PATRICIA MENDES DOS SANTOS

REGRESSÃO QUANTÍLICA APLICADA AO ESTUDO DE SELEÇÃO GENÔMICA PARA CARACTERÍSTICAS ASSIMÉTRICAS DE SUÍNOS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae.*

VIÇOSA
MINAS GERAIS – BRASIL
2016

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T

S237r
2016

Santos, Patricia Mendes dos, 1989-

Regressão quantilica aplicada ao estudo de seleção genômica para caracteristicas assimétricas de suínos / Patricia Mendes dos Santos. – Viçosa, MG, 2016.

x, 58f. : il. (algumas color.) ; 29 cm.

Inclui apêndices.

Orientador: Ana Carolina Campana Nascimento.

Dissertação (mestrado) - Universidade Federal de Viçosa.

Referências bibliográficas: f.41-43.

1. Suíno - Melhoramento genético. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-graduação em Estatistica Aplicada e Biometria. II. Título.

CDD 22. ed. 636.4

PATRICIA MENDES DOS SANTOS

REGRESSÃO QUANTÍLICA APLICADA AO ESTUDO DE SELEÇÃO GENÔMICA PARA CARACTERÍSTICAS ASSIMÉTRICAS DE SUÍNOS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae.*

APROVADA: 26 de fevereiro de 2016.

Moysés Nascimento
(Coorientador)

Camila Ferreira Azevedo

Rodrigo Reis Mota

Ana Carolina Campana Nascimento
(Orientadora)

A Deus e aos meus pais, Walter e Maria Aparecida,

Aos meus irmãos Walter e Mirian,

Ao meu namorado Bruno.

AGRADECIMENTOS

Primeiramente agradeço a Deus, por sempre estar presente em minha vida e por ter me sustentado durante toda essa jornada.

Aos meus pais, Walter e Maria Aparecida pelo amor, pelo apoio incondicional e por sempre acreditarem em mim. Vocês são o meu alicerce.

Aos meus irmãos, Mirian e Walter, pelo carinho, incentivo e por estarem sempre ao meu lado.

Ao Bruno pelo amor, companheirismo e por nunca me deixar desistir. Agradeço a Deus por ter você em minha vida.

À minha orientadora Ana Carolina Campana Nascimento pela paciência, pelos ensinamentos, por sua dedicação e apoio ao longo desta pesquisa.

Aos meus coorientadores Moysés Nascimento e Fabyano Fonseca e Silva por contribuírem diretamente no meu aprendizado, pelas valiosas sugestões e disponibilidade.

Aos amigos de mestrado, José Alfredo e Geraldo, pela amizade, pelos momentos de estudo, pela paciência e disposição em me ajudar. Vocês são pessoas muito especiais.

Aos meus amigos, Leonardo e Elizena, pela amizade, pelos momentos de alegria e por torcerem sempre por mim. Vocês são as melhores pessoas que conheci.

Aos professores e funcionários do departamento de estatística da Universidade Federal de Viçosa, por contribuíram com seus ensinamentos e por todo apoio dado durante minhas atividades acadêmicas.

Aos membros da banca, Camila Ferreira Azevedo e Rodrigo Reis Mota por aceitarem o convite e por estarem dispostos a dar suas contribuições para este trabalho.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria pela oportunidade.

À CAPES, pela concessão da bolsa de estudos.

Por fim, a todas as pessoas que de uma forma ou de outra contribuíram para a realização deste trabalho.

BIOGRAFIA

PATRICIA MENDES DOS SANTOS, filha de Maria Aparecida Mendes dos Santos e Walter dos Santos, nasceu na cidade de Visconde do Rio Branco, Minas Gerais, em 19 de Outubro de 1989.

Em Março de 2008, ingressou no curso de Licenciatura em Matemática na Universidade Federal de Viçosa, graduando-se em Março de 2014.

Em Março do mesmo ano, iniciou o curso de mestrado do Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa da dissertação em 26 de fevereiro de 2016.

SUMÁRIO

RESUMO

SANTOS, Patricia Mendes dos, M.Sc., Universidade Federal de Viçosa, fevereiro de 2016. **Regressão quantílica aplicada ao estudo de seleção genômica para características assimétricas de suínos**. Orientadora: Ana Carolina Campana Nascimento. Coorientadores: Moysés Nascimento e Fabyano Fonseca e Silva.

Em programas de melhoramento, seja para predição do mérito genético individual ou para identificar regiões genômicas responsáveis por fenótipos de interesse econômico, o uso de informações dos marcadores SNPs (*Single Nucleotide Polymorphisms*) tem se tornado uma importante ferramenta e para tanto, diferentes métodos estatísticos têm sido empregados. Dentre os métodos comumente utilizados, destaca-se o método Lasso bayesiano (BLASSO) que, como as demais metodologias apresentadas na literatura, estima apenas o efeito dos marcadores em termos do valor médio da característica de interesse. Porém, em algumas situações, os fenótipos avaliados não possuem distribuição simétrica e, portanto, uma modelagem considerando a estimação dos efeitos de marcadores em diferentes níveis da variável de interesse pode ser mais adequada. Uma metodologia alternativa e ainda pouco explorada na seleção genômica, é a Regressão Quantílica Regularizada, a qual incorpora naturalmente o estudo de diferentes níveis da distribuição dos valores fenotípicos, bem como a seleção de variáveis e a regularização do processo de estimação. Diante do exposto, este trabalho propõe a utilização da Regressão Quantílica Regularizada, visando predizer os valores genéticos de suínos para as características de rendimento de carcaça (RCARC), espessura de bacon (EBACON) e espessura de toucinho imediatamente após a última costela na linha dorso-lombar (ETUC) em diferentes níveis da distribuição dessas variáveis. Os resultados obtidos por esta metodologia, em termos de acurácia, da classificação dos animais para seleção e efeitos de marcadores, foram comparados com aqueles advindos do método BLASSO. Para tanto, foram utilizados dados de uma população F_2, referente a 345 suínos, obtidos pelo cruzamento das raças Piau x Comercial. As funções quantílicas foram ajustadas para $(\tau = 0{,}05 \text{ a } \tau = 0{,}95)$ e para fins de comparação das metodologias, considerou-se nos ajustes via RQR além dos valores do parâmetro de encolhimento (λ) estimado a partir do método

BLASSO, um grid de valores considerando valores variando de 0 até o valor fornecido pelo BLASSO, com intervalo de 0,5. Os modelos de regressão quantílica se apresentaram como uma alternativa interessante quando o interesse é predizer os valores genômicos dos animais, uma vez que os mesmos apresentaram valores de acurácia maiores ou iguais àqueles obtidos pelo BLASSO. Além disso, os resultados dos métodos para as variáveis ETUC e EBACON foram semelhantes e para RCARC foram diferentes no que se refere à classificação dos animais. Quanto ao padrão dos efeitos de marcadores, os resultados encontrados pelos métodos para as variáveis foram diferentes.

ABSTRACT

SANTOS, Patricia Mendes dos, M.Sc., Universidade Federal de Viçosa, February, 2016, **Quantile regression applied to genomic selection for asymmetric traits in pigs.** Advisor: Ana Carolina Campana Nascimento. Co-advisors: Moysés Nascimento and Fabyano Fonseca e Silva.

In breeding programs, used to predict the individual genetic merit or to identify genomic regions responsible for phenotypes of economic interest, the use of information from SNP markers (Single Nucleotide Polymorphisms) has become an important tool and due to it, different statistical methods have been employed. Among the methods commonly used for this purpose, the most used is the Bayesian Lasso method called BLASSO that, like other methods presented in the literature, only estimates the markers effect related to the average value of the characteristic of interest. However, in some situations, the evaluated phenotypes don't have symmetrical distribution and thus a modeling considering estimation of the effect on markers in different levels of variable of interest may be more suitable. An alternative methodology and still not much explored in genomic selection, is the regularized quantile regression, which naturally includes the study of different levels of the phenotypic values distribution as well as the selection of variables and the regularization of the estimation process. Given the situation above, this paper proposes the use of regularized quantile regression, aiming to predict the genetic values of pigs for characteristics carcass yield (CY), bacon depth (BD) and midline backfat thickness immediately after the last rib dorsolumbar (LRBF) at different levels of the distribution of these variables. Moreover, the results obtained by this method, in terms of accuracy of classification and selection of animals for significance markers are compared with those arising from the usual method, (BLASSO). To obtain this result, we used data from an F_2 population of 345 pigs, obtained by crossing the Piau x Commercial race. The quantile functions were adjusted for (τ = 0.05 to τ = 0.95) and for comparison of methodologies, considered in RQR via adjustments in addition to the values of the shrinkage parameter (λ) estimated from the BLASSO method, a grid values are values ranging from 0 to the value provided by BLASSO with an interval of 0,5. The

quantile regression models were presented as an interesting alternative when interest is to predict the genomic values of the animals, since they had higher accuracy than or equal to those obtained by BLASSO. Moreover, the results of the methods for LRBF and BD variables were similar and were CY different as regards the classification of the animals. Regarding the pattern of markers effects, the results found by the methods for the variables were different.

1. INTRODUÇÃO GERAL

A carne de porco é uma das mais consumidas no mundo. Segundo a Associação Brasileira de Proteína Animal (ABPA), foram produzidas 110.606 mil toneladas de carne suína no ano de 2014. O Brasil é o quarto maior produtor e exportador mundial deste tipo de carne, sendo responsável por cerca de 3% (3344 mil toneladas) da produção mundial e de 10% do volume exportado no mundo, chegando a lucrar mais de US$ 1 bilhão por ano (MAPA, 2015).

O contínuo aumento na produção e exportação tem tornado essa atividade cada vez mais importante para a economia brasileira. E dentre os fatores que têm contribuído para o aumento da produção, bem como para a melhoria da qualidade da carne, destacam-se os investimentos em pesquisa e na evolução genética da espécie, realizados nos últimos 20 anos (MAPA, 2015).

Nos programas de melhoramento, o uso de informações dos marcadores moleculares para predição do mérito genético dos indivíduos tem se tornado uma importante ferramenta. Em razão do desenvolvimento de novas classes de marcadores moleculares, como os SNPs (*Single Nucleotide Polymorphisms*), Meuwissen et al. (2001) idealizaram a seleção genômica ampla (*Genome Wide Selection* - GWS), a qual consiste em utilizar informações dos marcadores para estimar seus efeitos sobre características fenotípicas de interesse, visando predizer o mérito genético para posterior seleção de indivíduos.

Nesse sentido, muitos estudos voltados para o melhoramento genético de suínos têm sido realizados. Destacam-se os estudos de Azevedo et al. (2013) que utilizaram regressão via componentes independentes para estimação de valores genéticos genômicos e dos efeitos de marcadores SNPs para características de carcaça de uma população F_2 de suínos (Piau x Comercial), e Paixão et al. (2012) que utilizaram marcadores microssatélites para identificar QTL's (*Quantitative Trait Locus*) associados à características de carcaça e qualidade de carne desta mesma população.

No contexto da seleção genômica, a utilização direta dos marcadores SNPs na seleção dos melhores indivíduos ainda é um desafio, já que na

maioria das vezes não é possível estimar livremente o efeito de cada SNP sobre o fenótipo, devido a problemas de multicolinearidade e de alta dimensionalidade oriundos da quantidade, geralmente superior, desses marcadores em relação ao número de observações.

De acordo com Gianola et al. (2003), essa situação requer a utilização de métodos estatísticos que considerem a seleção de covariáveis (problema de multicolinearidade) e a regularização do processo de estimação (problema de dimensionalidade). Dentre os métodos comumente utilizados, destaca-se o método Lasso bayesiano (*Bayesian Least Absolute Shrinkage and Selection Operator* - BLASSO), que, como as demais metodologias apresentadas na literatura, considera o comportamento médio da variável em estudo para realizar predições e posteriormente selecionar indivíduos. Especificamente, aplicando essa metodologia, a relação funcional entre o fenótipo e os SNPs é explicada por meio de um comportamento médio, possibilitando ao pesquisador selecionar indivíduos apenas em termos médios da população.

