP não é significativo) contra
>
> H_1: $\beta_1 \neq 0$ (a média do fenótipo varia linearmente em função do genótipo do SNP, ou seja, o efeito do SNP é significativo).

Para testar essas hipóteses, utilizamos a estatística teste $F = \frac{QMReg}{QMRes}$, em que $QMReg = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$ e $QMRes = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$, sendo $\hat{y}_i$ o valor do fenótipo do i-ésimo indivíduo predito e $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$. Sob H_0, essa estatística segue uma distribuição F de Snedecor com 1 grau de liberdade no numerador e $n–2$ graus de liberdade no denominador.

Para rejeitarmos ou não a hipótese nula, utilizamos o valor-p definido como a probabilidade de obtermos um valor da estatística do teste igual ou mais desfavorável à hipótese nula (H_0) que o valor observado quando H_0 é verdadeira. Mais detalhes sobre esse teste de hipóteses são encontrados em Morettin e Bussab (2017).

Um SNP é considerado significativo, ou seja, ele afeta o fenótipo, quando $\beta_1 \neq 0$. Fixado um nível de significância α, rejeitamos H_0 se o valor-p é menor que α e não rejeitamos H_0, caso contrário.

Nesse estudo, cada teste de hipóteses é realizado para diferentes níveis de significância (α), são eles: 5% e 10%, ambos com e sem a correção de Bonferroni para múltiplos testes parciais. Essa correção tem como finalidade reduzir a probabilidade de concluir simultânea e erroneamente que muitos SNPs são significativos e, consequentemente, identificar muitos SNPs falso positivos. O método é executado do seguinte modo: para que o nível de significância na decisão global (considerando conjuntamente todos os testes realizados, um para cada SNP) seja α, o nível de significância de cada teste deve ser o resultado da divisão de α pelo número de testes realizados (Abdi, 2007). Quando muitos testes devem ser feitos, como nesse caso em que testamos o efeito de milhares de SNPs, o nível de significância de cada teste é tão pequeno que o seu poder também cai drasticamente e, assim, podemos não selecionar muitos SNPs verdadeiramente significativos.

2.2 LASSO

Com o propósito de identificar e selecionar os SNPs que influenciam o fenótipo de interesse, o LASSO (do inglês *least absolute shrinkage and selection operator*) também tem sido uma das metodologias mais utilizadas. Ele se trata de uma regressão linear múltipla que seleciona as variáveis preditoras utilizando uma restrição através da qual a soma do valor absoluto dos coeficientes da regressão seja menor que uma constante fixada. Desse modo, esse modelo permite analisar os SNPs conjuntamente para cada cromossomo.

Nesse modelo, o genótipo de cada SNP é tido como variável preditora e o fenótipo é considerado como variável resposta. Seja $Y = (Y_1, Y_2, ..., Y_n)$ um traço quantitativo a ser observado em n indivíduos e $x_i = (x_{1i}, ..., x_{pi})$ o genótipo de p SNPs para o i-ésimo indivíduo. O fenótipo Y_i do i-ésimo indivíduo pode ser modelado pelo seguinte modelo de regressão múltipla:

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} + \epsilon_i, \qquad (2)$$

em que β_0 é o valor esperado do fenótipo quando o genótipo de todos os SNPs valem zero, β_k é o efeito aditivo do k-ésimo SNP no valor esperado do fenótipo, x_{ki} é o genótipo do k-ésimo SNP do i-ésimo indivíduo codificado como –1,0 ou 1 para *aa*, *Aa* ou *AA*, respectivamente, $k = 1,...,p$ e $i = 1,...,n$, ε_i é o erro aleatório do i-ésimo indivíduo e ε_i e $\varepsilon_{i'}$ são supostamente independentes para $i \neq i'$.

O LASSO é um método que seleciona as variáveis preditoras mais influentes através da adição de uma restrição na fórmula dos mínimos quadrados. Essa restrição é da forma: $\sum_{k=1}^{p} |\beta_k| \leq c$, em que $c = c(\lambda)$. Assim, por conta dessa limitação, temos que algumas das estimativas desses coeficientes (β_ks) serão aproximadamente 0. Desse modo, SNPs associados a um coeficiente com estimativa diferente de zero são selecionados como relevantes e SNPs com estimativa zero são descartados da análise.

Na sua forma lagrangiana, a metodologia LASSO busca

$$(\hat{\beta}_0, \hat{\beta}) = \arg \min_{\beta_0, \beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{p} \beta_k x_{ki} \right)^2 + \lambda \sum_{k=1}^{p} |\beta_k|, \tag{3}$$

em que $\lambda \geq 0$ é um parâmetro de regularização. Além disso, quanto maior o valor de λ, mais restrita é a limitação e, portanto, menor a quantidade de SNPs significativos. Consequentemente, menos complexo é o modelo. Ademais, se $\sum_{k=1}^{p} |\tilde{\beta}_k| \leq c(\lambda)$, em que $\tilde{\beta}_k$s são as estimativas do método de mínimos quadrados, então as estimativas dos coeficientes do LASSO também serão os $\tilde{\beta}_k$s, pois não haverá restrição para as estimativas dos parâmetros (Rodrigues, 2018).

De acordo com Rodrigues (2018) temos que, se $\lambda = 0$, então o estimador obtido pelo modelo LASSO é idêntico ao obtido pelo método de mínimos quadrados. Ainda, para cada valor de λ, obtemos diferentes valores de $\hat{\beta}$.

Desse modo, a fim de obtermos o melhor modelo, é necessário escolher um valor para λ que não assuma o valor 0 nem valores muito grandes, uma vez que, no primeiro caso, todas as covariáveis seriam selecionadas, e, no segundo caso, provavelmente não seriam selecionadas covariáveis e o modelo possuiria apenas o intercepto β_0 (Rodrigues, 2018). Assim, o valor de λ escolhido, para cada cromossomo, para compor o modelo é aquele que possui o menor erro de predição dentre todos os valores testados para o λ.

