Análise dos dados do Cromossomo 7 do GAW17

Adriana Eva Fernandes

23 de outubro de 2023

Descrição do conjunto de dados

O conjunto de dados faz parte do Workshop de Análise Genética 17 (GAW17) e foi criado artificialmente com o objetivo de possibilitar a investigação genética de características particulares em uma população fictícia de 697 indivíduos sem parentesco, composta por 327 homens e 370 mulheres. O foco principal da análise está direcionado ao cromossomo 7 e ao fenótipo quantitativo Q1.

- O cromossomo 7, representado neste conjunto de dados simulados, contém 1063 marcadores SNPs (polimorfismos de nucleotídeo único), refletindo a variação na densidade genética ao longo do cromossomo. Alguns marcadores podem ser raros, introduzindo complexidades realistas na análise.
- O fenótipo quantitativo Q1 é um dos três fenótipos contínuos disponíveis neste conjunto de dados. Ele representa uma medida numérica associada a uma característica específica. No contexto deste trabalho, a atenção está voltada para a análise da relação entre o cromossomo 7 e as variações observadas no fenótipo Q1.

Objetivos da análise

1. **Análise da Frequência Alélica e Identificação de Variantes Raras**
   - Calcular a Frequência Relativa do Alelo Menor (MAF) para todos os marcadores presentes no cromossomo 7 e avaliar a presença de variantes raras na base de dados genômicos, analisando se há alelos com frequência alélica muito baixa e sua proporção em relação à frequência alélica total.
2. **Seleção de Marcadores Associados ao Fenótipo Q1**
   - Utilizar pelo menos dois métodos diferentes de seleção para identificar os marcadores associados ao fenótipo Q1 considerando apenas covariáveis de efeito aditivo (excluindo covariáveis de dominância ou epistasia).
   - Incluir covariáveis ambientais e de comportamento, para melhorar a precisão das associações identificadas e determinar a quantidade e quais as covariáveis relevantes para a associação fenotípica.
3. **Análise do MAF dos Marcadores Selecionados**
   - Analisar a Frequência Relativa do Alelo Menor (MAF) nos marcadores identificados como associados ao fenótipo Q1.
   - Avaliar se marcadores com MAF muito baixo foram selecionados para a associação fenotípica e justificar a inclusão ou exclusão desses marcadores com base na relação entre MAF e associação fenotípica.

MAF

O MAF representa a proporção de indivíduos dentro de uma população que carregam um determinado alelo em um locus genético específico.

É uma medida importante, pois está relacionado com a variabilidade genética em uma população. Regiões com frequências alélicas muito baixas podem indicar mutações raras ou variações genéticas pouco comuns. Por outro lado, regiões com frequências alélicas muito altas são mais comuns na população e podem ser responsáveis por características mais predominantes. Geralmente, valores de MAF abaixo de 1% ou 5% podem ser considerados baixos.

MAF

Para o cálculo do MAF, inicialmente, foi calculada a frequência alélica dos alelos na base nitrogenada ATCG em cada SNP.

A frequência alélica é dada por:

$$\hat{p}_A = \frac{2n_{AA} + n_{Aa}}{2n}$$

$$\hat{p}_a = \frac{2n_{aa} + n_{Aa}}{2n}$$

em que:

- $\hat{p}_A$ é a frequência amostral do alelo dominante.
- $\hat{p}_a$ é a frequência amostral do alelo recessivo.
- $n = n_{AA} + n_{Aa} + n_{aa}$ é o número total de genótipos.
- $n_A = 2n_{AA} + n_{Aa}$ é o número total de alelos dominantes.
- $n_a = 2n_{aa} + n_{Aa}$ é o número total de alelos recessivos.
- $2n$ é o número total de alelos.

MAF

O alelo menor em um marcador é o alelo que aparece com menos frequência, logo, é o que possui a menor frequência alélica. Portanto, após o cálculo das frequências alélicas, foi possível encontrar o **alelo menor** e a sua respectiva **frequência alélica** em cada um dos SNP's.

O gráfico a seguir mostra a frequência absoluta do MAF dos marcadores do cromossomo 7.