Entretanto, em algumas situações, os fenótipos avaliados não possuem distribuição simétrica e, portanto, uma modelagem considerando a estimação dos efeitos dos SNPs sobre o valor médio da característica pode não ser a estratégia mais adequada. Neste sentido, uma metodologia alternativa e ainda pouco explorada na seleção genômica, é a utilização da Regressão Quantílica Regularizada, a qual incorpora naturalmente o estudo de diferentes níveis da distribuição dos valores fenotípicos, bem como a seleção de variáveis e a regularização do processo de estimação, de modo que a mesma possibilite a obtenção de valores genéticos genômicos para diferentes níveis (quantis) da distribuição do fenótipo em estudo e não somente em relação à média, como nos métodos usuais.

1.1 Objetivos

1.1.1. Objetivo Geral

O objetivo principal deste trabalho é propor a utilização da Regressão Quantílica Regularizada em estudos de seleção genômica, visando predizer os valores genéticos de suínos para características assimétricas associadas à carcaça de suínos em diferentes níveis da distribuição das variáveis de interesse. Além disso, os resultados obtidos por esta metodologia, em termos de acurácia, concordância dos animais para seleção e efeitos de marcadores, serão comparados com aqueles advindos do método, Lasso bayesiano (BLASSO).

1.1.2. Objetivos Específicos

Especificamente, pretende-se:

- Identificar a potencialidade da Regressão Quantílica Regularizada para a estimação de valores genéticos genômicos para suínos que possuem fenótipos com distribuição assimétricas.
- Utilizar os 10% maiores valores genéticos genômicos preditos pela Regressão quantílica regularizada e o BLASSO para verificar o grau de concordância na classificação dos animais.
- Utilizar a Regressão Quantílica Regularizada para estimar efeitos de marcadores em relação a um conjunto de características de carcaça, para verificar se estes efeitos diferem dos efeitos estimados pelo BLASSO.

2. REVISÃO DE LITERATURA

2.1. Regressão Quantílica

2.1.1. Introdução

A análise de regressão é uma técnica estatística utilizada com o objetivo de investigar e modelar a relação funcional entre uma variável resposta e uma ou mais variáveis explicativas (MONTGOMERY et al., 2001). Em situações onde há somente uma variável explicativa, é comum o uso do modelo de regressão linear simples (RLS), que analisa a relação entre a variável explicativa X e a variável resposta Y e essa relação é representada por uma equação de reta:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1)$$

em que β_0 é o intercepto e β_1 é o coeficiente angular, sendo estas constantes desconhecidas e e_i são os erros aleatórios. Para este modelo, considera-se que os erros têm média zero, variância desconhecida σ^2 e são homocedásticos. Ademais, eles devem ser normalmente distribuídos e independentes entre si.

Os parâmetros do modelo (1) são estimados por meio do método dos mínimos quadrados ordinários (MQO), em que se assumem como estimativas dos parâmetros os valores que minimizam a soma de quadrados dos erros.

Entretanto, apesar de sua facilidade de implementação, a RLS apresenta algumas restrições, tais como: Os pressupostos do modelo nem sempre são válidos para dados reais. Neste caso, por exemplo, se o pressuposto de normalidade fosse violado, poderia ocorrer viés nos p-valores e assim ocasionaria testes de hipóteses inválidos, uma vez que o cálculo dos p-valores se baseiam nesse pressuposto (HAO e NAIMAN, 2007). Tem-se ainda, que nem sempre o pressuposto de homocedasticidade é válido. Além disso, quando a distribuição apresenta caudas pesadas, a média condicional pode tornar-se uma medida inadequada de localização central, pois está

fortemente influenciada por valores discrepantes (outliers) (HAO e NAIMAN, 2007).

Nesse sentido Koenker e Bassett (1978) introduziram a regressão quantílica (RQ), uma metodologia baseada no método da minimização dos erros absolutos ponderados, que, diferentemente dos métodos de regressão linear clássica - os quais ajustam modelos para funções médias condicionais - possibilita o ajuste de modelos para funções de quantis condicionais, por meio de uma ponderação na minimização dos erros.

Deste modo, a RQ possibilita analisar o impacto das variáveis explicativas em diferentes pontos da distribuição condicional da variável resposta, permitindo obter maiores informações de localização, ao contrário de quando utiliza-se apenas a localização central da distribuição. Assim, é possível examinar uma localização na cauda inferior (por exemplo, o quantil 0,1) ou na cauda superior (por exemplo, o quantil 0,9) de acordo com o interesse do pesquisador (HAO e NAIMAN, 2007). Ademais, a RQ não requer a pressuposição de erros homocedásticos, ou seja, ela pode ser utilizada quando a distribuição dos erros é heterocedástica. Também é uma técnica robusta à presença de outliers e, outra vantagem é que possui uma representação de programação linear, o que facilita a estimação dos parâmetros (KOENKER e BASSET, 1978; SILVA e PORTO JUNIOR, 2006; MISSIO et al., 2009).

Diante das potencialidades da técnica, muitos estudos utilizando a regressão quantílica têm sido realizados, sendo estes com aplicações em diferentes áreas, podendo-se citar, os estudos de Silva e Porto Junior (2006) que analisaram teórica e empiricamente a suposta relação positiva existente entre desenvolvimento financeiro e crescimento econômico aplicando a técnica de regressão quantílica. Nascimento et al. (2012) que analisaram a influência de variáveis técnicas e econômicas sobre os índices de eficiência técnica de produtores de leite de Minas Gerais ao longo de pontos distintos da distribuição dos índices de eficiência. Hammoudeh et al. (2014) que utilizaram a regressão quantílica para investigar o impacto das mudanças nos preços do petróleo bruto, gás natural, carvão e os preços da eletricidade sobre a distribuição dos preços das licenças de emissão de CO_2 nos Estados Unidos e, Barroso et al. (2015), que desenvolveram e validaram uma metodologia de análise da adaptabilidade e da estabilidade fenotípica baseada em regressão quantílica.

2.1.2. Definição dos quantis

O quantil de ordem τ de uma população ou de uma amostra é o valor m tal que 100τ dos valores populacionais ou amostrais são inferiores a ele, com $0 < \tau < 1$. Mais formalmente, qualquer variável aleatória X pode ser caracterizada pela sua função de distribuição acumulada, $F(x) = P(X \leq x)$. Então, considerando a função inversa da distribuição acumulada no ponto τ, tem-se que $F^{-1}(\tau) = \inf\{x: F(x) \geq \tau\}$ é o quantil de ordem τ da variável aleatória X. Assim, a mediana pode ser definida por $F^{-1}\left(\frac{1}{2}\right)$ (KOENKER, 2005).

Segundo Hao e Naiman (2007), um quantil também pode ser considerado uma solução para um determinado problema de minimização. Seja Y uma variável aleatória com função de distribuição acumulada F. Sabe-se que é possível medir o quanto Y é distante de m, por meio do distância absoluta $|Y - m|$. Além disso, o valor esperado da distância absoluta pode ser dada por $E|Y - m|$. Assim, os autores demostraram que o valor de m que minimiza esse valor esperado é a mediana.

A solução da minimização é encontrar o ponto onde a derivada em relação a m é zero ou onde as derivadas laterais mudam de sinal. Este ponto é a mediana da distribuição. A prova deste resultado encontra-se no Apêndice A. Similarmente, para uma amostra também pode-se definir a distância média absoluta de m até os pontos amostrais por: $f(m) = \frac{1}{n}\sum_{i=1}^{n} |y_i - m|$. A função f assume um valor mínimo quando a derivada for igual a $-\frac{1}{n}$ para $m < y_i$ e $\frac{1}{n}$ para $m > y_i$. Porém, a função f não é diferenciável em $m = y_i$, assim, ela irá admitir derivada lateral de $-\frac{1}{n}$ na direção negativa e $\frac{1}{n}$ na direção positiva (HAO e NAIMAN, 2007).

Este resultado pode ser generalizado para qualquer quantil de interesse, pois para qualquer $\tau \in (0,1)$, a distância de Y para um dado q é medida pela distância absoluta ponderada entre Y e q. Assim, a distância de Y para um dado q é definida por (2) :

$$d_\tau(Y, q) = \begin{cases} (1-\tau)|Y - q|, & Y < q \\ \tau|Y - q|, & Y \geq q \end{cases} \qquad (2)$$

O valor de q que minimiza a distância média de Y, $E[d_\tau(Y,q)]$, é o τ -ésimo quantil. Da mesma maneira, em uma amostra, o τ-ésimo quantil é o valor de q que minimiza a distância média ponderada (HAO e NAIMAN, 2007):

$$\frac{1}{n}\sum_{i=1}^{n} d_\tau(y_i,q) = \frac{1-\tau}{n}\sum_{y_i<q}|y_i-q| + \frac{\tau}{n}\sum_{y_i>q}|y_i-q| \qquad (3)$$

em que $d_\tau(y_i,q)$ é a distância de y_i à q, n é o tamanho da amostra e y_i são os valores observados.

2.1.3. Modelos de Regressão Quantílica

Conforme já mencionado, os modelos de regressão quantílica oferecem uma visão mais completa da relação entre as variáveis estudadas, uma vez que possibilita observar a relação funcional em diferentes níveis da variável resposta. Este modelo, proposto por Koenker e Bassett (1978), pode ser representado conforme a equação (4):

$$y_i = \beta_0(\tau) + \beta_1(\tau)x_{i1} + \cdots + \beta_p(\tau)x_{ip} + e_i \qquad (4)$$

em que $\beta_0(\tau)$ é a constante da regressão no quantil τ, $\beta(\tau)$ são os coeficientes da regressão e e_i são os erros aleatórios independentes e identicamente distribuídos com quantil de ordem τ igual a zero. Assim, o quantil condicional de ordem τ de $X|Y$ é dado por:

$$Q_\tau(Y|X) = \beta_0(\tau) + \beta_1(\tau)x_1 + \cdots + \beta_p(\tau)x_p. \qquad (5)$$

Conforme é apresentado em Koenker (2005), esses modelos são capazes de incorporar uma possível heterocedasticidade, que seria detectada a partir da variação das estimativas dos parâmetros $\beta(\tau)$ para os diferentes quantis. Ademais, os erros padrão para a construção dos intervalos de

confiança para os parâmetros do modelo de RQ podem ser obtidos pelo método bootstrap.

Na regressão quantílica podem ser ajustadas retas para cada quantil de interesse, fornecendo informações sobre mudanças na distribuição da variável resposta, o que facilitaria a interpretação dos resultados para um conjunto de dados com assimetria, já que através da RQ é possível traçar a relação em regiões centrais da distribuição, por meio da mediana, e nas caudas da distribuição condicional, conforme o interesse do pesquisador.

2.1.4. Estimação dos Parâmetros

Uma diferença significativa entre o estimador da RQ e da regressão linear simples é que na RQ a distância de pontos observados à reta ajustada é medida minimizando a média ponderada da soma das distâncias verticais, onde o peso é $1-\tau$ para pontos abaixo da reta e τ para pontos acima da reta (HAO e NAIMAN, 2007).

Cada escolha para o peso τ, dá origem a uma função ajustada do quantil condicional. Assim, é necessário encontrar um estimador com a propriedade desejada para cada possível τ.

Dessa forma, busca-se encontrar os estimadores de $\beta_0(\tau), \beta_1(\tau), \cdots, \beta_p(\tau)$ que minimizem a seguinte equação:

$$\begin{aligned}\sum_{i=1}^{n} d_\tau\,(y_i, \hat{y}_i) = \sum_{y_i \geq \hat{y}_i} \tau \left| y_i - [\beta_0(\tau) + \beta_1(\tau)x_{i1} + \cdots + \beta_p(\tau)x_{ip}] \right| \\ + \sum_{y_i < \hat{y}_i} (1-\tau) \left| y_i - [\beta_0(\tau) + \beta_1(\tau)x_{i1} + \cdots + \beta_p(\tau)x_{ip}] \right|\end{aligned} \tag{6}$$

em que d_τ é a distância entre y_i e $\hat{y_i}$, e ainda, $\hat{y}_i = \hat{\beta}_0(\tau) + \hat{\beta}_1(\tau)x_{i1} + \cdots + \hat{\beta}_p(\tau)x_{ip}$. Minimizando esta equação para cada τ tem-se como resultados as estimativas dos coeficientes de regressão dos diferentes quantis de interesse. Detalhes do algoritmo de minimização da soma dos erros absolutos ponderados encontram-se no Apêndice B.