Esses valores do parâmetro de regularização foram definidos através da função "cv.glmnet" (do pacote "glmnet"do software R) que realiza a validação cruzada *k*-fold para obtê-los. Essa validação cruzada é um método no qual a amostra é dividida em *l* partes (subamostras) mutuamente exclusivas e aproximadamente de mesmo tamanho. O modelo é estimado com *l* – 1 partes e testado na única parte que não é utilizada para estimar o modelo. Para cada valor de λ, são propostos *l* modelos e, para cada um, é calculado o erro de predição (*EP*) como

$$EP = \sum_{i} \left(y_i - \hat{y}_i \right)^2, \tag{4}$$

em que y_i e $\hat{y}_i$ são, respectivamente, o valor observado e o valor predito, pelo modelo de regressão múltipla, do fenótipo dos indivíduos que fazem parte da subamostra não utilizada na estimação.

Esse procedimento é repetido *l* vezes, alterando, a cada vez, as *l* – 1 subamostras

sobre as quais o modelo é estimado e, consequentemente, a subamostra na qual ele é testado. Portanto, para cada valor de λ, temos *l* erros de predição e calculamos a média desses *EP*s. O valor de *l* utilizado para análise foi 10, que é o valor padrão da função no R e o λ escolhido, para o específico cromossomo, é aquele que apresenta a menor média de *EP*. Para mais detalhes, ver Ióca e Zuanetti (2021).

2.3 SPLS

Outra metodologia que tem sido muito utilizada para identificar e selecionar SNPs que possuem efeito significativo para o fenótipo de interesse é o SPLS (do inglês *sparse partial least squares regression*). Assim como o LASSO, o SPLS se trata de uma regressão linear múltipla que realiza a seleção das variáveis através de uma restrição na soma do valor absoluto dos coeficientes.

Essa metodologia consiste na redução de dimensão dos dados através da inclusão de variáveis latentes (combinações lineares das variáveis originais) e seleção de quais dessas variáveis são relevantes no modelo. Logo, esse método realiza a colapsagem de SNPs que possuem alta correlação em variáveis ocultas e verifica quais delas são realmente significativas para explicar a variabilidade do fenótipo em questão, de modo a analisar conjuntamente os SNPs para cada cromossomo.

Seja $Y = (Y_1, Y_2, ..., Y_n)$ um traço quantitativo a ser observado em n indivíduos. Supondo que p SNPs são considerados, sendo $W_{p\times M}$ a matriz de carga fatorial (matriz com os coeficientes do genótipo de cada SNP em um cromossomo) de M vetores $(w_1, w_2, ..., w_M)$ de tamanho p e sendo $X_{n\times p}$ a matriz que, para um cromossomo, contém o genótipo dos SNPs do i-ésimo indivíduo, $i = 1,...,n$, codificados em –1,0 ou 1 para *aa*, *Aa* ou *AA*, respectivamente. Dessa maneira, XW contém, para cada indivíduo, grande parte das informações originais dos SNPs colapsadas em $M < min(n,p)$ combinações lineares de seus genótipos. Cada uma dessas M combinações lineares do genótipo dos SNPs são denominadas componentes latentes, cujos valores não são observados no conjunto de dados, mas preditos durante o processo de ajuste do modelo.

De maneira similar ao LASSO, o SPLS impõe uma restrição em cada w_m para que haja uma solução esparsa, ou seja, para que muitos SNPs sejam identificados como não significativos e poucos tenham cargas fatoriais diferente de zero dentro de cada componente latente. Logo, a solução para cada estimativa de w_m é dada por:

$$\hat{\boldsymbol{w}}_m = \max_{\boldsymbol{w}_m} \boldsymbol{w}_m^t \boldsymbol{Q} \boldsymbol{w}_m \tag{5}$$

em que $Q = X^tYY^tX$ e sujeita a $w^t_m w_m = 1$ e $w^t_m S_{XX} \hat{w}_h = 0$, $h = 1,...,m-1$ e $m = 1,...,M-1$, em que S_{XX} representa a matriz de covariância amostral dos SNPs. Desse modo, a solução $\hat{w}_m$ traz como resultado uma regressão de mínimos quadrados parcial, em que diferentes componentes latentes explicam diferentes partes da variabilidade do fenótipo.

Apesar disso, uma solução esparsa não é alcançada e w_{mk} geralmente é diferente

de zero para muitos SNPs. Chung e Keles (2010) propõe, então, uma formulação impondo uma penalização em um vetor de cargas substituto (*c*) e não no vetor de cargas fatoriais original (*w*), enquanto mantém *w* e *c* próximos um do outro. A formulação é:

$$\tilde{\boldsymbol{w}}_m, \tilde{\boldsymbol{c}}_m = \arg\min_{\boldsymbol{w}_m, \boldsymbol{c}_m} \left\{ -d\boldsymbol{w}_m^t \boldsymbol{Q} \boldsymbol{w}_m + (1-d)(\boldsymbol{c}_m - \boldsymbol{w}_m)^t \boldsymbol{Q}(\boldsymbol{c}_m - \boldsymbol{w}_m) + \lambda_1 \sum_{k=1}^{p} |c_{mk}| + \lambda_2 \sum_{k=1}^{p} c_{mk}^2 \right\}$$

para $0<d\leq\frac{1}{2}$, tal que $w^t_m w_m=1$ e $w^t_m SXX\hat{w}_h=0$ para $h=1,\ldots, m$; $m=1,\ldots, M-1$, e em que $Q=X^tYY^tX$.