Análise Descritiva

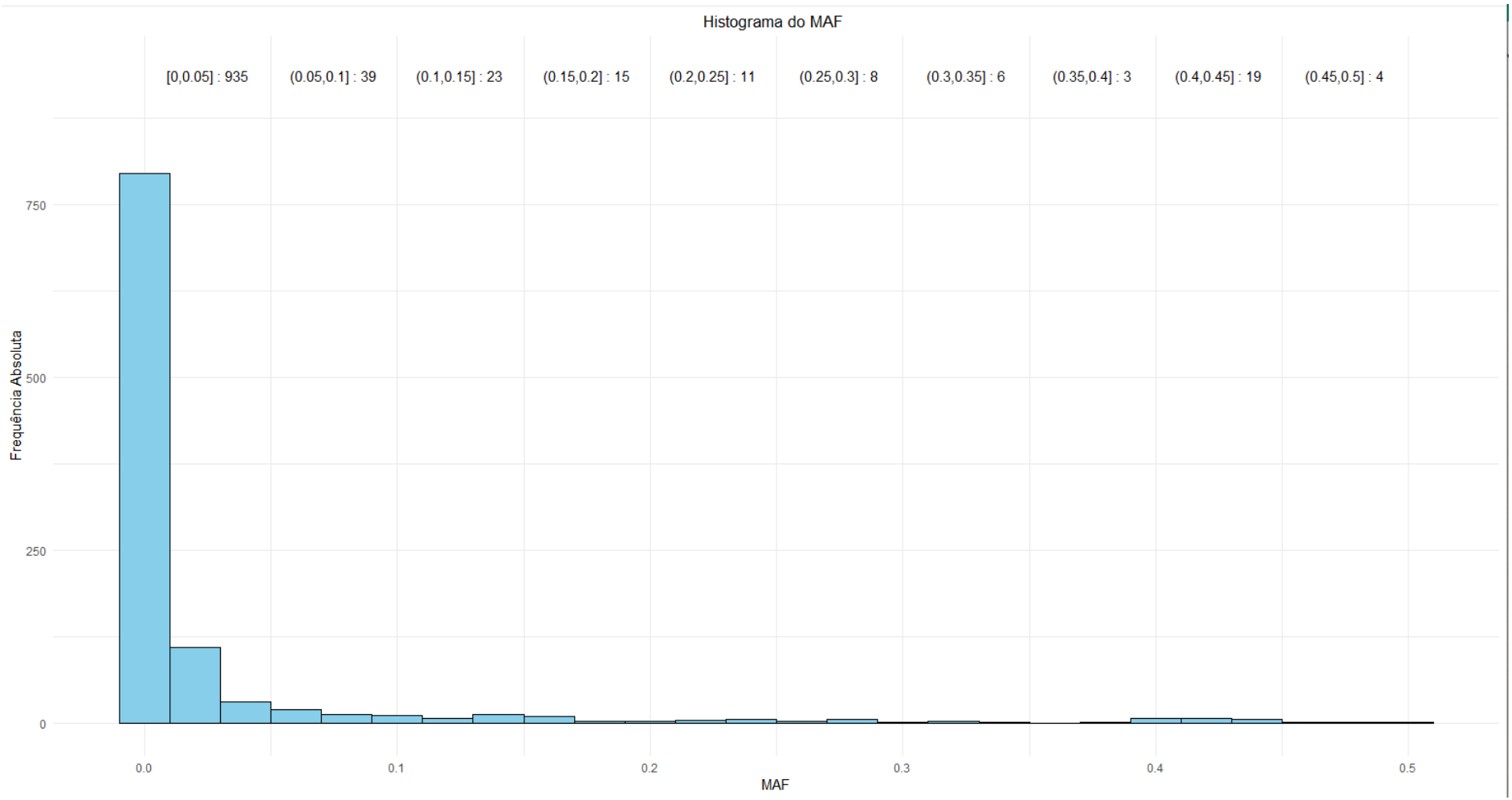

Figura: Histograma da frequência absoluta do MAF dos marcadores do cromossomo 7.

Análise Descritiva

Com base na análise do **Histograma da frequência absoluta do MAF dos marcadores do cromossomo 7**, é nítida a presença de variantes raras na base de dados, uma vez que a proporção de marcadores com a frequência relativa do alelo menor, menores que 5%, é de aproximadamente 88%.