2.1.5. Regressão Quantílica Regularizada

A Regressão Quantílica Regularizada (RQR) é uma metodologia estatística que ajusta modelos de predição em diferentes níveis da variável de interesse, mas que impõe restrição no processo de estimação. Dessa forma, através da restrição imposta, o método combina a seleção de variáveis e regularização via encurtamento dos coeficientes de regressão (*shrinkage*) (LI e ZHU, 2008). De modo geral, o ajuste do modelo regularizado para encontrar a $100\tau\%$ função quantílica consiste na obtenção de estimadores de coeficientes de regressão que resolvam o seguinte problema de otimização:

$$min\left\{\sum_{i=1}^{n}\rho_\tau\left(y_i-\mu-\sum_{j=1}^{p}x_{ij}g_j\right)+\lambda\sum_{j=1}^{p}|g_j|\right\} \qquad (7)$$

em que $\sum_{j=1}^{p}|g_j|$ é a soma dos valores absolutos dos coeficientes de regressão, λ é o parâmetro que controla a força da regularização e $\rho_\tau(\,)$, denotada função *check* por Koenker e Bassett (1978), é definida por:

$$\rho_\tau\left(y_i-\mu-\sum_{j=1}^{p}x_{ij}g_j\right)=\begin{cases}\tau\left(y_i-\mu-\sum_{j=1}^{p}x_{ij}g_j\right), se\ \ y_i-\mu-\sum_{j=1}^{p}x_{ij}g_j\ >0\\ (1-\tau)\left(y_i-\mu-\sum_{j=1}^{p}x_{ij}g_j\right), \text{caso contrário}\end{cases}$$

em que $\tau \in (0,1)$ indica o quantil de interesse.

A RQR pode ser utilizada em situações que ocorrem problemas de multicolinearidade e de dimensionalidade decorrentes da quantidade, geralmente superior, de parâmetros a serem estimados em relação ao número de observações. Estudos sobre a RQR tem sido cada vez mais frequentes e merecem destaque os estudos realizados por He et al. (2015) que utilizaram a

RQR para identificar as características genéticas que influenciam caracteres quantitativos a fim de descobrir a causa de doenças e Li et al. (2004) que estudaram a regressão quantílica regularizada sob uma perspectiva bayesiana.

2.1.6. Análise da Qualidade de Ajuste da Regressão Quantílica

Na análise de modelos de regressão, uma medida bastante utilizada para verificar a qualidade de ajuste do modelo é o coeficiente de determinação, (R^2), que pode ser calculado pela seguinte expressão:

$$R^2 = \frac{SQT - SQE}{SQT} \qquad (8)$$

em que $SQT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ é a soma de quadrados totais e $SQE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ a soma de quadrados dos resíduos.

Este coeficiente pode ser interpretado como a proporção da variabilidade da variável resposta explicada pelas variáveis independentes do modelo. Ele varia entre 0 e 1 $(0 \leq R^2 \leq 1)$, e quanto maior o seu valor melhor é o ajuste do modelo (HAO e NAIMAN, 2007).

Koenker e Machado (1999) sugeriram medir a qualidade do ajuste em modelos de Regressão Quantílica comparando a soma das distâncias ponderadas de um modelo completo com a soma das distâncias de um modelo reduzido, que inclui apenas o intercepto. Desta forma, seja $V^1(\tau)$, a soma das distâncias ponderadas para o τ-ésimo modelo de regressão quantilica completo, e $V^0(\tau)$ a soma da distância ponderada para um modelo reduzido. Por exemplo, utilizando um modelo com uma variável independente, temos:

$$\begin{aligned} V^1(\tau) &= \sum_{i=1}^{n} d_\tau\,(y_i, \hat{y}_i) \\ &= \sum_{y_i \geq \hat{y}_i} \tau |y_i - [\beta_0(\tau) + \beta_1(\tau)x_{i1}]| + \sum_{y_i < \hat{y}_i} (1-\tau)|y_i - [\beta_0(\tau) + \beta_1(\tau)x_{i1}]| \end{aligned} \qquad (9)$$

e

$$V^0(\tau) = \sum_{i=1}^{n} d_\tau\,(y_i, \widehat{Q}(\tau)) = \sum_{y_i \geq \bar{y}} \tau \left|y_i - \widehat{Q}(\tau)\right| + \sum_{y_i < \bar{y}} (1-\tau)\left|y_i - \widehat{Q}(\tau)\right| \qquad (10)$$

Para o modelo que possui somente o termo constante, a constante ajustada é o τ-ésimo quantil amostral $\hat{Q}(\tau)$ para a amostra $y_1, y_2, \cdots, y_n$. Assim, a qualidade do ajuste é então medida por:

$$R(\tau) = 1 - \frac{V^1(\tau)}{V^0(\tau)}, \qquad (11)$$

em que o $V^0(\tau)$ e $V^1(\tau)$ são valores não - negativos e $R(\tau)$ é, no máximo, igual a 1.

Além disso, como a soma das distâncias ponderadas é minimizada para um modelo completo ajustado, $V^1(\tau) \leq V^0(\tau)$ e $R(\tau) \geq 0$. Assim, $R(\tau) \in [0,1]$ e quanto maior o $R(\tau)$ melhor é o ajuste do modelo.

O $R(\tau)$ constitui uma medida local de qualidade de ajuste para um quantil particular, e não uma medida global da qualidade de ajuste ao longo de toda a distribuição condicional. O $R(\tau)$ conforme definido acima, permite a comparação de modelos ajustados com qualquer número de covariáveis. Esta é uma forma restrita de comparação de qualidade de ajuste introduzida por Koenker e Machado (1999) para os modelos aninhados.

2.2. Seleção Genômica Ampla

A genética molecular tem beneficiado o melhoramento animal e vegetal através da utilização direta de informações do DNA. Dessa forma, os programas de melhoramento buscam identificar os locos gênicos que estão envolvidos na determinação de características fenotípicas de interesse e incorporam essas informações às metodologias de melhoramento, visando obter indivíduos geneticamente superiores.

A Seleção genômica ampla *(Genome Wide Selection* - GWS*)* foi proposta por Meuwissen et al. (2001) e consiste em utilizar informações de marcadores para estimar seus efeitos sobre características fenotípicas de interesse, visando predizer o mérito genético para posterior seleção de indivíduos. Dessa forma, a GWS possibilita aumentar a eficiência da seleção de genótipos e a rapidez na obtenção de ganhos genéticos (RESENDE, 2012).

As novas tecnologias no processo de genotipagem permitiram o desenvolvimento de novas classes de marcadores moleculares, como os SNPs (*Single Nucleotide Polymorphisms*), que são variações da seqüência de DNA que ocorrem quando somente uma base na seqüência do genoma é alterada (BROOKES,1999). Esses marcadores são as mais abundantes variações genéticas encontradas nos genomas, sendo mais vantajosos em relação a outros marcadores em razão de sua baixa taxa de mutação e reduzido custo de genotipagem (RESENDE, 2012). Os SNPs ocorrem com grande frequência no genoma e, por esse motivo são potencialmente úteis para a associação de mapeamento de características de interesse.

Os efeitos desses marcadores são estimados por uma metodologia estatística, e são utilizados na predição de valores genéticos genômicos (GBV's) e posterior seleção de indivíduos. A seleção pode ser realizada logo após o nascimento, uma vez que através dos marcadores são identificados os alelos que estão associados a uma característica de interesse, acelerando o processo de melhoramento genético (RESENDE, 2012).

Dentre as metodologias estatísticas utilizadas em GWS, as técnicas que envolvem inferência bayesiana têm se destacado muito, devido, principalmente, à evolução dos recursos computacionais.

2.3. Inferência Bayesiana

Na abordagem bayesiana os parâmetros de interesse são tratados como variáveis aleatórias e não mais como constantes fixas e desconhecidas. Além disso, essa abordagem considera o fato de que o pesquisador tem algum conhecimento sobre o parâmetro de interesse θ antes de se observar os valores da amostra (informação à priori), e que tal conhecimento pode ser incorporado ao estudo através de uma função densidade $p(\theta)$ (GAMERMAN e LOPES, 2006). Ademais, para a análise Bayesiana também é fundamental conhecer a função de verossimilhança ou densidade conjunta dos dados $f(y|\theta)$, que é obtida após se observar as informações contidas nos mesmos.

A teoria bayesiana foi fundamentada nos trabalhos desenvolvidos por Thomas Bayes em 1761, que só foram publicados em 1763, e por isso os métodos de inferência bayesiana baseiam-se no Teorema de Bayes, apresentado a seguir.

Teorema: Suponha que os eventos $A_1, A_2, \cdots, A_n$ formam uma partição do espaço amostral Ω e que todos tenham probabilidade positiva. Seja B um evento qualquer com $P(B) > 0$. Então para $j = 1,2, \ldots, n$, tem-se:

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(B|A_j)P(A_j)}. \qquad (12)$$

Assim, considerando uma amostra aleatória $(y_1, y_2, \cdots, y_n)$, cujas informações são utilizadas na análise bayesiana por meio da função de verossimilhança $f(y|\theta)$, e a densidade a priori $p(\theta)$, que contém a distribuição de probabilidade de θ antes de se observar os dados, a distribuição a posteriori dos dados, $p(\theta|y)$, é obtida por meio do Teorema de Bayes (12), que resulta em:

$$p(\theta|y) = \frac{f(y|\theta)_{p(\theta)}}{p(y)} \propto f(y|\theta)p(\theta), \qquad (13)$$

em que $p(y) = \int_{\Theta} f(y|\theta)\, p(\theta) d\theta$ em que Θ é o espaço paramétrico de θ . Como $p(y)$ não depende de θ, este termo é apenas uma constante normalizadora.

Toda a inferência a respeito do parâmetro θ é feita por meio da distribuição a posteriori, uma vez que esta contém toda a informação probabilística a respeito deste parâmetro.

A Estatística Bayesiana requer o uso de métodos computacionais de Monte Carlo via Cadeias de Markov (MCMC) dentre os quais se destacam o amostrador de Gibbs e algoritmo de Metrópolis-Hastings. Isso ocorre devido à complexidade dos problemas matemáticos que se pode ter em algumas situações, dependendo de qual é a distribuição a priori ou a função de verossimilhança.

O método MCMC é fundamentado em simulações iterativas, que tem como objetivo obter amostras aleatórias da distribuição conjunta dos parâmetros, isto é, gerar amostras aleatórias da distribuição à posteriori. Para maiores detalhes ver (GAMERMAN e LOPES, 2006).

2.4. Métodos Bayesianos na GWS

Os métodos bayesianos foram introduzidos na GWS por Meuwissen et al. (2001). Esses métodos são utilizados para a predição do mérito genético baseado em informações genômicas. Dentre estas metodologias destacam-se Bayes A e Bayes B (MEUWISSEN et al., 2001), Bayes Cπ (HABIER et al., 2011), GBLUP (Genomic *Best Linear Unbiased Predictor*) (VAN RADEN, 2008), LS (*Least Squares*) (LANDE e THOMPSON, 1990), RR-BLUP (WHITTAKER et al., 2000; MEUWISSEN et al., 2001), Bayesian LASSO (PARK e CASELLA, 2008; DE LOS CAMPOS et al., 2009).

O modelo geral para a seleção genômica ampla sugerido por Meuwissen et al. (2001) é descrito em (14):

$$y_i = \mathbf{1}\mu + \sum_{j=1}^{p} x_{ij} g_j + \boldsymbol{e_i}, \qquad (14)$$

em que $\boldsymbol{y}$ é o vetor de fenótipos, $\mathbf{1}$ é o vetor de mesma dimensão de $\boldsymbol{y}$ com todas as entradas iguais a 1, μ é a média da característica estudada, g_j é o efeito marcador SNP $(j = 1,2, \dots, p)$, x_{ij} são os elementos da matriz de incidência de cada marcador j, parametrizada em 0,1 e 2, e $\boldsymbol{e}$ é o vetor de erros aleatórios do modelo.