Enquanto a penalidade λ_1 promove a esparsidade, a penalidade λ_2 evita que a matriz *Q* seja singular (não inversível). Para *Y* univariado, como neste estudo, e assumindo $\lambda_2=\infty$, a solução não depende de *d* e resulta no limite

$$\tilde{\boldsymbol{w}}_m = \left(|\hat{\boldsymbol{w}}_m| - \eta \max_{1\leq k\leq p} |\hat{w}_{mk}| \right) \mathbb{I}_{(|\hat{\boldsymbol{w}}_m| \geq \eta \max_{1\leq k\leq p} |\hat{w}_{mk}|)} sign(\hat{\boldsymbol{w}}_m), \qquad (6)$$

em que $0 \leq \eta \leq 1,\ldots,\lambda_1$ e $\hat{w}_m$ é a solução da regressão múltipla parcial encontrada na Equação (5) e $\tilde{w}_m$ é a solução parcial e esparsa da Equação (6), como mostra Feng et al. (2012).

Logo, para cada cromossomo são encontrados valores de *M* e η que minimizam o erro de predição (*EP*). Esses valores foram definidos via validação cruzada através da função "cv.spls"(do pacote "spls"do software R) de modo que a amostra foi dividida em *l* = 3 partes mutuamente exclusivas e aproximadamente de mesmo tamanho. O modelo é estimado com *l* – 1 partes e testado na única parte que não é utilizada para estimar o modelo. Para cada valor de *M* e η são ajustados *l* modelos e, para cada um, é calculado o erro de predição (*EP*) da seguinte maneira

$$EP = \sum_i \left(y_i - \hat{y}_i \right)^2,$$

em que y_i e $\hat{y}_i$ são, respectivamente, o valor observado e o valor predito, pelo modelo de regressão múltipla, do fenótipo dos indivíduos que fazem parte da subamostra não utilizada na estimação.

Esse procedimento é repetido *l* vezes, alterando, em cada momento, as *l*–1 subamostras em que o modelo é estimado e, consequentemente, a subamostra na qual ele é testado. Portanto, para cada valor de *M* e η, temos *l* erros de predição e calculamos a média desses *EP*s. Os valores de *M* e η escolhidos, para o específico cromossomo, são aqueles que apresentam a menor média de *EP*. Os SNPs relevantes são os que apresentam cargas fatoriais diferentes de zero nos *M* componentes latentes que explicam a variabilidade do fenótipo. Para mais detalhes, ver Feng et al. (2012).

2.4 Medidas de desempenho

A fim de medir e comparar o desempenho das diferentes metodologias testadas na seleção de SNPs influentes, adotamos duas métricas denominadas especificidade e sensibilidade (Lee et al., 2011).

A sensibilidade é calculada como a proporção de estimativas de β diferentes de 0 dentre os verdadeiros elementos β que são não nulos. Já a especificidade é a razão da quantidade de estimativas de β iguais a 0 dentre os verdadeiros elementos β que são nulos. De maneira mais direta, o numerador da especificidade pode ser calculado como a diferença entre a quantidade de SNPs que não são verdadeiramente significativos e a quantidade de SNPs que foram erroneamente identificados como significativos e o denominador dessa fração é a quantidade de SNPs que realmente não são significativos.

Uma seleção de variáveis ideal ocorre quando tanto a sensibilidade, quanto a especificidade são iguais a um. Todavia, geralmente há uma compensação entre essas medidas, ou seja, quando a especificidade tende a 1, a sensibilidade tende a 0, e vice-versa.

3 | ANALISANDO O GAW 17

O banco de dados utilizado nesse estudo é denominado *Genetic Analysis Workshop 17* (GAW 17) e é formado por dados simulados, a partir de características reais da população, de 697 indivíduos sem parentesco, sendo eles 327 homens e 370 mulheres.

Como base para a simulação de uma doença comum e complexa, de fenótipos quantitativos e dos fatores de risco (os marcadores SNPs) foram utilizados os dados reais contidos no *1000 Genomes Project,* que consideram variações genéticas de vários grupos de populações humanas, sendo eles: Europa, Leste Asiático, do sul da Asia,´ Africa Ocidental e de índios´ americanos.

Um fenótipo binário que indica presença ou não de uma doença e três fenótipos quantitativos contínuos: Q_1,Q_2 e Q_3 foram simulados. Para mais detalhes, ver lóca e Zuanetti (2021) e Almasy et al. (2011).

Originalmente, o banco de dados nos fornecia as informações dos genótipos dos marcadores SNPs através de 16 possíveis pares de bases nitrogenadas, sendo eles: A/A, T/T, C/C, G/G, A/T, A/C, A/G, T/A, T/C, T/G, C/A, C/T, C/G, G/A, G/T, G/C. Todavia, a fim de possibilitarmos esse estudo, classificamos essas informações do seguinte modo:

- C/C ou G/G: homozigoto dominante codificado como 1;
- A/T, A/C, A/G, T/A, T/C, T/G, C/A, C/T, C/G, G/A, G/T, G/C: heterozigoto codificado como 0; e
- A/A ou T/T: homozigoto recessivo codificado como -1.

Nesse estudo, analisamos o fenótipo Q_1 que originalmente é impactado por 39

SNPs em 9 genes. Nosso objetivo é verificar o desempenho das metodologias estudadas em identificar e selecionar os 39 SNPs relevantes entre os 24487 SNPs disponíveis para análise. O boxplot do fenótipo estudado Q_1 está disponível na Figura 1.