Análise Descritiva

Tabela: Medidas de resumo do MAF

Min.	1ºQuartil	Mediana	Média	3ºQuartil	Max.
0.0007174	0.0007174	0.0021521	0.0290440	0.0100430	0.4971306

Análise Descritiva

Com base nas **Medidas de resumo do MAF** vemos que :

- O valor mínimo do MAF observado é 0.0007174, o que significa que há pelo menos um marcador genético com uma frequência muito baixa do alelo menor.
- A mediana é 0.0021521, o que indica que metade dos marcadores tem um MAF igual ou inferior a esse valor. Isso confirma que uma proporção considerável dos marcadores tem frequências baixas do alelo menor.
- O valor máximo do MAF observado é 0.4971306. Isso indica que há pelo menos um marcador com uma frequência do alelo menor relativamente alta. No entanto, ainda é abaixo de 0.5, o que significa que nenhum marcador possui um MAF maior que 50%

As medidas resumo indicam que a maioria dos marcadores apresentam frequências do alelo menor baixas, o que é consistente com a presença de variantes raras na base de dados. A média um pouco mais alta pode indicar que alguns marcadores têm frequências de alelo menor mais substanciais, mas em geral, a tendência é de que muitas variantes sejam raras nessa população.

Categorização das variáveis

Antes de prosseguir com a análise de associação entre as variáveis, foi necessário realizar uma etapa de transformação nos genótipos presentes nas bases nitrogenadas. Essa etapa visou simplificar a representação dos genótipos para facilitar a avaliação das relações entre eles e o fenótipo em estudo. A transformação foi realizada atribuindo valores numéricos aos diferentes genótipos, resultando em uma representação padronizada.

A tabela apresentada a seguir ilustra a conversão dos genótipos em seus respectivos valores numéricos:

Tabela: Conversão dos genótipos em valores numéricos nas bases nitrogenadas.

Genótipo	Valor	Genótipo	Valor	Genótipo	Valor
C/C	1	A/A	-1	G/G	1
T/T	-1	G/A	0	T/C	0
C/T	0	A/G	0	G/C	0
T/C	0	C/A	0	C/G	0
T/A	0	A/T	0	G/T	0

em que -1 representa os genótipos homozigotos recessivos, 0 representa os genótipos heterozigotos e 1 representa os genótipos homozigotos dominantes.

Associação entre marcadores e fenótipos

No contexto da análise de dados genéticos, muitas vezes estamos interessados em saber se existe alguma associação entre uma região genética e um fenótipo. Nesse estudo estamos interessados em saber quais dos 1063 marcadores do cromossomo 7, de fato, tem associação com o fenótipo Q1.

Como temos um número considerável de SNP's, é necessário que metodologias específicas sejam utilizadas para encontrar possíveis associações entre marcadores e fenótipos, entre todas as disponíveis e em teste. Em vez de testar todos os subconjuntos possíveis de covariáveis (o que se torna inviável uma vez que $p = 1063$), e escolher o melhor deles, podemos utilizar métodos que fazem essa busca de uma maneira mais eficiente.

Dito isso, a descrição e aplicação dos métodos utilizados para a seleção de covariáveis, estão descritas nos próximos slides.

Metodologias utilizadas

No contexto dessa análise, utilizamos as abordagens de seleção de marcadores associados ao fenótipo Q1, bem como a inclusão de covariáveis ambientais e de comportamento. Entre os métodos de seleção Forward, Stepwise, Backward, Lasso e Regressão Ridge.

Justificativas:

Considerando o cenário em que lidamos com um conjunto de 1066 covariáveis, abrangendo regiões genéticas, variáveis de comportamento e ambientais, e tendo à nossa disposição apenas 697 observações ($p > n$), o método Backward seria impraticável devido ao grande número de covariáveis. Optar por métodos como Forward, Stepwise e Lasso e Ridge seria apropriado. Tanto Lasso quanto regressão Ridge ajudam a evitar overfitting ao encolher estimativas para zero. Devido à sua seleção mais restrita, Lasso é preferível. Portanto, os métodos escolhidos são Forward, Stepwise e Lasso.