A diferença entre os métodos bayesianos na GWS está nas distribuições a priori que são atribuídas aos $g_j's$ (efeitos dos SNP's) e das suas respectivas variâncias ($\sigma^2_{g_j}$). Neste estudo, será descrito apenas o método LASSO Bayesiano (BLASSO) que posteriormente será utilizado para realizar comparações com o método da Regressão Quantílica Regularizada.

2.4.1. Lasso Bayesiano

A regressão bayesiana (MEUSSIWEN et al., 2001) tem sido muito utilizada para solucionar os problemas de multicolinearidade, em que alguns marcadores são altamente correlacionados, e a alta dimensionalidade, no qual o número de marcadores (covariáveis) é muito maior que o número de observações, pois algumas distribuições *a priori* impõem regularização no ajuste do modelo, sob forma de encurtamento dos coeficientes de regressão (*shrinkage*).

Dentre os métodos utilizados para este fim, destaca-se o método de regressão Lasso (*Least Absolute Shrinkageand Selection Operator*), proposto por Tibshirani (1996), que tem se tornado uma alternativa ao método dos mínimos quadrados ordinários (MQO) com o objetivo de melhorar a precisão de predição e a interpretação do modelo, uma vez que combina seleção de variáveis e a regularização via encurtamento dos coeficientes de regressão.

O Lasso pode ser utilizado para encontrar os estimadores dos coeficientes de regressão do modelo (14) que resolvam o seguinte problema de otimização (DE LOS CAMPOS et al., 2009) :

$$min\left\{\sum_{i=1}^{n}\left(y_i - \mu - \sum_{j=1}^{p} x_{ij}g_j\right)^2 + \lambda\sum_{j=1}^{p}|g_j|\right\} \qquad (15)$$

em que $\sum_{j=1}^{p}|g_j|$ é a soma dos valores absolutos dos coeficientes de regressão e λ é o parâmetro de encurtamento que controla a força da regularização, de modo que quando $\lambda = 0$, não há regularização.

A versão Bayesiana da regressão Lasso (PARK e CASELLA, 2008), conhecida como Lasso bayesiano ou BLASSO, foi proposta por de Los Campos et al. (2009) para ser aplicada na seleção genômica. Na implementação Bayesiana do Lasso, a prática desta regularização envolve um encurtamento mais forte permitindo que as estimativas dos coeficientes de regressão sejam aproximadamente zero. Ademais, impõe-se que os coeficientes de regressão (efeitos dos marcadores) tenham distribuições a priori de Laplace idênticas e independentes. Dessa forma admite-se como distribuição *a priori* dos p coeficientes de regressão é um produto de densidades exponenciais duplas: $p(g|\lambda) = \prod_{j=1}^{p} \frac{\lambda}{2\sigma_e} exp\left(\frac{-\lambda|g_j|}{\sigma_e}\right)$.

A densidade exponencial dupla apresenta maior massa de densidade no valor 0 e caudas mais robustas, realizando maior encurtamento sobre coeficientes de regressão próximos à 0 e menor encurtamento sobre aqueles distantes de zero (TIBSHIRANI, 1996). Conforme pode ser observado na Figura 2, quando comparada a densidade normal.

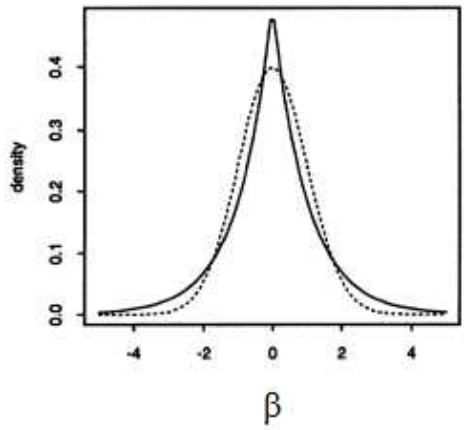


Figura 1: Densidade da distribuição exponencial dupla (——) e densidade da distribuição normal(------).

Fonte: Tibshirani (1996).

A distribuição exponencial dupla pode ser expressa como uma mistura na escala de distribuições normais com variâncias que seguem distribuições exponenciais independentes (ANDREWS e MALLOWS, 1974). Dessa forma, ao construir a distribuição a priori conjunta dos parâmetros, a distribuição assumida pelos coeficientes de regressão regularizada por Lasso (DE LOS CAMPOS, 2009) será: $\prod_{j=1}^{p} N(g_{j1}|0, \sigma_e^2\tau_j^2)$, o que resulta em variância específica para cada efeito SNP j, igual a $\sigma_e^2\tau_j^2$. Ademais, a distribuição a priori do parâmetro τ_j^2 é dada por: $\prod_{j=1}^{p} exp(\tau_j^2|\lambda)$, em que o parâmetro λ influencia o ajuste dos coeficientes da regressão. A informação a priori para λ é dada pelas distribuições, gama ou beta, com hiperparâmetros conhecidos.

2.5. Valores genéticos genômicos, herdabilidade, capacidade preditiva e acurácia

Valores genéticos genômicos (GBV's) são valores atribuídos aos indivíduos e, que se refere ao seu potencial genético. Especificamente, são estimativas do valor real fenotípico, em função do genótipo e dos efeitos de marcadores. Os GBV's são obtidos pela seguinte expressão:

$$G\hat{B}V_i = \hat{y}_i(\tau) = \sum_{j}^{p} x_{ij}\hat{g}_j \qquad (16)$$

em que p é o número de marcadores, x_{ij} são os elementos da matriz de genótipos e $\hat{g}_j$ é o vetor que contém as estimativas dos efeitos dos SNP's.

A capacidade preditiva corresponde ao coeficiente de correlação entre os valores genéticos genômicos estimados e os valores fenotípicos observados, sendo representada por $r_{y,\hat{y}}$.

A herdabilidade é representada por h^2 e sua principal função é medir quanto da variação fenotípica é atribuída à variação genotípica (FALCONER e MACKAY, 1996). A herdabilidade pode ser definida por $h^2 = \frac{v_g}{v_f}$, istó é, a razão entre a variância genotípica (v_g) e a variância fenotípica (v_f).

A acurácia, é uma medida para avaliar a qualidade do ajuste do modelo, ou seja, ela mede o quanto a estimativa obtida é próxima do valor real. A acurácia varia entre 0 e 1 e quanto maior for o seu valor melhor será o ajuste do modelo. Essa medida é definida pela razão entre capacidade preditiva e a raiz da herdabilidade, e é representada por (LEGARRA, 2008):

$$r_{q,\hat{q}} = \frac{r_{y,\hat{y}}}{\sqrt{h^2}} \qquad (17)$$

em que $r_{y,\hat{y}}$ é a capacidade preditiva do modelo, dada por $r_{y,\hat{y}} = \frac{cov(y,\hat{y})}{\sqrt{Var(y)Var(\hat{y})}}$ e h^2 é a herdabilidade do caráter.

2.6. Validação Cruzada

O método de validação cruzada consiste em dividir o conjunto de dados originais em n subconjuntos, e realizar n análises, de modo que em cada uma das análises um dos subconjuntos seja retirado e utilizado para validar a análise que foi realizada. Dessa forma, os valores preditos pelas equações estimadas em cada uma das análise podem ser comparados com os valores observados retirados.

Nos estudos de seleção genômica ampla, três populações podem ser definidas: população de estimação, validação e seleção (RESENDE, 2012). A população de estimação ou treinamento é utilizada para estimar os efeitos dos marcadores, e assim obter as equações de predição dos valores genéticos genômicos (GBV's). Nesta população é considerado um grande número de marcadores avaliados em um número moderado de indivíduos, que devem ter os seus fenótipos avaliados para as várias características de interesse (RESENDE, 2012).

A população de validação é utilizada para testar e verificar as acurácias das equações de predição que foram estimadas. Esse conjunto de dados pode ser menor ou igual que a população de estimação.

Para avaliar a capacidade de generalização do modelo preditivo e evitar a superestimação das medidas que avaliam o modelo, uma vez que as mesmas podem ser estimadas na população de estimação e, neste caso, o modelo seria estimado e validado com base no mesmo conjunto de dados, a validação cruzada torna-se de grande importância.

2.7. Teste D'Agostino –Pearson

O teste D'Agostino–Pearson, também conhecido como o teste k^2 de D'Agostino-Pearson é um teste abrangente que combina os testes de assimetria e de curtose .

Considere uma amostra de tamanho n, $x_1, x_2, \dots x_n$ em que $\sqrt{b_1}$ e b_2 são os coeficientes de assimetria e a curtose da amostra. $m'_k s$ são os k-ésimos momentos centrais da amostra. Podemos definir os coeficientes de assimetria e curtose de uma amostra como:

$$\sqrt{b_1} = \frac{m_3}{m_2^{3/2}} \quad \text{e} \quad b_2 = \frac{m_4}{m_2^2} \tag{18}$$

em que $m_k = \frac{\sum(X_i - \bar{X})^k}{n}$ e $\bar{X}$ é a média da amostra.

2.7.1. Testes de Assimetria e Curtose

D'Agostino et al. (1990) descreve um teste de normalidade com base nos coeficientes de assimetria, $(\sqrt{b_1})$ e curtose (b_2). Como a distribuição normal é simétrica, $\sqrt{b_1}$ é igual a zero para dados normais e o valor de b_2 é igual a 3. Assim, o teste foi desenvolvido para determinar se o valor de $\sqrt{b_1}$ é significativamente diferente de zero e o valor de b_2 é significativamente diferente de 3, ou seja, se os dados são não- normais.

A assimetria e a curtose da amostra são assintoticamente normais. Entretanto, a taxa de convergência para o limite da distribuição é muito lenta. Para solucionar esse problema, D'Agostino (1970) propôs uma transformação para $\sqrt{b_1}$ de modo que a sua distribuição seja o mais próximo possível da

normal padrão. A transformação sugerida para a assimetria da amostra é dada por:

$$Z(\sqrt{b_1}) = \delta ln\left({}^{Y}/_{\alpha} + \left\{ \left({}^{Y}/_{\alpha} \right)^2 + 1 \right\}^{1/2} \right) \qquad (19)$$

em que: $Y = \sqrt{b_1}\left\{ \frac{(n+1)(n+3)}{6(n-2)} \right\}^{1/2}$,

$$W^2 = -1 + \left\{ 2\left(\beta_2 \left(\sqrt{b_1} \right) - 1 \right) \right\}^{1/2}$$

e a constante δ é igual a $\delta = \frac{1}{\sqrt{lnW}}$.

Analogamente, Anscombe e Glynn (1983) sugeriram uma transformação para o coeficiente de curtose b_2:

$$Z(b_2) = \frac{\left(\left(1 - \frac{2}{9A}\right) - \left[\frac{1 - {}^{2}/_{A}}{1 + x\sqrt{2/(A-4)}} \right]^{1/3} \right)}{\sqrt{{}^{2}/_{(9A)}}} \qquad (20)$$

$$com \quad A = 6 + \frac{8}{\sqrt{\beta_1(b_2)}}\left[\frac{2}{\sqrt{\beta_1(b_2)}} + \sqrt{\left(1 + \frac{4}{\beta_1(b_2)}\right)} \right] \quad \text{e} \quad x = \frac{(b_2 - E(b_2))}{\sqrt{var(b_2)}} \quad .$$

O terceiro momento padronizado de b_2 é dado por:

$$\sqrt{\beta_1(b_2)} = \frac{6(n^2 - 5n + 2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}} \qquad (21)$$

A média e a variância de b_2 são iguais:

$$E(b_2) = \frac{3(n-1)}{n+1} \quad \text{e} \quad var(b_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)} \qquad (22)$$

2.7.2. Teste abrangente

D'Agostino e Pearson (1973) apresentaram uma estatística que combina $\sqrt{b_1}$ e b_2 para produzir um teste abrangente de normalidade. Pelo teste, é possível detectar desvios da normalidade, devido à assimetria ou a curtose. A estatística de teste é dada por:

$$K^2 = Z^2\left(\sqrt{b_1}\right) + Z^2(b_2) \qquad (23)$$

em que $Z^2\left(\sqrt{b_1}\right)$ e $Z^2(b_2)$ são as aproximações normais para $\sqrt{b_1}$ e b_2. A estatística K^2 possui aproximadamente uma distribuição qui-quadrado, com 2 graus de liberdade quando a população é normalmente distribuída (D'AGOSTINO et al. 1990).