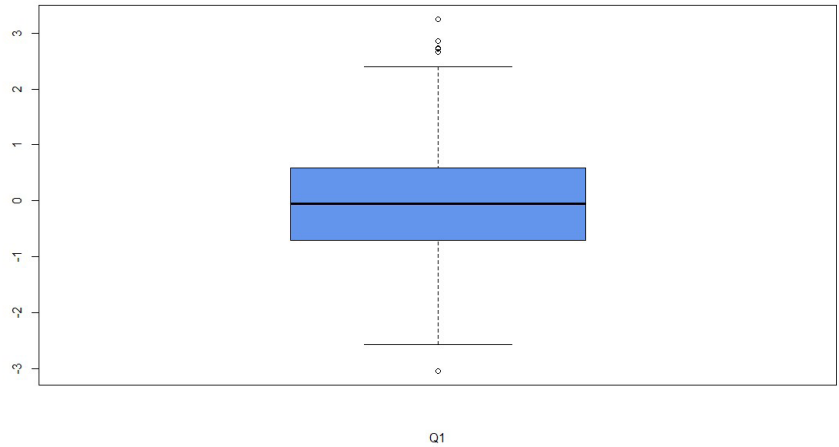


Figura 1: Boxplot do fenótipo Q_1.

Notamos, através da Figura 1 que o fenótipo Q_1 é uma variável contínua que assume valores no intervalo (–3.0,3.5) e possui 5 outliers, sendo que 4 deles ultrapassam o limite superior e 1 ultrapassa o limite inferior. Além disso, Q_1 possui comportamento simétrico e 50% dos valores observados estão entre –1 e 1.

Ademais, de acordo com a simulação realizada, os cromossomos que possuem SNPs significativos são os cromossomos 1, 4, 5, 6, 13, 14 e 19. Esses SNPs que realmente influenciam o fenótipo Q_1 são listados a seguir de acordo com o cromossomo em que eles se encontram (de modo que a notação CaSb identifica o SNP b presente no cromossomo a):

- Cromossomo 1: C1S6533, C1S6537, C1S6540, C1S6542, C1S6561, C1S3181 e C1S3182;
- Cromossomo 4: C4S1861, C4S1873, C4S1874, C4S1877, C4S1878, C4S1879, C4S1884, C4S1887, C4S1889, C4S1890 e C4S4935;
- Cromossomo 5: C5S5133 e C5S5156;
- Cromossomo 6: C6S2981;
- Cromossomo 13: C13S320, C13S399, C13S431, C13S479, C13S505, C13S514, C13S522, C13S523, C13S524, C13S547 e C13S567;
- Cromossomo 14: C14S1718, C14S1729, C14S1734 e C14S1736; e
- Cromossomo 19: C19S4799, C19S4815 e C19S4831.

Vale ressaltar que apenas 7 entre os 39 SNPs verdadeiramente relevantes possuem genótipos diferentes em mais que 1% do total de indivíduos, ou seja, grande parte dos SNPs são variantes raras e dificilmente identificados por metodologias de seleção de variáveis tradicionais.

3.1 Resultados

Nessa seção aplicamos os três métodos discutidos para identificar e selecionar os SNPs relevantes no fenótipo simulado Q_1, disponível nos dados GAW17. Mostramos os resultados obtidos e comparamos as metodologias em termos de sensibilidade e especificidade na seleção dos SNPs.

3.2 Regressão linear simples

Na primeira metodologia, foram realizados testes de hipóteses para verificar a significância do coeficiente de regressão associado ao genótipo de cada SNP em um modelo de regressão linear simples. Consideramos os níveis de significância α de 0.10, 0.05 e suas correções de Bonferroni. Os resultados obtidos nesses testes para cada cromossomo autossômico e para cada α são expostos a seguir e comparados com os 39 SNPs que são realmente significativos para Q_1 de acordo com Almasy et al. (2011).

Utilizando o nível de significância de 10%, o modelo identificou que todos os cromossomos autossômicos possuem SNPs significativos, detectando, no total, 3575 SNPs como significativos. Além disso, dos 39 SNPs que realmente tem efeito significativo sobre Q_1, apenas 16 foram identificados pelo modelo. Desse modo, 3559 SNPs foram identificados incorretamente como significativos pela metodologia utilizada. Para o nível de significância de 10%, a sensibilidade (S) e a especificidade (E) valem, respectivamente:

$$S = \frac{16}{39} = 0.410 \; e \; E = \frac{20889}{24448} = 0.854.$$

Os valores acima possuem a seguinte interpretação: 41% dos SNPs que são realmente significativos são identificados pelo modelo e 85.4% dos SNPs que não possuem efeito importante sobre o fenótipo não são identificados pelo modelo.

Aplicando agora o nível de significância de 5%, o modelo identificou que todos os cromossomos autossômicos possuem SNPs significativos, em um total de 2177 SNPs relevantes. Comparando esse valor com a quantidade de SNPs identificados com α = 10%, observamos uma redução na quantidade de SNPs selecionados, uma vez que, quanto menor o α, mais restritiva é a condição para considerar a hipótese H_1 como sendo verdadeira, e, sendo assim, menor a quantidade de SNPs falsos positivos. Dos 39 SNPs significativos, apenas 15 são identificados pelo modelo, ou seja, há 24 SNPs que são significativos e não foram identificados.

A sensibilidade e a especificidade ao nível de significância de 5% valem, respectivamente:

$$S = \frac{15}{39} = 0.385 \; e \; E = \frac{22286}{24448} = 0.912.$$

Ao fazer uso do nível de significância de 10% com correção de Bonferroni, foi apontado que os cromossomos 1, 2, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 19, 20 e 22 possuem, ao todo, 42 SNPs que afetam o fenótipo Q_1. Entre os 42 SNPs selecionados, apenas 4 SNPs no cromossomo 13 foram corretamente identificados como significativos. São eles: C13S431, C13S522, C13S523 e C13S524. Os SNPs identificados como significativos pelo modelo foram: C1S1683, C1S11539, C2S2355, C7S1598 , C7S1632, C7S4110, C8S2899, C9S4121, C10S103, C12S703, C12S704, C12S706, C12S707, C12S708, C12S711, C12S718, C12S719, C12S792, C12S831, C12S971, C12S2028, C12S2798, C13S431, C13S522, C13S523, C13S524, C14S1137, C15S1580, C16S3109, C17S3017, C17S4607, C19S1163, C19S3512, C19S3829, C19S5085, C19S5879, C20S667, C22S116, C22S1507, C22S1508, C22S1540 e C22S1901.