Forward

Sua abordagem é incremental, começando com um modelo só com o intercepto e adicionando variáveis uma de cada vez, escolhendo aquela que melhora mais o ajuste do modelo em cada etapa.

A covariável (variável preditora) que deve ser incluída no modelo é a que mais reduz os valores de AIC ou BIC ou que apresenta menor valor-p no teste de significância.

1. **Vantagem - Justificativa computacional**: para p grande, não podemos calcular a melhor sequência de subconjuntos, mas podemos sempre calcular a sequência progressiva passo a passo (mesmo quando $p > n$).
2. **Vantagem - Justificativa estatística**: um preço é pago em termos de variância do estimador quando queremos selecionar o melhor subconjunto de variáveis para a base de estimação. Esse método faz uma procura de variáveis mais restrita que, consequentemente, apresenta menor variância quando aplicado em outras bases de dados, mas talvez mais viés.
3. **Desvantagem**: produz uma sequência de modelos encaixados. Nesse sentido, pode não fazer a seleção ótima de variáveis.

Stepwise

O método de seleção Stepwise alterna entre passos forward e backward. O modelo começa apenas com o intercepto e usa passos forward para incluir variáveis. No entanto, a cada nova variável preditora incluída, o método realiza passos backward para avaliar se alguma variável que já esteja no modelo poderia ser excluída.

O procedimento termina quando nenhuma variável preditora mereça ser incluída ou excluída do modelo atual.

1. **Vantagem**: é preferido em relação aos anteriores por, geralmente, ser mais parcimonioso e selecionar conjuntos menores de variáveis.
2. **Desvantagem:** o método pode facilmente levar ao overfitting. Isso pode resultar em um modelo que se ajusta excessivamente aos dados de treinamento, mas que não generaliza bem para novos dados.

Lasso

A ideia principal do Lasso é adicionar a restrição

$$\sum_{j=1}^{p} |\beta_j| \leq c,$$

à fórmula tradicional dos mínimos quadrados usada para estimar o valor de β_j's. Com esta adição, as estimativas de alguns β_j's são aproximadamente zero e, portanto, o método seleciona como relevantes as covariáveis que apresentam $\hat{\beta}_j \neq 0$.
Quando c é muito grande, os coeficientes tendem a se aproximar dos valores que seriam obtidos pelo método dos mínimos quadrados, sem a regularização.

Lasso

O objetivo é encontrar os valores dos coeficientes β que minimizam a soma dos quadrados dos resíduos e ao mesmo tempo estão sujeitos a uma penalização proporcional à soma dos valores absolutos dos coeficientes.

$$\hat{\beta}_{\text{Lasso}} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{1}$$

onde $\lambda \geq 0$ é um parâmetro de regularização que controla a quantidade de encolhimento. Quanto maior o valor de λ, maior será a quantidade de coeficientes que se tornarão iguais a zero.
A validação cruzada é uma estratégia valiosa para determinar o melhor valor de λ no modelo Lasso. Ela equilibra a habilidade do modelo em se ajustar aos dados com a regularização para prevenir overfitting.

Critério de seleção

Optamos por adotar o critério de seleção AIC (Critério de Informação de Akaike) para a escolha do modelo.O AIC é fundamentado em princípios estatísticos que levam em consideração dois fatores essenciais ao construir um modelo estatístico: a qualidade do ajuste e a complexidade do modelo. A ideia é que, ao adicionar mais variáveis ao modelo, a capacidade de ajuste geralmente melhora, mas a complexidade do modelo aumenta. O AIC busca encontrar um equilíbrio entre esses dois fatores. Ele recompensa um bom ajuste aos dados, mas ao mesmo tempo penaliza por cada variável adicional incluída no modelo. Isso impede que o modelo se torne excessivamente complexo, evitando problemas de sobreajuste.