2.8 Coeficiente de correlação de postos de Spearman

O coeficiente de correlação de postos de Spearman, representado pela letra grega ρ (rho), é uma medida de correlação não-paramétrica, isto é, não exige nenhum pressuposto de distribuição normal . Ele pode ser utilizado para variáveis medidas em nível ordinal, de modo que os elementos em estudo formem duas séries ordenadas (SIEGEL, 1975). Além disso, segundo GUILFORD (1950) este coeficiente é conveniente para amostras pequenas. O ρ é dado por:

$$\rho = 1 - \frac{6\sum {d_i}^2}{(n^3 - n)}$$

em que:

d_i= a diferença entre cada posto de valor correspondentes de x e y, e
n= o número dos pares dos valores.

O coeficiente de correlação de postos de Spearman nada mais é que o coeficiente de correlação linear de Pearson aplicado aos postos dos dados, obtidos de maneira independente para cada variável. Ademais, ao contrário do coeficiente de correlação de Pearson, o coeficiente de correlação de Spearman não requer a suposição que a relação entre as variáveis seja linear e também é menos sensível a "outliers". Ele varia de -1 (maior correlação negativa) e 1 (maior correlação positiva) e esses valores não são raros de ocorrer na prática.

REFERÊNCIAS BIBLIOGRÁFICAS

ABPA: Associação Brasileira de Proteína Animal. Disponível em: <http://www.abipecs.org.br/pt/estatisticas/mundial/producao-2.html>. Acesso em: Mar. 2015.

ANDREWS, D. F.; MALLOWS, C. L. Scale mixtures of normal distributions. **Journal of the Royal Statistical Society**, London, Ser. B, v.36, p. 99-102, 1974.

AZEVEDO, C. F.; RESENDE, M. D. V.; SILVA, F. F.; Lopes, P. S.; GUIMARÃES, S. E. F. Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. **Pesquisa Agropecuária Brasileira,** v. 48, p. 619-626, 2013.

BARROSO, L. M. A.; NASCIMENTO, M.; NASCIMENTO, A. C. C.; SILVA, F. F.; CRUZ, C.D.; LEONARDO LOPES BHERING, L. L.; FERREIRA, R.P.; Metodologia para análise de adaptabilidade e estabilidade por meio de regressão quantílica. **Pesquisa Agropecuária brasileira**, Brasília, v.50, p. 290-297, 2015.

BROOKES, A. J. The essence of SNP. **Gene**, Amsterdam, v. 234, p. 177-186, 1999.

D'AGOSTINO, R. B.; BELANGER, A.; D'AGOSTINO JR, R. B. A Suggestion for Using Powerful and Informative Tests of Normality. **The American Statistician**, November , v. 44, n. 4, p. 316-321, 1990.

D'AGOSTINO, R. B.; PEARSON, E. S. Tests for Departure from Normality. **Biometrika,** v. 60, p. 613–22, 1973.

DE LOS CAMPOS, G.; NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J. M.. Predicting quantitative traits with regression models for dense molecular makers. **Genetics**, Austin, v. 182, p. 375-385, 2009.

EVERTON, N. S.; SABINO, S. P. J. Sistema financeiro e crescimento econômico: uma aplicação de regressão quantílica*. **Economia Aplicada**, v. 10, p. 425-442, 2006.

FALCONER, D. S.; MACKAY, T. F.C. **Introduction to quantitative genetics**. 4.ed. New York: Longman, 464p., 1996.

GAMERMAN, D.; LOPES, H. L. **Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference Second Edition(Chapman & Hall/CRC Texts in Statistical Science)**. 2. Ed. Hardocover, 2006.

GIANOLA, D.; PEREZ-ENCISO, M.; TORO, M. A. On marker-assisted prediction of genetic value: beyond the ridge. **Genetics,** v.163, p. 347-365, 2003.

GUILFORD, J. P. **Fundamental statistics in psychology and education**. 4.ed. New York: McGraw-hill Book, 605p., 1950.

HABIER, D.; FERNANDO, R.L.; KIZILKAYA, K.; GARRICK, D.J. Extension of the bayesian alphabet for genomic selection. **BMC Bioinformatics**, v.12, p.186, 2011.

HAMMOUDEH, S.; NGUYEN, D. K.; SOUSA, R. M. Energy prices and CO2 emission allowance prices: A quantile regression approach. **Energy Policy**, v. 70, p. 201-206, 2014.

HAO, L.; NAIMAN, D. Q. **Quantile regression**. Sage publications. 126p., 2007.

KOENKER, R.; BASSETT, G. Regression quantiles. **Econometrica,** v. 46, p. 33-50, 1978.

KOENKER, R.; MACHADO, J. Goodness of Fit and Related Inference Processes for Quantile Regression. **Jornal of the American Statistical Association**, v.94, p.1296-1310, 1999.

KOENKER, R. **Quantile Regression**, Cambridge University Press, 2005.

LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics**, v. 124, p. 743-756, 1990.

LEGARRA, A.; MISZTAL, I. Computing strategies in genome-wide selection. **Jornal of Dairy Science**, v.91, n.1, p.360-366, 2008.

LI, Q.; XI, R.; LIN, N. Bayesian Regularized Quantile Regression, **Bayesian Analysis**, p. 1–26, 2004.

LI, Y.; ZHU, J. L1-Norm Quantile Regression. **Journal of Computational and Graphical Statistics**, v.17, p. 163–185, 2008.

MAPA. Ministério da Agricultura, Pecuária e Abastecimento. 2015. Disponível em: <http://www.agricultura.gov.br/animal/especies/suinos>. Acesso em: Jan. 2015.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction os total genetic value using Genome-Wide dense marker maps. **Genetics Society of America**, v. 157, p. 1819-1829, 2001.

MISSIO, F. .J.; JAYME JR, F. G.; OLIVEIRA, A. M. H. C. **Desenvolvimento financeiro e crescimento econômico: Teoria e evidência empírica para os estados brasileiros (1995-2004).** Belo Horizonte: UFMG/ Cedeplar. 33p., 2009.

MONTGOMERY, D.; PECK, E.; VINING, C. **Introduction to Linear Regression Analysis**, Wiley, 2001.

NASCIMENTO, A. C. C.; LIMA, J. E.; BRAGA, M. J.; NASCIMENTO, M.; GOMES, A. P.; Eficiência técnica da atividade leiteira em Minas Gerais: Uma aplicação de regressão quantílica. **Revista Brasileira de Zootecnia**, v.41, p. 783-789, 2012.

PAIXÃO, D. M.; CARNEIRO, P. L. S.; PAIVA, S. R.; SOUSA, K. R. S.; VERARDO, L. L.; BRACCINI NETO, J. ; PINTO, A. P. G.; HIDALGO, A. M.; NASCIMENTO, C. S.; PÉRISSÉ, I. V.; LOPES, P. S.; GUIMARÃES, S. E. F. Mapeamento de QTL nos cromossomos 1, 2, 3, 12, 14, 15 e X em suínos: características de carcaça e qualidade de carne. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 64, p. 974-982, 2012.

PARK, T.; CASELLA, G. The Bayesian Lasso. **Journal of the American Statistical Association**,103(482), p. 681–686, 2008.

QIANCHUAN HE.; LINGLONG K.; YANHUA, W.; SIJIAN W.; TIMOTHY A. C.; ERIC H. Regularized quantile regression under heterogeneous sparsity with application to quantitative genetic traits. **Computational Statistics and Data Analysis**, 2015.

RESENDE, M. D. V.; SILVA, F. F.; VIANA, J. M.; PETTERNELLI, L. A.; RESENDE JR, M. F. R.; VALLE, P. **Computação da Seleção Genômica Ampla**. Colombo: EMBRAPA Florestas, 79p. 2010.

RESENDE, M. D.; SILVA, F. F.; LOPES, P. S.; AZEVEDO, C. F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial**. Viçosa: Universidade Federal de Viçosa/Departamento de Estatística, 291p., 2012. Disponível em:<http://www.det.ufv.br/ppestbio/corpo_docente.php>. Acesso em Fev. 2015.

SIEGEL, S. **Estatística Não-Paramétrica: Para as Ciências do Comportamento.** São Paulo: McGraw-Hill do Brasil, 350 p. 1975.

SILVA, E. N.; PORTO JÚNIOR, S. S. Sistema financeiro e crescimento econômico: Uma aplicação de regressão quantílica**. Economia Aplicada,** Ribeirão Preto, v. 10, n.3, p. 425-442, 2006.

TEIXEIRA, F.R.F.; NASCIMENTO, M.; NASCIMENTO, A. C. C.; PAIXÃO, D. M.; AZEVEDO, C. F.; SILVA, F. F.; CRUZ, C. D.; LOPES, P. S.; GUIMARÃES, S. E. F.; determinação de fatores em características de suíno. **Revista Brasileira de Biometria**, São Paulo, v.33, n.2, p.130-138, 2015.

TIBSHIRANI, R. Regression shrinkage and selection via the LASSO, J. R. **Stat. Soc. B 58,** p. 267-288, 1996.

VAN RADEN, P.M. Efficient methods to compute genomic predictions. **Jornal of Dairy Science**, v.91, n.11, p. 4414-4423, 2008.

WHITTAKER, J.C.; THOMPSON, R.; DENHAM, M.C. Marker-assisted selection using ridge regression. **Genetics Research**, v. 75, p. 249-252, 2000.

CAPÍTULO 1

Seleção Genômica para características assimétricas em suínos via regressão quantílica

Resumo: Nos programas de melhoramento, o uso de informações dos marcadores SNPs (*Single Nucleotide Polymorphisms*) para predição do mérito genético ou para identificar regiões genômicas responsáveis por fenótipos de interesse, tem se tornado uma importante ferramenta. Dentre os métodos comumente utilizados para este fim, destaca-se o método bayesiano BLASSO que, como a maioria das metodologias apresentadas na literatura, estima apenas o efeito dos marcadores em termos do valor médio da característica de interesse. Porém, em algumas situações, a característica de interesse pode não ter distribuição simétrica, ou mesmo, o interesse recai na seleção de animais que apresentam maiores valores para a característica em estudo. Desta forma, uma modelagem considerando a estimação dos efeitos de marcadores em diferentes níveis dessa variável pode ser mais adequada do que aquela em que se considera apenas o efeito médio. Diante do exposto, este trabalho propõe a utilização da Regressão Quantílica Regularizada, na predição do mérito genético de suínos em diferentes níveis de características de carcaça. Além disso, objetiva-se verificar se os resultados obtidos por esta metodologia, em termos da classificação dos animais para seleção, diferem daqueles advindos do método usual, BLASSO. Neste estudo, consideraram-se dados de uma população F_2 de suínos Piau x Comercial e foram ajustadas funções quantílicas para $(\tau = 0{,}05 \text{ a } \tau = 0{,}95)$. Como resultados, tem-se que modelos de regressão quantílica apresentaram, em algumas situações, valores de acurácia maiores que aqueles obtidos pelo BLASSO. Além disso, os resultados dos métodos para as variáveis ETUC e EBACON foram semelhantes e para RCARC foram diferentes no que se refere à classificação dos animais. Quanto ao padrão dos efeitos de marcadores, os resultados encontrados pelos métodos para as variáveis foram diferentes.

Palavras-chave: melhoramento animal, regressão quantílica regularizada, blasso.