Além disso, ao calcularmos a sensibilidade e a especificidade ao nível de significância de 10% com correção de Bonferroni, temos que essas são, respectivamente, iguais a 0.102 e 0.998. Portanto, 10.2% dos SNPs que são realmente importantes para Q_1 são identificados pelo modelo e 99.8% dos SNPs que de fato não têm efeito significativo sobre o fenótipo não são identificados pelo modelo de regressão linear simples.

Ao utilizar o nível de significância de 5% com correção de Bonferroni, os cromossomos 1, 2, 7, 8, 12, 13, 14, 15, 17, 19, 20 e 22 foram caracterizados como detentores de SNPs significativos, possuindo, ao todo, 27 SNPs significativos. Entre os 27 SNPs selecionados, apenas 2 SNPs no cromossomo 13 foram corretamente identificados como significativos, são eles: C13S522 e C13S523. A seguir estão listados quais SNPs o modelo identificou como significativos: C1S11539, C2S2355, C7S1598, C7S4110, C8S2899, C12S704, C12S707, C12S711, C12S719, C12S831, C12S971, C12S2028, C12S2798, C13S522, C13S523, C14S1137, C15S1580, C17S3017, C17S4607, C19S3512, C19S3829, C19S5085, C19S5879, C20S667, C22S1507, C22S1508 e C22S1901.

Ao calcular a sensibilidade e especificidade ao nível de significância de 5% corrigido por Bonferroni, a fim de comparar qual seria o melhor nível de significância para o modelo de regressão linear simples, temos que a proporção de estimativas de β diferentes de 0 dentre os SNPs que são realmente significativos e a proporção de estimativas de β iguais a zero dentre os SNPs que na realidade não são significativos são, respectivamente: 0.051 e 0.999.

Vale destacar que os 4 SNPs verdadeiro positivos identificados praticamente em todos os níveis de significância são os SNPs que simultaneamente apresentam *MAF* (frequência do menor alelo) maior que 1% e coeficientes de regressão maior que 0.60 na simulação do Q_1. Enquanto que SNPs com baixíssima frequência do alelo menor (maior parte dos SNPs realmente influentes) ou SNPs com efeito mais baixo no fenótipo não foram selecionados por essas metodologias. SNPs com coeficientes maiores que 1 na simulação,

tais como: C4S4935, C6S2981, C4S1889, e C4S1877, mas com *MAF* < 0.2% não foram identificados por essa metodologia.

Apesar de ser falso positivo, o SNP que foi identificado no cromossomo 14 pelos 4 níveis de significância está próximo a 4 SNPs relevantes, assim como os 4 SNPs identificados no cromossomo 19 estão próximos a 3 SNPs influentes.

3.3 LASSO

Os resultados obtidos através do LASSO para cada cromossomo são apresentados a seguir e comparados com os SNPs que realmente influenciam Q_1 segundo Almasy et al. (2011). A sensibilidade e especificidade atingidas por esse método também são expostas a seguir.

O modelo identificou que todos os cromossomos possuem SNPs significativos, detectando, ao todo, 799 SNPs como significativos. Ademais, dos 39 SNPs que são realmente significativos para Q_1, apenas 9 foram identificados pelo modelo, são eles: C4S1877, C4S1884, C4S1889, C6S2981, C13S320, C13S431, C13S522, C13S523 e C14S1734. Desse modo, 790 SNPs foram erroneamente identificados como relevantes pela metodologia utilizada. A sensibilidade (*S*) e a especificidade (*E*) valem, respectivamente:

$$S = \frac{9}{39} = 0.231 \; e \; E = \frac{23658}{24448} = 0.968. \tag{7}$$

Os valores observados acima possuem a seguinte interpretação: 23.1% dos SNPs que são realmente significativos são identificados pelo modelo e 96.8% dos SNPs que não são relevantes para o fenótipo não foram identificados pelo modelo. Notamos que o LASSO identificou 3 SNPs verdadeiros do cromossomo 13 que também foram identificados pelo valor-p do teste de hipóteses do modelo de regressão simples e outros SNPs que, apesar de apresentarem *MAF* bem baixo, apresentam grande efeito sobre o fenótipo, tais como C6S2981 e C4S1889. Outros 4 SNPs com pequeno *MAF* e razoável efeito sobre o fenótipo também foram selecionados.

3.4 SPLS

Para conseguir determinar os valores ótimos de *M* e η para cada cromossomo usando a validação cruzada no SPLS, foi necessário excluir da base e não analisar SNPs que possuíam menos que 5 indivíduos com variação nos cromossomos 2, 4, 6, 7, 14, 16, 17, 19, 20, 21 e 22; 6 indivíduos com variação nos cromossomos 1, 3, 5, 9, 10, 11, 13 e 18; 7 indivíduos com variação nos cromossomos 12 e 15; e 8 indivíduos com variação no cromossomo 8. Essa remoção foi necessária pois, ao dividir a base original em subamostras, esses SNPs muito raros não apresentavam variabilidade em alguma ou algumas das subamostras e o algoritmo não funcionava. Ademais, na validação cruzada o valor de *M* variava de 1 a 20, porém para todos os cromossomos foi selecionado *M* igual a 1 como o valor ótimo de componentes latentes.