Procedimentos feitos no método Forward

O conjunto de dados foi segmentado em 10 partes, uma vez que havia um grande número de marcadores. O método forward não apresentou um desempenho satisfatório nesse cenário, mas a abordagem de dividir o conjunto de dados em 10 partes e executar 10 modelos distintos, um para cada segmento, mostrou-se eficaz. Após a seleção de variáveis em cada um dos 10 modelos, consolidamos as variáveis escolhidas em cada um deles. Com essa lista combinada de variáveis selecionadas, realizamos uma etapa subsequente do método forward, agora empregando apenas essas variáveis. Essa abordagem refinada permitiu uma seleção mais precisa, levando em consideração as variáveis que foram consistentemente identificadas como relevantes em diferentes partes do conjunto de dados.

Resultados - Forward

O método de seleção Forward selecionou 173 covariáveis como significativas:

- 171 marcadores (C7S1598, C7S1362, C7S1033, C7S4110, C7S163, C7S1247, C7S1070, C7S164, C7S2330, C7S4743, C7S1829, C7S3613, C7S4515, C7S5250, ...).
- uma variável de ambiente (idade).
- uma variável de comportamento (smoke).
  Entretanto, o método forward tendeu a selecionar um número excessivo de covariáveis como significativas. Essa abordagem frequentemente resultava na inclusão de covariáveis que talvez não fossem verdadeiramente relevantes, uma vez que, uma vez adicionadas, elas permaneciam no modelo permanentemente.

Resultados - Forward

Para superar essa limitação, optaremos por utilizar o método de seleção stepwise. Ao contrário do método forward, o stepwise tem a capacidade de tanto adicionar quanto remover covariáveis ao modelo durante suas iterações. Isso proporciona uma vantagem, pois podemos eliminar covariáveis que podem ter sido adicionadas inicialmente, mas que posteriormente se mostram não tão significantes. Essa flexibilidade torna o método stepwise mais eficiente e apropriado para o nosso cenário de análise de dados.

Resultados - Stepwise

Step	Variable	A/R	R-Sq	C(p)	AIC	RMSE
1	idade	add	0.116	569.986	1897.301	0.9411
2	smoke	add	0.179	481.816	1847.674	0.9075
3	C7S1598	add	0.218	428.083	1815.739	0.8863
4	C7S1362	add	0.235	404.902	1801.848	0.8769
5	C7S1033	add	0.250	385.707	1790.185	0.8690
6	C7S4110	add	0.265	366.315	1778.118	0.8609
...	...	...	...	...	...	...
102	C7S4983	rem	0.640	14.799	1464.418	0.6471
103	C7S4696	add	0.642	13.901	1462.476	0.6458
104	C7S3291	add	0.644	13.299	1460.917	0.6447
105	C7S1917	add	0.646	12.769	1459.439	0.6436
106	C7S2063	add	0.648	11.186	1456.484	0.6419
107	C7S2306	add	0.650	10.610	1454.899	0.6408
108	C7S5013	rem	0.650	9.233	1453.768	0.6406
109	C7S1047	add	0.652	8.490	1451.935	0.6394
110	C7S76	add	0.654	7.850	1450.227	0.6382
111	C7S2126	add	0.655	7.442	1448.827	0.6372
112	C7S3590	add	0.657	7.087	1447.486	0.6362

Resultados - Stepwise

Posteriormente, seguindo também o critério AIC, optamos por aplicar o método stepwise utilizando as variáveis que haviam sido identificadas como significantes nos 10 modelos anteriores. Esse processo resultou na seleção de um total de 109 variáveis consideradas relevantes. É notável a diferença significativa entre essa seleção e o resultado obtido pelo método forward, no qual 173 variáveis foram escolhidas como significativas.

No caso em que o método forward escolheu um número maior de variáveis (173), ele tendeu a ser mais inclusivo, incorporando um conjunto mais amplo de covariáveis ao modelo. Por outro lado, o método stepwise resultou em uma seleção mais restrita de variáveis (109), mostrando uma abordagem mais seletiva. Isso implica que o método stepwise, ao optar por um conjunto menor de variáveis, pode estar favorecendo uma abordagem mais parcimoniosa e evitando a supercomplexidade do modelo.

Em relação à eficiência computacional, observou-se que o método stepwise foi significativamente mais rápido em comparação com o método forward.