1. Introdução

No Brasil, o mercado de carne suína apresenta grande destaque, com uma produção de 110.606 mil toneladas no ano de 2014. Este total representa cerca de 3% da produção mundial, o que classifica o Brasil como quarto maior produtor deste tipo de carne (Associação Brasileira de Proteína Animal - ABPA). Segundo o Ministério da Agricultura, Pecuária e Abastecimento (MAPA), os investimentos em pesquisa e na evolução genética da espécie, realizados nos últimos 20 anos têm sido responsáveis pelo grande aumento na produção e na qualidade da carne produzida, destacando a importância dos programas de melhoramento (MAPA, 2015).

Dentre as diversas características avaliadas nos programas de melhoramento de suínos, aquelas associadas à carcaça são importantes para o desenvolvimento da suinocultura, por serem exigências do mercado consumidor (ZANGERONIMO et al., 2009), o que faz com que estas sejam, frequentemente, consideradas como critérios de seleção.

Entretanto, as distribuições de muitas dessas características apresentam comportamento assimétrico e, nestes casos, o uso de metodologias usuais de Seleção Genômica Ampla (*Genome Wide Selection* – GWS), as quais se baseiam em esperanças condicionais, além de impossibilitar predições em toda a distribuição dos valores fenotípicos, podem sub ou superestimar os efeitos de marcadores. Isso ocorre, uma vez que, nestas situações, a média não é a melhor medida representativa da distribuição dos dados e, como consequência, as estimativas geradas podem levar a erros na seleção dos animais.

Embora inúmeras metodologias estatísticas tenham sido propostas para GWS, como RR-BLUP, Lasso, Lasso bayesiano, alfabeto bayesiano (Bayes A, Bayes B...), Quadrados Mínimos Parciais, regressão baseada em Kernel, dentre outras, as quais lidam com problemas de multicolinearidade e dimensionalidade, apenas a regressão quantílica regularizada (RQR) (NASCIMENTO et al., 2016) possibilita a obtenção de valores genéticos em diferentes níveis da distribuição do fenótipo de interesse.

A RQR permite um estudo mais completo das relações entre a variável resposta (fenótipo) e as explicativas (marcadores), uma vez que, diferentemente dos métodos tradicionais, que se baseiam em esperanças

condicionais, a mesma é baseada em quantis condicionais (NASCIMENTO et al., 2016) e, desta forma, pode-se estudar a relação existente para diferentes níveis do fenótipo de interesse e não somente em relação ao seu valor médio.

Em seleção genômica, esta metodologia foi apresentada visando a obtenção de uma "melhor" explicação sobre a relação funcional entre os marcadores e fenótipos cujas distribuições apresentam assimetria e que são comumente encontrados no melhoramento vegetal, como por exemplo, o número de dias para o florescimento (TUBEROSA, 2012) e também no melhoramento animal, como a concentração de hormônios (MATHUR et al., 2012).

Nesse sentido, este trabalho propôs a utilização da RQR na predição de valores genéticos de suínos para variáveis associadas à carcaça, em diferentes níveis da distribuição das variáveis de interesse. Além disso, os resultados obtidos por esta metodologia, em termos de acurácia, concordância dos animais para seleção e efeitos de marcadores, serão comparados com aqueles advindos do método usual, Lasso bayesiano (BLASSO).

2. Material e Métodos

A coleta dos dados fenotípicos utilizados neste estudo, foi realizada na Granja de Melhoramento de Suínos do Departamento de Zootecnia da Universidade Federal de Viçosa (UFV), no período de novembro de 1998 a julho de 2001. Esses dados são provenientes de uma população composta por 345 indivíduos F_2 de suínos, obtidos pelo cruzamento de animais da raça Piau com animais da raça Comercial.

O DNA do animal foi extraído no Laboratório de Biotecnologia do Departamento de Zootecnia da Universidade Federal de Viçosa, e os procedimentos utilizados estão descritos em Peixoto et al. (2006). A genotipagem para os 384 SNPs foi realizada no Laboratório de Genética Animal (LGA) da Embrapa Recursos Genéticos e Biotecnologia (Brasília, DF) via tecnologia Golden Gate/Vera Code® utilizando o leitor de BeadXpress Ilumina, conforme relatado por Hidalgo et al. (2013).

Os SNPs utilizados para o mapeamento fino foram selecionados de acordo com seu espaçamento entre cromossomos que continham QTLs previamente detectados nessa população e foram distribuídos da seguinte forma nos cromossomos de *Sus scrofa*: SSC1 (85), SSC4 (71), SSC7 (84), SSC8 (42), SSC17 (36) e SSCX (66). Destes, 66 SNPs foram descartados devido à ausência de amplificação, e dos 318 SNPs restantes, 81 foram descartados por apresentar menor frequência alélica. Após estes procedimentos,os marcadores SNPs foram distribuídos da seguinte forma: SSC 1 (56), SSC4 (54), SSC7 (59), SSC8 (31), SSC17 (25) e SSCX(12), que corresponde a um total de 237 marcadores que foram identificados para esta população (HIDALGO et al., 2013).

As características analisadas foram: rendimento de carcaça (RCARC), espessura de bacon (EBACON) e espessura de toucinho imediatamente após a última costela na linha dorso-lombar (ETUC). Os valores fenotípicos foram corrigidos para efeitos fixos de sexo, lote e presença do gene halotano.

Para avaliar a assimetria das distribuições das características estudadas, utilizou-se o teste de normalidade D'Agostino-Person (D'AGOSTINO e PEARSON, 1973), implementado no software R (Development Core Team, 2015) por meio da função *dagoTest* do pacote *fbasics* do R.

O modelo geral de predição de valores genéticos genômicos adotado, foi proposto por Meuwissen et al. (2001) e pode ser descrito como:

$$y_i = 1\mu + \sum_{j=1}^{237} x_{ij} g_j + e_i \qquad (1)$$

em que y é o vetor de fenótipos com dimensão $I \times 1$, em que $I = 345$ é o número de indivíduos, 1 é um vetor de mesma dimensão de y com todas as entradas iguais a 1, μ é a média da característica estudada, g_j é o j-ésimo efeito do marcador SNP (j=1,2,...,237), x_{ij} são os elementos da matriz de incidência de cada marcador j com parametrização 0,1 e 2 e, e é o vetor de erros aleatórios do modelo. Para obter as estimativas dos valores genéticos genômicos dos indivíduos (GBV's), adotou-se a Regressão Quantílica Regularizada (RQR) (LI e ZHU, 2008) considerando diferentes quantis ($\tau =$

$0{,}05$ a $\tau = 0{,}95)$ e o LASSO Bayesiano (BLASSO) (DE LOS CAMPOS et al. 2009).

A obtenção das estimativas dos coeficientes nos diferentes quantis de interesse, por meio da RQR, consiste na resolução do seguinte problema de otimização:

$$min\left\{\sum_{i=1}^{n}\rho_\tau\left(y_i-\mu-\sum_{j=1}^{p}x_{ij}g_j\right)+\lambda\sum_{j=1}^{p}|g_j|\right\} \qquad (2)$$

em que $\sum_{j=1}^{p}|g_j|$ é a soma dos valores absolutos dos coeficientes de regressão, λ é o parâmetro de suavização que controla a força da regularização, $n = 345$, $p = 237$ e $\rho_\tau(\,)$, denotada função *check* (KOENKER e BASSETT, 1978) definida por:

$$\rho_\tau\left(y_i-\mu-\sum_{j=1}^{p}x_{ij}g_j\right)=\begin{cases}\tau\left(y_i-\mu-\sum_{j=1}^{p}x_{ij}g_j\right), se\ \ y_i-\mu-\sum_{j=1}^{p}x_{ij}g_j\ >0\\ (1-\tau)\left(y_i-\mu-\sum_{j=1}^{p}x_{ij}g_j\right), \text{caso contrário}\end{cases}$$

em que $\tau \in (0{,}1)$ indica o quantil de interesse.

Após a estimação dos parâmetros (efeitos de marcadores) os valores genéticos genômicos são obtidos por meio da seguinte expressão:

$$G\hat{B}V(\tau)=\hat{y}_i(\tau)=\sum_j x_{ij}\hat{g}_j(\tau) \qquad (3)$$

em que $G\hat{B}V(\tau)$ é o valor genético genômico do i-ésimo indivíduo, e $\hat{g}_j(\tau)$ é o efeito do j-ésimo marcador SNP, sendo todos definidos pela relação funcional obtida para o $100\tau\,\%$ quantil de interesse. A obtenção das estimativas dos coeficientes por meio do LASSO Bayesiano (DE LOS CAMPOS et al. 2009) se dá a partir da solução do seguinte problema de otimização:

$$min\left\{\sum_{i=1}^{n}\left(y_i-\mu-\sum_{j=1}^{p}x_{ij}g_j\right)^2+\lambda\sum_{j=1}^{p}|g_j|\right\} \qquad (4)$$

em que $\sum_{j=1}^{p}|g_j|$ é a soma dos valores absolutos dos coeficientes de regressão e λ é o parâmetro de suavização que controla a força da regularização, de forma que, quando $\lambda = 0$ não há regularização. Na implementação do método, impõe-se como distribuição marginal *a priori* dos *p* coeficientes de regressão um produto de densidades exponenciais duplas: $p(g|\lambda) = \prod_{j=1}^{p} \frac{\lambda}{2\sigma_e} exp\left(\frac{-\lambda|g_j|}{\sigma_e}\right)$. Ademais, da mesma forma que para RQR, as estimativas dos valores genéticos genômicos ($G\hat{B}V$'s) via BLASSO são obtidas por meio de $G\hat{B}V = \hat{y}_i = \sum_j x_{ij}\hat{g}_j$, em que $G\hat{B}V$ é um vetor que contém o valor genético genômico estimado para todos os indivíduos.

Para fins de comparação das metodologias, considerou-se nos ajustes via RQR além dos valores do parâmetro de encolhimento (λ) estimado a partir do método BLASSO um grid de valores considerando valores variando de 0 até o valor fornecido pelo BLASSO, com intervalo de 0,5.

Após a obtenção dos GBV's por cada método, as metodologias apresentadas foram comparadas por meio da acurácia, calculada a partir da seguinte expressão:

$$r_{q,\hat{q}} = \frac{r_{y,\hat{y}}}{\sqrt{h^2}} \qquad (5)$$

em que $r_{y,\hat{y}}$, é a capacidade preditiva do modelo, dada por $r_{y,\hat{y}} = \frac{cov(y,\hat{y})}{\sqrt{Var(y)Var(\hat{y})}}$ e h^2 é a herdabilidade do caráter estimada pelo método REML sobre fenótipos em um modelo unicaracterístico, sendo definida por, $h^2 = \frac{v_g}{v_f}$ em que v_g é a variância genética e v_f é a variância fenotípica (RESENDE et al., 2010).

Por fim, foram calculados os coeficientes de correlação de postos de Spearman (SIEGEL,1975) para verificar o grau de concordância na classificação dos animais entre as metodologias.

Este coeficiente pode ser definido por:

$$\rho = 1 - \frac{6\sum {d_i}^2}{(n^3 - n)}$$

em que:

d_i= a diferença entre cada posto de valor correspondentes de x e y,

n= o número dos pares dos valores.

Ademais, foram selecionados os 10% indivíduos que apresentaram os maiores valores genéticos genômicos preditos pela RQR e pelo BLASSO e, a partir dos quais, foram calculados os percentuais de concordância de classificação entre os métodos.

Para avaliar a capacidade de generalização do modelo preditivo e evitar a superestimação das medidas ao avaliar o modelo estimado utilizando os mesmos dados da estimação, todo o processo descrito foi realizado utilizando validação cruzada. Especificamente, a população de suínos foi dividida em três populações distintas, onde se utilizou duas destas para estimação dos efeitos de marcadores e a outra para validação das estimativas obtidas a partir das equações de predição. As três combinações possíveis para essa situação foram utilizadas e o resultado final refere-se aos valores médios para acurácia e coeficiente de concordância.

A fim de verificar o comportamento dos efeitos dos marcadores SNPs nos cromossomos nas duas técnicas utilizadas, foram construídos gráficos de Manhattan, por meio da função *mhtp*, do pacote *gap* (ZHAO, 2015).