A seguir, estão listados por cromossomo a quantidade de SNPs que o cromossomo possuía, a quantia que foi filtrada (devido a quantidade mínima de variações necessária) e o número de SNPs que foram identificados como significativos, respectivamente: Cromossomo 1: 2237, 738, 252; cromossomo 2: 1599, 595, 7; cromossomo 3: 1211, 349, 1; cromossomo 4: 944, 304, 8; cromossomo 5: 1074, 361, 47; cromossomo 6: 1425, 550, 45; cromossomo 7: 1063, 417, 4; cromossomo 8: 982, 333, 17; cromossomo 9: 1166, 517, 34; cromossomo 10: 1396, 549, 423; cromossomo 11: 2102, 959, 3; cromossomo 12: 1435, 526, 102; cromossomo 13: 425, 135, 2; cromossomo 14: 795, 262, 1; cromossomo 15: 933, 268, 1; cromossomo 16: 844, 308, 14; cromossomo 17: 1223, 470, 2; cromossomo 18: 634, 204, 4; cromossomo 19: 1649, 792, 164; cromossomo 20: 591, 239, 1; cromossomo 21: 251, 86, 21 e cromossomo 22: 508, 175, 10.

Vale destacar que, apesar de excluirmos apenas SNPs muito raros, 9137 (37.3% do total) não foram considerados por essa metodologia.

O modelo identificou que todos os cromossomos possuem SNPs significativos e, apesar de descartar uma quantidade grande de SNPs raros, detectou, ao todo, 1163 SNPs como significativos. Dos 39 SNPs que realmente possuem efeito sobre o fenótipo Q_1, apenas 4 foram identificados pelo modelo, são eles: C1S6533, C4S1884, C13S522 e C13S523, sendo que os dois SNPs do cromossomo 13 também foram identificados pelas duas metodologias anteriores. Logo, 1159 SNPs foram incorretamente identificados como significativos pela metodologia implementada. A sensibilidade (*S*) e a especificidade (*E*) valem, respectivamente:

$$S = \frac{4}{39} = 0.102 \; e \; E = \frac{23289}{24448} = 0.952. \tag{8}$$

Assim, 10.2% dos SNPs que são realmente significativos são identificados pelo modelo e 95.2% dos SNPs que não são relevantes para o fenótipo não foram identificados pelo modelo.

Vale destacar que o SPLS apenas identificou os SNPs verdadeiro positivos que apresentam *MAF* > 1% e moderado efeito sobre o fenótipo e selecionou uma grande quantidade de falso positivos, mesmo que quase 40% dos SNPs não tenham sido analisados.

3.5 Comparações dos resultados obtidos pelos métodos empregados

Com o propósito de verificar o desempenho das metodologias para selecionar os SNPs que influenciam o fenótipo Q_1, comparamos o número de SNPs identificados como significativos e os valores de sensibilidade e especificidade, todos advindos de cada um dos métodos estudados anteriormente. Os valores são mostrados na Tabela 1.

	Regressão Linear Simples				LASSO	SPLS
	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha_{BF} = 10\%$	$\alpha_{BF} = 5\%$		
sensibilidade	0.410	0.385	0.102	0.051	0.231	0.102
especificidade	0.854	0.912	0.998	0.999	0.968	0.952
n° de SNPs identificados como significativos	3575	2177	42	27	799	1163

Tabela 1: Valores de sensibilidade, especificidade e número de SNPs identificados como significativos de cada um dos métodos estudados. Aqui $\alpha_{BF} = 10\%$ representa o nível de significância de 10% corrigido por Bonferroni.

Através da Tabela 1, notamos que a regressão linear simples com nível de significância de 5% corrigido por Bonferroni possui a melhor especificidade dentre os métodos comparados, pois seu valor é praticamente 1. Em contrapartida, a sua sensibilidade é baixa, indicando que esse modelo não considera como significativo grande parte dos SNPs que realmente influenciam o fenótipo. Notamos ainda que a regressão linear simples com nível de significância de 10% é a metodologia que possui maior sensibilidade dentre as utilizadas, embora seu valor para a especificidade seja o mais baixo de todos e isso se deve à grande quantidade de SNPs falso positivos.

Ao analisarmos os SNPs que são realmente significativos para o fenótipo Q_1, percebemos que os SNPs C13S522 e C13S523 foram identificados por todos os métodos utilizados e o SNP C4S1884 foi identificado pelo LASSO, pela regressão linear simples aos níveis de significância de 5 e 10% e pelo SPLS. Por sua vez, o SNP C1S6533 foi identificado como significativo pelo SPLS e pela regressão linear simples com α de 5% e de 10%. A identificação desses SNPs por mais de um método pode ocorrer devido à maior variação dos seus genótipos na base de dados ou devido ao elevado valor do efeito desses marcadores, os quais valem, respectivamente, 0.623466, 0.653351, 0.318125 e 0.589734, de modo que se distanciam de zero (caso em que o SNP não é considerado relevante para o fenótipo em questão).

Comparado com o SPLS e com a regressão a nível 5%, 10%, 10% Bonferroni e 5% Bonferroni, o LASSO é o método que apresenta simultaneamente bons valores de sensibilidade e especificidade, além de ter identificado (mesmo sem aumentar consideravelmente o número de falso positivos) SNPs influentes com baixo *MAF*. O SPLS, por sua vez, apresentou o pior desempenho no geral, com baixa sensibilidade e alto número de SNPs falso positivos, mesmo não considerando quase 40% dos SNPs.

A escolha entre utilizar LASSO ou regressão a um nível 5% para esse conjunto de dados ou conjuntos que tenham características parecidas depende de quão tolerante está o pesquisador em aumentar o número de SNPs falso positivos para identificar mais SNPs verdadeiro positivos.