Resultados - Lasso

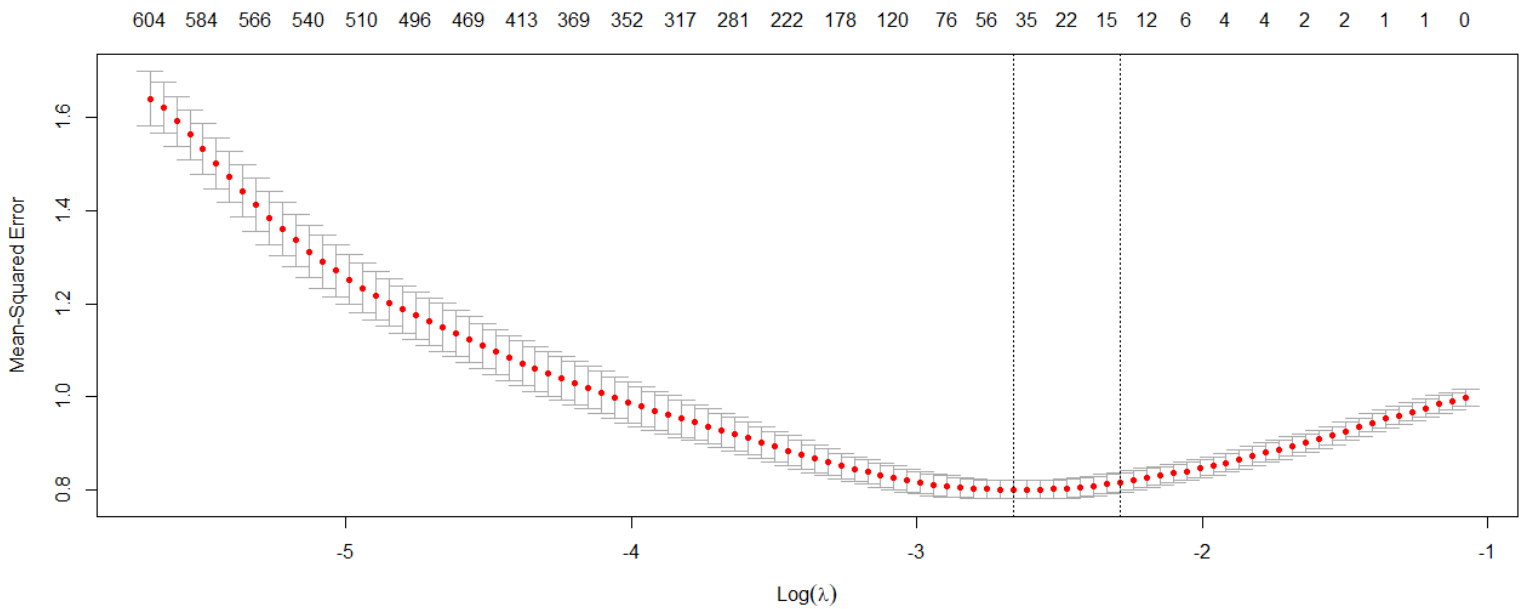

Figura: Gráfico de risco

Resultados - Lasso

O gráfico da figura 2 apresenta como o Erro Quadrático Médio varia em função do lambda. O melhor lambda vai selecionar em torno de 35 variáveis. Esse gráfico nos mostra que quanto maior o lambda menos variáveis ele seleciona e quanto menor o lambda, mais variáveis serão selecionadas e mais próximo fica do método de mínimos quadrados.

O valor $\lambda = 0.07322843$ obtido indica que, nesse ponto, o modelo alcança a minimização do erro, ao mesmo tempo em que mantém uma penalização moderada nos coeficientes. Isso sugere que o seu modelo Lasso está encontrando um equilíbrio adequado entre o ajuste aos dados e a simplicidade do modelo, resultando em um desempenho otimizado.