Para a estimação dos parâmetros das regressão e as demais análises, foi utilizada a função *rq* do pacote *quantreg* (KOENKER, 2006). Por outro lado, o modelo bayesiano foi ajustado utilizando-se a função *bglr* (com 100.000 interações, 20.000 de *burn-in* e thin assumindo o valor 10), implementada no pacote *BGLR* (CAMPOS e RODRIGUEZ, 2015). As rotinas computacionais implementadas no programa R estão apresentadas no Apêndice D.

3. Resultados e Discussões

Com base no teste de normalidade D'Agostino, verificou-se que as características estudadas apresentaram distribuição com comportamento assimétrico ao nível de significância $\alpha = 5\%$ (Tabela1). Dessa forma, a mediana ou algum quantil pode ser mais adequado para explicar a relação funcional entre os marcadores e as variáveis estudadas, uma vez que a média não é a medida que melhor representa as distribuições assimétricas (HAO e NAIMAN, 2007).

Tabela 1. P-valor associado ao teste de normalidade de D'Agostino-Person

Variável	P-valor
RCARC	0,0000
EBACON	0,0005
ETUC	0,0101

Notas: RCARC- rendimento de carcaça; EBACON - espessura de bacon; ETUC- espessura de toucinho imediatamente após a última costela na linha dorso-lombar.

Os valores de herdabilidade estimados foram: 0,20 para o rendimento de carcaça (RCARC), 0,35 para a espessura de toucinho (ETUC) e 0,34 para a espessura de bacon (EBACON). Esses valores estão de acordo com aqueles obtidos por Almeida Netto et al. (1993), Azevedo (2012) e Mendonça et al. (2012).

Segundo Lopes (2005), estas herdabilidades são consideradas de baixa à média magnitude (0,20 a 0,35). De modo geral, a baixa herdabilidade pode ser explicada quando a correlação entre o genótipo e o fenótipo é pequena e o efeito do ambiente se torna mais importante. Na prática isso significa que o desempenho apresentado pelo animal está associado ao ambiente a ele proporcionado. E, dessa forma, a resposta à seleção genética será muito lenta (TAMIOSO et al., 2013).

Para os diferentes ajustes que foram realizados para o caráter rendimento de carcaça (RCARC), considerando diferentes lambdas para RQR, tem-se que o modelo via RQR com $\tau = 0{,}15$ e o valor de $\lambda = 2{,}5$, foi aquele que apresentou maior valor de acurácia (0,48), conforme pode ser verificado na figura 1C do apêndice C, indicando a superioridade da RQR sobre o BLASSO (0,45) na predição dos GBV's para esta característica (Tabela 2).

Tabela 2. Acurácias obtidas pelos métodos BLASSO e RQR considerando diferentes valores para o parâmetro de encolhimento (λ), para as variáveis rendimento de carcaça (RCARC), espessura de bacon (EBACON) e espessura de toucinho imediatamente após a última costela na linha dorso-lombar (ETUC).

	RQR			**BLASSO**
Variável	**Quantil**	**Acurácia**	λ	**Acurácia**
RCARC	τ=0,15	0,48	2,5	0,45
EBACON	τ=0,45	0,54	14,5	0,45
ETUC	τ=0,50	0,39	11,5	0,39

A RQR também mostrou-se adequada para prever valores genéticos genômicos da variável EBACON quando se utilizou o quantil $\tau = 0{,}45$ e $\lambda = 14{,}5$ (Figura 2C do Apêndice C), cujo valor de acurácia foi igual à 0,54, novamente, sendo este superior ao obtido a partir do BLASSO (0,45) (Tabela 2).

Por outro lado, para a característica ETUC, o modelo de RQR que apresentou maior valor de acurácia (0,39) foi aquele utilizando $\tau = 0{,}50$ e $\lambda = 11{,}5$ (Figura 3C do Apêndice C), sendo este valor idêntico àquele fornecido pelo método BLASSO (Tabela 2).

Em geral, a RQ encontrou valores superiores ou iguais aqueles encontrados pelo BLASSO, o que indica que os métodos usuais podem não ser a melhor alternativa de análise em todos os casos.

Para fins de comparação, estimou-se os modelos de RQR considerando os valores de λ fornecidos pelo método BLASSO (Tabela 3). Como resultados, tem-se que estes modelos apresentaram valores de acurácia inferiores aos obtidos pelo BLASSO e pelos modelos já apresentados. De acordo com esses resultados, tem-se que a utilização de valores idênticos de λ nos ajustes via RQR e BLASSO não é uma boa estratégia para a RQR uma vez que a força de regularização (*shrinkage*) é diferente nos dois métodos. Dessa forma a escolha da metodologia deve ser baseada entre a comparação do método BLASSO com diversos ajustes obtidos para diferentes quantis e valores de λ.

Tabela 3. Acurácias obtidas pelo método BLASSO e RQR considerando o valor do parâmetro de encolhimento (λ), fornecido pelo BLASSO, para as variáveis rendimento de carcaça (RCARC), espessura de bacon (EBACON) e espessura de toucinho imediatamente após a última costela na linha dorso-lombar (ETUC)

	RQR			**BLASSO**
Variável	**Quantil**	**Acurácia**	λ	**Acurácia**
RCARC	τ=0,15	-0,093	43,22	0,45
EBACON	τ=0,45	0,260	28,44	0,45
ETUC	τ=0,50	0,00993	31,37	0,39

De acordo com as estimativas das correlações dos valores genéticos genômicos estimados pelo BLASSO e RQR observou-se a correlação entre o BLASSO e a RQR para o quantil $\tau = 0{,}15$ e valor de $\lambda = 2{,}5$, considerando a variável RCARC, foi alta e igual a 0,58 (Tabela 4).

Considerando a espessura de bacon (EBACON), observou-se a maior correlação entre o BLASSO e a RQR para o quantil $\tau = 0{,}45$ (Tabela 4), cujo valor foi igual à 0,72.

Por outro lado, a correlação entre o BLASSO e a RQR para o quantil $\tau = 0{,}50$ e valor de $\lambda = 11{,}5$ para a variável ETUC apresentou-se moderada (0,42) (Tabela 4).

Tabela 4. Correlações das estimativas dos GBV's (Valores Genéticos Genômicos) provenientes dos métodos BLASSO e RQR para $\tau = 0{,}15, 0{,}45\ e\ 0{,}50$, considerando valores de $\lambda = 2{,}5, 14{,}5\ e\ 11{,}5$, para as características de carcaça

Variável	MÉTODOS	
	RQR	**BLASSO**
RCARC	τ =0,15	0,58
EBACON	τ =0,45	0,72
ETUC	τ =0,50	0,42

Adicionalmente, calculou-se o coeficiente de correlação de postos de Spearman dos GBV's estimados pelos métodos BLASSO e RQR, a fim de compará-los quanto à concordância na classificação dos animais. Para tanto, foram considerados os modelos via RQR para os quantis $\tau = 0{,}15, 0{,}45\ e\ 0{,}50$ conforme apresentado nas Tabela 5.

Tabela 5. Coeficientes de correlação de postos de Spearman obtidas para os métodos BLASSO e RQR considerando $\lambda = 2{,}5, 14{,}5\ e\ 11{,}5$, para as variáveis rendimento de carcaça (RCARC), espessura de bacon (EBACON) e espessura de toucinho imediatamente após a última costela na linha dorso-lombar (ETUC)

Variável	Tau	Coeficiente de Correlação de Spearman
RCARC	τ =0,15	0,05
EBACON	τ =0,45	0,35
ETUC	τ =0,50	0,69

Foi possível notar um baixo valor para o coeficiente de correlação de Spearman (0,05) entre o BLASSO e a RQR para $\tau = 0{,}15\ e\ \lambda = 2{,}5$ para a característica RCARC, indicando que existe baixa correlação entre os métodos na classificação dos animais para esta característica, ou seja, os métodos estão classificando os animais de forma diferente (Tabela 5).

Analisando a variável EBACON, verifica-se que o coeficiente de correlação de Spearman obtido foi igual a 0,35 para RQR ($\tau = 0{,}45\ e\ \lambda = 14{,}5$)(Tabela 5). Esses valores mostram que há moderada correlação entre os métodos na classificação dos animais considerando esta característica.

Por outro lado, considerando a variável ETUC pode-se notar que o coeficiente de correlação de Spearman entre o BLASSO e a RQR ($\tau = 0{,}50\ e\ \lambda = 11{,}5$) obtido foi um valor alto e igual à 0,69 (Tabela 5), ou seja, as classificações dos animais foram semelhantes pelos modelos. Neste caso, os resultados foram satisfatórios em relação a concordância entre as metodologias na classificação dos animais, uma vez que para modelo de RQR($\tau = 0{,}50\ e\ \lambda = 11{,}5$) considerando a variável ETUC a concordância foi maior, conforme era esperado.

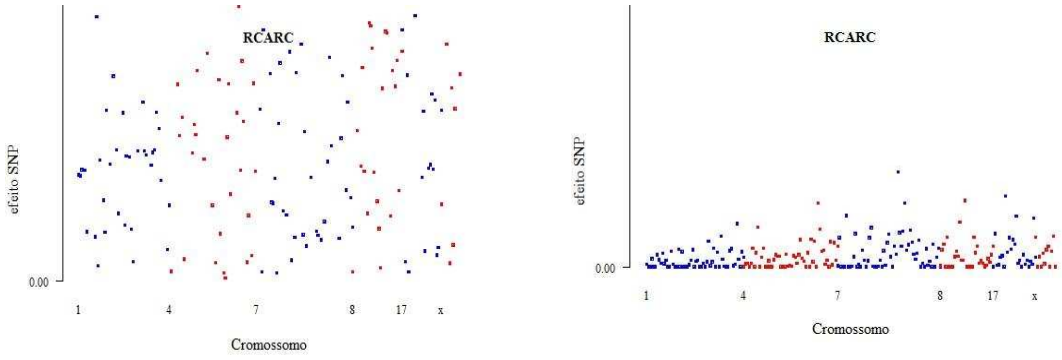
Com relação à concordância entre os métodos RQR e BLASSO na classificação dos 10% indivíduos que apresentaram os maiores valores genéticos genômicos, observou-se que para a variável RCARC os métodos BLASSO e a RQR ($\tau = 0{,}15\ e\ \lambda = 2{,}5$) apresentaram o menor percentual de coincidência, coincidência igual à 11,4% (Tabela 6). Analisando a variável EBACON, tem-se que os métodos foram mais coincidentes, coincidência igual à 34,3%, quando se utilizou o modelo de RQR para $\tau = 0{,}50\ e\ \lambda = 11{,}5$ (Tabela 6). Para ETUC, os métodos BLASSO e a RQR ($\tau = 0{,}50\ e\ \lambda = 11{,}5$) apresentaram o maior percentual de coincidência, sendo este valor igual à 54,3%, na classificação dos melhores animais (Tabela 6). Este resultado era esperado para a RQR para $\tau = 0{,}50$ uma vez que o valor desse quantil é próximo ao valor médio dessa variável. Ademais, como a RQR tem melhores valores de acurácia, é possível que ela esteja classificando melhor os indivíduos.

Tabela 6. Porcentagens de classificações coincidentes entre os métodos BLASSO e RQR considerando $\lambda = 2{,}5, 14{,}5\ e\ 11{,}5$, para as variáveis rendimento de carcaça (RCARC), espessura de bacon (EBACON) e espessura de toucinho imediatamente após a última costela na linha dorso-lombar (ETUC)

Variável	Quantil	Percentual de coincidência (%)
RCARC	τ =0,15	11,4
EBACON	τ =0,45	34,3
ETUC	τ =0,50	54,3

Finalmente, são apresentados os gráficos de Manhattan para analisar o comportamento dos efeitos dos SNPs nos cromossomos.