4 | CONCLUSÃO E DISCUSSÃO

Nesse trabalho, estudamos, aplicamos e avaliamos o desempenho de três diferentes metodologias para identificação de SNPs relevantes no fenótipo Q_1 dos dados GAW17.

Ao analisarmos os desempenhos das metodologias SPLS, LASSO e regressão linear simples ao nível de significância de 5% e 10%, ambos com e sem a correção de Bonferroni, todas utilizadas para selecionar os SNPs que influenciam o fenótipo de interesse, notamos que o LASSO além de apresentar um equilíbrio entre a sensibilidade e a especificidade, ou seja, identifica relativamente bem os SNPs que realmente possuem efeito sobre Q_1 e não identifica aqueles que não são significativos, também identifica SNPs que afetam Q_1 e que são muito raros (possuem $MAF < 1\%$), são eles: C13S320, C4S1877, C4S1889 e C6S2981.

Os níveis de significância 5% e 10% corrigido por Bonferroni do teste de hipóteses em um modelo de regressão linear simples também apresentaram performance razoável. O SPLS, apesar de reduzir dimensão e selecionar variáveis simultaneamente, teve o pior desempenho no geral, além de não conseguir analisar SNPs muitos raros.

Esse trabalho foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) através da Bolsa PIBIC.

REFERÊNCIAS

Abdi, H. (2007), 'Bonferroni and Šidák corrections for multiple comparisons', *Encyclopedia of Measurement and Statistics* **3**, 103–107. 4

Almasy, L., Dyer, T. D., Peralta, J. M., Kent, J. W., Charlesworth, J. C., Curran, J. E. e Blangero, J. (2011), Genetic Analysis Workshop 17 mini-exome simulation, *in* 'BMC Proceedings', Vol. 5, BioMed Central, p. S2. 2, 8, 10, 11

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32. 2

Chun, H. e Kele¸s, S. (2010), 'Sparse partial least squares regression for simultaneous dimension reduction and variable selection', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(1), 3–25.

Chung, D. e Keles, S. (2010), 'Sparse partial least squares classification for high dimensional data', *Statistical Applications in Genetics and Molecular Biology* **9**(1). 6

Feng, Z. Z., Yang, X., Subedi, S. e McNicholas, P. D. (2012), 'The lasso and sparse least squares regression methods for SNP selection in predicting quantitative traits', *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **9**(2), 629–636. 2, 7

Goldberg, D. E. (1989), 'Genetic algorithm in search optimization and machine learning', *Addison Wesley* . 2

Ióca, M. P. e Zuanetti, D. A. (2021), 'Selection of SNP markers: Analyzing GAW17 data using different methodologies', *Brazilian Journal of Biometrics* **39**(1), 71–88. 5, 8

Lee, D., Lee, W., Lee, Y. e Pawitan, Y. (2011), 'Sparse partial least-squares regression and its applications to high-throughput data analysis', *Chemometrics and Intelligent Laboratory Systems* **109**(1), 1–8. 2, 7

Morettin, P. A. e Bussab, W. O. (2017), *Estatística básica*, Saraiva Educação SA. 3

Oliveira, F. C. (2015), 'Um método para seleção de atributos em dados genômicos', *Tese de Doutorado. Programa de Pós-graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora* . 2

Park, T. e Casella, G. (2008), 'The Bayesian lasso', *Journal of the American Statistical Association* **103**(482), 681–686. 2

Rodrigues, K. A. S. (2018), Lasso clássico e bayesiano, Technical report, IME-USP. 5

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Yazdani, A. (2014), 'Statistical approaches in genome-wide association studies', *Tese de*

Doutorado. Dipartimento di Scienze Statistiche - Scuola de Dottorato di Ricerca in Scienze Statistiche . 2

ANIELE DOMINGAS PIMENTEL SILVA - Possui graduação em Licenciatura Plena em Matemática e especialização em Educação Matemática pela Universidade Federal do Pará - UFPA. Mestre em Educação pela Universidade Federal do Oeste do Pará - UFOPA, na linha de pesquisa “Práticas Educativas, Linguagens e Tecnologias”, com ênfase em Modelagem Matemática e Tecnologias, atualmente é doutoranda pelo Programa de Pós-Graduação em Educação na Amazônia PGEDA/EDUCANORTE (2022) da Universidade Federal do Oeste do Pará – UFOPA, atuando na linha de pesquisa “Educação na Amazônia: formação do educador, práxis pedagógica e currículo”. Integra o Grupo de Estudos e Pesquisas em Educação Matemática e Interdisciplinaridade na Amazônia - GEPEIMAZ. Tem experiência como professora de matemática na educação básica pela Secretaria Municipal de Educação de Santarém-PA, professora colaboradora na UFOPA no programa PARFOR nos cursos de Licenciatura integrada em Matemática e Física e professora substituta no Instituto Federal do Amapá – IFAP em turmas do ensino médio integrado e de ensino superior.