Resultados - Lasso

Tabela: Variáveis selecionadas e estimativas viesadas dos coeficentes via Lasso

Covariável	Coeficiente	Covariável	Coeficiente	Covariável	Coeficiente
C7S163	-0.8355	C7S164	-0.0436	C7S612	0.2069
C7S723	-0.0724	C7S995	0.0000	C7S1033	0.4545
C7S1070	0.9027	C7S1247	-1.1004	C7S1362	-0.8298
C7S1393	-0.0242	C7S1582	-0.0006	C7S1589	0.0184
C7S1598	-0.0835	C7S1632	-0.0390	C7S1638	-0.0875
C7S1697	-0.1144	C7S1829	0.2648	C7S2330	-0.4629
C7S2893	0.0037	C7S3012	-0.0685	C7S3131	0.1338
C7S3462	-0.0259	C7S3974	0.0111	C7S3996	-0.0237
C7S4110	0.1346	C7S4515	-0.1359	C7S4631	-0.0041
C7S4743	-0.0528	C7S4983	-0.0182	C7S5013	-0.0518
C7S5091	0.0133	C7S5211	-0.1108	C7S5250	-0.0051
				smoke	0.3971
				idade	0.0152

Estimativas não viesadas para os coeficientes das variáveis que foram selecionadas no método Lasso

O método Lasso não fornece estimativas precisas, pois tende a gerar estimativas enviesadas. Como solução, após a seleção das covariáveis pelo Lasso, é comum refinar as estimativas utilizando o método de mínimos quadrados. Isso é feito para obter estimativas mais confiáveis das variáveis selecionadas.

A tabela a seguir mostra as covariáveis e os respectivos valores estimados dos coeficientes do modelo de regressão.

Estimativas não viesadas para os coeficientes das variáveis que foram selecionadas no método Lasso

Tabela: Valores das Covariáveis e Coeficientes do Modelo de Regressão

Covariável	Coeficiente	Covariável	Coeficiente	Covariável	Coeficiente
(Intercept)	6.21015	C7S163	-2.19208	C7S164	-0.14578
C7S612	1.83379	C7S723	-1.70543	C7S1033	1.14805
C7S1070	2.87798	C7S1247	-3.06114	C7S1362	-1.83544
C7S1393	-0.03828	C7S1582	-0.06822	C7S1589	0.08157
C7S1598	-0.06967	C7S1632	-0.06560	C7S1638	-0.15052
C7S1697	-0.23612	C7S1829	1.24575	C7S2330	-2.42434
C7S2893	0.12887	C7S3012	-0.66213	C7S3131	0.99617
C7S3462	-1.93609	C7S3974	0.05428	C7S3996	-0.71483
C7S4110	0.16252	C7S4515	-2.18434	C7S4631	-0.06785
C7S4743	-0.13271	C7S4983	-0.18833	C7S5013	-0.24588
C7S5091	0.05348	C7S5211	-0.78380	C7S5250	-0.35152
				smoke	0.54572
				idade	0.01958

Escolha do melhor método de seleção para nosso problema

Após a seleção de covariáveis através dos métodos de seleção Forward, Stepwise e Lasso, o método Lasso se mostrou mais eficiente na seleção.

Uma das razões pelas quais o método Lasso se mostra mais eficiente na seleção de variáveis é que ao forçar alguns coeficientes a zero, consegue lidar de forma mais robusta com multicolinearidade do que os métodos Forward e Stepwise, que podem incluir variáveis redundantes devido a essa interdependência.

Além disso, o método Lasso apresentou uma maior eficiência computacional. Ele foi capaz de realizar a seleção de variáveis em um tempo relativamente curto comparado aos outros métodos.

Outro aspecto crucial é a regularização oferecida pelo método Lasso. Ao introduzir o parâmetro de penalização λ, o Lasso consegue controlar o grau de regularização aplicado aos coeficientes, permitindo um melhor equilíbrio entre o ajuste do modelo aos dados e a simplificação do mesmo, evitando overfitting.

Sendo assim, a seguir será apresentada uma análise do MAF dos marcadores selecionados pelo Lasso.

Análise do MAF dos Marcadores Selecionados

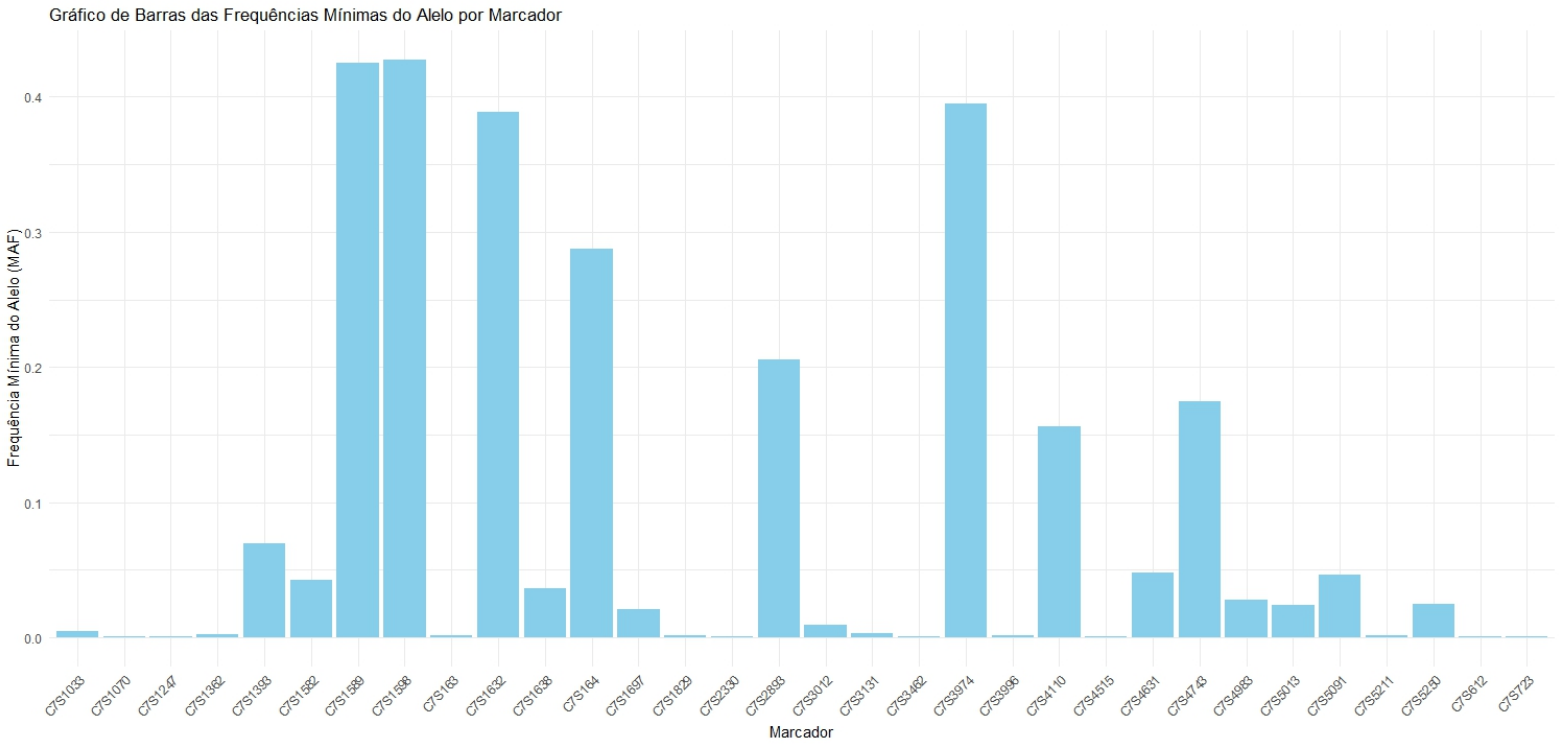

Figura: Gráfico de barras do MAF com os marcadores selecionados pelo Lasso

Análise do MAF dos Marcadores selecionados

A análise do **Gráfico de barras do MAF com os marcadores selecionados pelo Lasso** revela que alguns dos marcadores selecionados apresentam uma Frequência Alélica Mínima (MAF) muito baixa, indicando a presença de variantes raras entre os marcadores escolhidos pelo método Lasso.

A baixa frequência das variantes raras pode levar a um poder estatístico insuficiente para detectar associações verdadeiras. Métodos estatísticos tradicionais podem não ser sensíveis o suficiente para detectar associações com variantes raras.

Portanto, é necessário o desenvolvimento de métodos mais sensíveis que levem em consideração a baixa frequência dessas variantes. Isso permitirá uma compreensão mais profunda das bases genéticas de doenças complexas e características humanas.