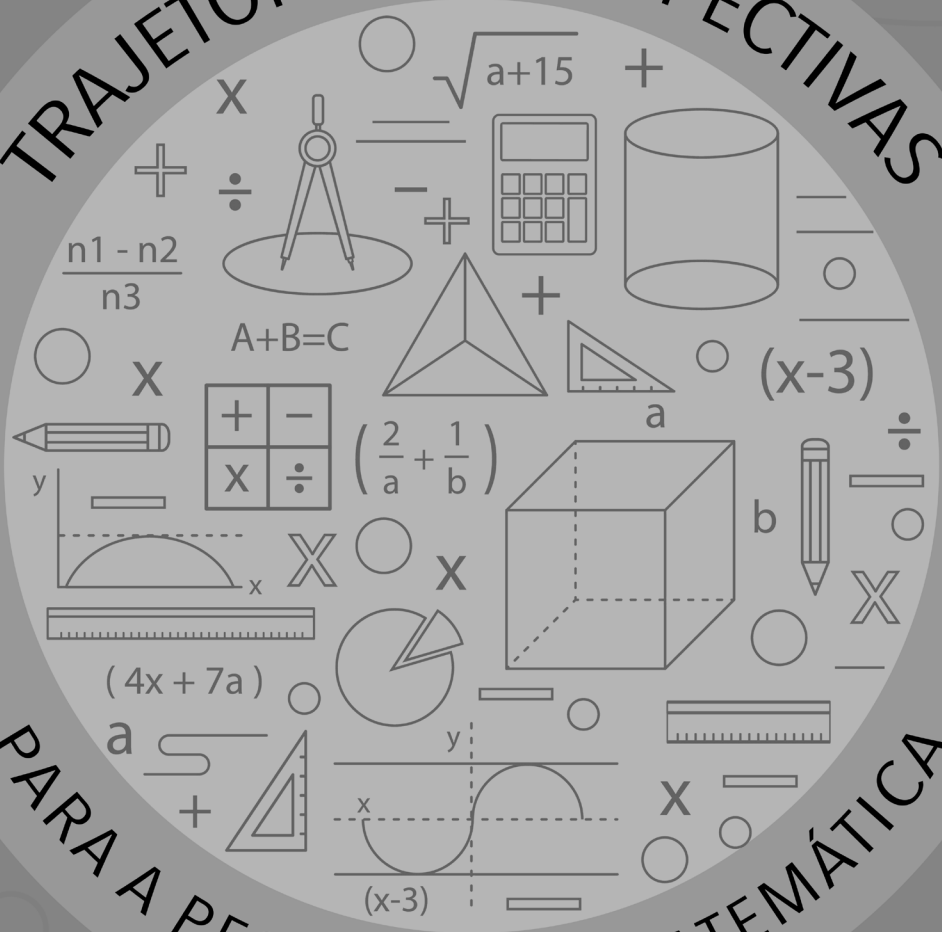
ÍNDICE REMISSIVO

ÍNDICE REMISSIVO

ÍNDICE REMISSIVO

www.atenaeditora.com.br
contato@atenaeditora.com.br
@atenaeditora
www.facebook.com/atenaeditora.com.br

TRAJETÓRIAS E PERSPECTIVAS PARA A PESQUISA EM MATEMÁTICA

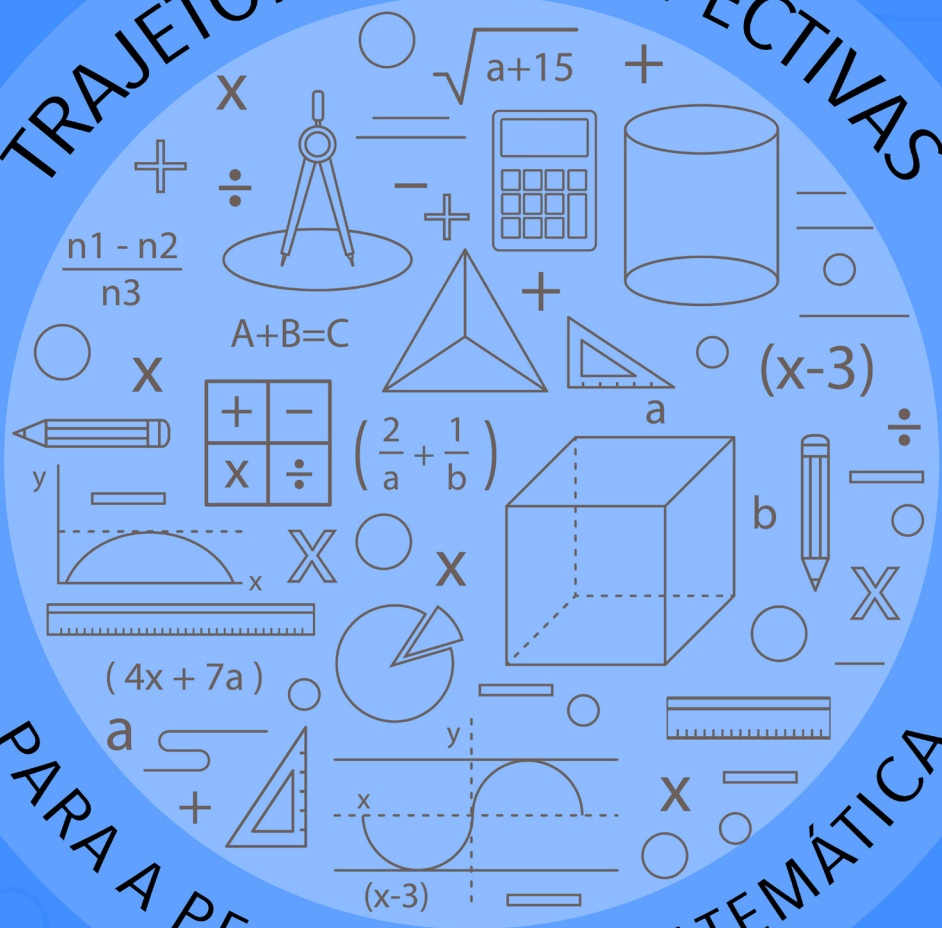


2

Atena Editora
Ano 2023

www.atenaeditora.com.br
contato@atenaeditora.com.br
@atenaeditora
www.facebook.com/atenaeditora.com.br

TRAJETÓRIAS E PERSPECTIVAS PARA A PESQUISA EM MATEMÁTICA



Atena Editora
Ano 2023

2