(a) (b)



BLASSO $\tau = 0.15$

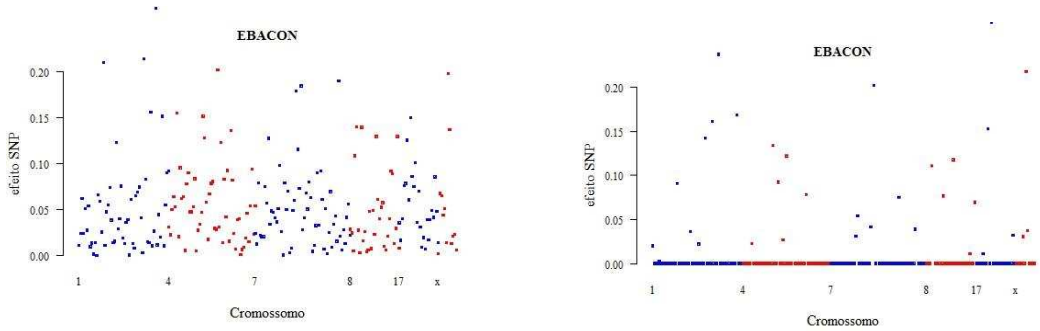
Figura 1. Gráficos de Manhattan dos efeitos dos marcadores considerando os modelos estimados pelo BLASSO (a) e RQR ($\tau = 0{,}15$ e $\lambda = 2{,}5$)(b) para a variável RCARC.

Observou-se que marcadores com efeitos relevantes na característica RCARC pelo método BLASSO (Figura 1a) encontram-se nos cromossomos 1, 4, 7, 8 e 17.

No gráfico da RQR para o quantil τ =0,15 (Figura 1b) percebe-se que SNP's de maiores efeitos encontram-se no cromossomo 7. Este resultado é

corroborado com o que foi encontrado por Evans et al. (2004), que identificou um QTL significativo para suínos de linhagem comercial para essa característica neste mesmo cromossomo.

(a) (b)

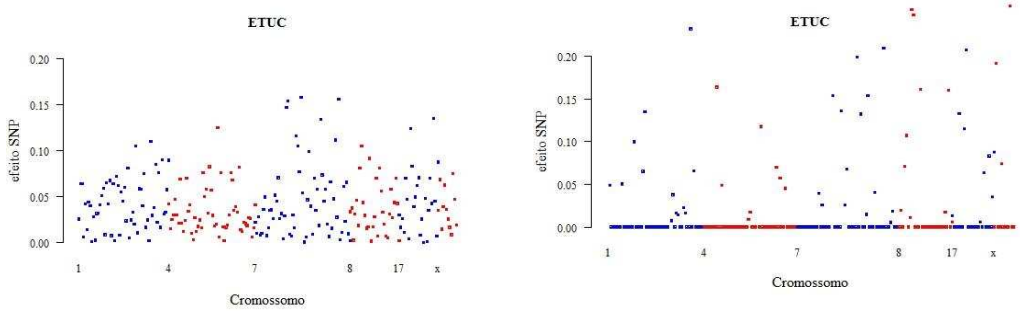


BLASSO $\tau = 0{,}45$

Figura 2. Gráficos de Manhattan dos efeitos dos marcadores considerando os modelos estimados pelo BLASSO e RQR ($\tau = 0{,}45$ e $\lambda = 14{,}5$) para a variável EBACON.

Para EBACON pode-se notar que SNPs de maiores efeitos encontram-se nos cromossomos 1, 4 e 7 pelo método BLASSO (Figura 2a), concordando com resultados encontrados por Silva et al. (2011) que identificou QTL significativo para essa característica no cromossomo 4 em uma população F_2 (Pial x Comercial) de suínos. No gráfico da RQR pode ser verificado efeitos mais relevantes de SNPs nos cromossomos 1, 7 e 17.

(a) (b)



BLASSO $\tau = 0{,}50$

Figura 3. Gráficos de Manhattan dos efeitos dos marcadores considerando os modelos estimados pelo BLASSO e RQR ($\tau = 0.50$ e $\lambda = 11.5$) para a variável ETUC.

Os SNPs de maiores efeitos na variável ETUC foram encontrados nos cromossomos 1, 4, 7 e 17 pelo método BLASSO (Figura 3a). Este resultado foi concordante com o estudo realizado por Fan et al. (2011), que identificaram QTL's nos mesmos cromossomos em uma população de fêmeas da raça Large White e Hidalgo et al. (2013) que identificaram um QTL no cromossomo 4 para essa mesma característica que identificaram a presença de QTL's para essa mesma variável nos mesmos cromossomos na população F_2 de suínos da raça Piau x Comercial. Na RQR (Figura 3b) os SNP's mais relevantes foram encontrados nos cromossomos 1, 7, 8 e 17 (Figura 3b).

O padrão dos efeitos de marcadores encontrados para as variáveis pelo método BLASSO foi diferente ao encontrado pela RQR. Percebe-se ainda que os efeitos foram muitos pequenos para a característica RCARC, conforme apresentado na Figura 1. Para EBACON, esses efeitos apresentaram maior magnitude para ambos os métodos (Figura 2). Já para a característica ETUC os efeitos estimados pela RQR para τ=0.50 foram de maior magnitude em relação ao do BLASSO (Figura 3). Ademais, é possível verificar que o poder de encurtamento dos efeitos dos marcadores na RQR é maior que no método BLASSO conforme mostrado nas Figuras.

4. Conclusões

A utilização de modelos de regressão quantílica em estudos de seleção genômica é uma abordagem interessante, uma vez que os mesmos apresentaram valores de acurácia maiores ou iguais que aqueles obtidos pelo BLASSO. Em relação à classificação dos animais, os métodos classificaram os animais de forma semelhante para as variáveis EBACON e ETUC e de forma diferente para o RCARC. Ademais, o padrão dos efeitos de marcadores encontrados para as variáveis pelo método BLASSO foi diferente ao encontrado pela RQR.

REFERÊNCIAS BIBLIOGRÁFICAS

ABPA: Associação Brasileira de Proteína Animal. Disponível em: <http://www.abipecs.org.br/pt/estatisticas/mundial/producao-2.html>. Acesso em: Nov. 2014.

ALMEIDA NETO, P. P.; OLIVEIRA, A. I. G.; ALMEIDA, A. J. L.; LIMA, A. F.; SILVA, M. A.; COSTA, C. N. Parâmetros genéticos e fenotípicos de características de carcaça de suínos. **Rev. Soc. Bras. Zootec**, v. 22, p. 624-633, 1993.

AZEVEDO, C. F. **Métodos de redução de dimensionalidade aplicados na seleção genômica para características de carcaça em suínos**. Dissertação (Mestrado em Estatística Aplicada e Biometria), Universidade Federal de Viçosa, Viçosa, 2012.

CAMPOS e *RODRIGUEZ, Bayesian Generalized Linear Regression.*

URL: http://cran.r-project.org/web/packages/BGLR/index.html, 2015.

D'AGOSTINO, R. B.; PEARSON, E. S. Tests for Departure from Normality. **Biometrika,** v. 60, p. 613–22, 1973.

DE LOS CAMPOS, G.; NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J. M. Predicting quantitative traits with regression models for dense molecular makers. **Genetics**, Austin, v. 182, p. 375-385, 2009.

FAN, B.; ONTERU, S. K.; DU, Z. Q.; GARRICK, D. J.; STALDER, K. J.; ROTHSCHILD, M. F. Genome-wide association study identifies Loci for composition and sctructural soundness traits in pigs. **PlosOne**. v. 6, p. 1-11, 2011.

HAO, L.; NAIMAN, D. Q. **Quantile regression**. Sage publications. 126p, 2007.

HIDALGO, A.M.; Lopes, P. S.; Paixão, D. M.; Silva, F. F.; Bastiaansen, J. W.M.; Paiva, S. R.; Faria, D. A.; Guimarães, S. E.F. Fine mapping and single nucleotide polymorphism effects estimation on pig chromossomes 1,4,7,8,17 and x. **Genetics and Molecular Biology**, v. 36, nº 4, p. 511-519, 2013.

KOENKER, R.; BASSET, G. Regression Quantiles, **Econometrica**, v. 46, p. 33–50, 1978.

KOENKER, R. **Quantile regression in R**: A vignette, 2006. Disponível em: <http://www.econ.uiuc.edu/~roger/research/rq/vig.pdf>. Acesso em: Mar. 2015.

LI, Y.; ZHU, J. L1-Norm Quantile Regression. **Journal of Computational and Graphical Statistics**, v.17, p. 163–185, 2008.

LOPES, P. S. **Teoria do melhoramento animal**. Belo Horizonte: FEPMVZ, 118p. , 2005.

MAPA. Ministério da Agricultura, Pecuária e Abastecimento. 2014. Disponível em: <http://www.agricultura.gov.br/animal/especies/suinos>. Acesso em: Jan. 2015.

MATHUR P.K.; TEN, N.J.; BLOEMHOF, S.; HERES, L.; KNOL, E.;, MULDER, H.A. A 322 human nose scoring system for boar taint and its relationship with 323 androstenone and skatole. **Meat Sci,** v. 91, p. 414–422, 2012.

MENDONÇA, P. T.; LOPES, P. S.; BRANCCINI NETO, J.; CARNEIRO, P. L. S.; TORRES, R. de A.; GUIMARÃES, S. E. F.; VERONEZE, R. Estimação de parâmetros genéticos de uma população F_2 de suínos. **Revista Brasileira de Saúde e Produção Animal,** v.13, p. 330-343, 2012.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction os total genetic value using Genome-Wide dense marker maps. **Genetics Society of America**, v. 157, p. 1819-1829, 2001.

NASCIMENTO, M.; SILVA, F.F. E; RESENDE, M.D.V. DE; CRUZ, C.D.; NASCIMENTO, A.C.C.; VIANA, J.M.S.; BARROSO, L.M.A. Regularized quantile regression applied to genome-enabled prediction for skewness quantitative traits. **PlosOne**, No prelo 2016.

PEIXOTO, J. O.; GUIMARAES, S. E. F.; LOPES, P. S.; SOARES, M. A. M.; PIRES, A. V.; SILVA, M. V.; TORRES, R. A.; SILVA, M. A. E. Associations of leptin gene polymorphisms with production traits in pigs. **Journal of Animal Breeding and Genetics**, v.123, p. 378-383, 2006.

RESENDE, M. D. V.; SILVA, F. F.; VIANA, J. M.; PETTERNELLI, L. A.; RESENDE JR, M. F. R.; VALLE, P**. Computação da seleção genômica ampla.** Colombo: Embrapa Florestas, 79p. 2010.

R DEVELOPMENT CORE TEAM. **R: A language and environment for statistical computing.** Vienna, Austria: R Fundation for Statistical Computing, 2009. Disponível em: <http://r-project .org>. Acesso em: Jan. 2016.

SIEGEL, S. **Estatística Não-Paramétrica: Para as Ciências do Comportamento**. São Paulo: McGraw-Hill do Brasil, 350 p. 1975.

SILVA, F. F.; ROSA, G. J. M.; GUIMARÃES, S.E.F.; LOPES, P.S.; DE LOS CAMPOS, G. Thee-step Bayesian factor analysis applied to QTL detection in crosses between outbred pig populations. **Livestock Science**, v.42, p. 210-215, 2011.

TAMIOSO, P. R.; DIAS, L. T. **Herdabilidade e sua importância na seleção de ovinos de corte**. Disponível em: http://www.milkpoint.com.br/radar-tecnico/ovinos-e-caprinos, 2013. Acesso em: Set. 2015.

TUBEROSA, R. Phenotyping for drought tolerance of crops in the genomicsera. **Front Physio**, p. 3-347, 2012.

ZANGERONIMO, M. G.; FIALHO, E. T.; LIMA, J. A. F.; GIRÃO, L. V. C.; AMARAL, N. O.; SILVEIRA, H. Desempenho e características de carcaça de suínos dos 20 aos 50kg recebendo rações com reduzido teor de proteína bruta e diferentes níveis de lisina digestível verdadeira. **Ciência Rural**, v.39, p.1507-1513, 2009.

ZHAO, Genetic analysis package.

URL: http://cran.r-project.org/web/packages/gap/index.html, 2015.

CONSIDERAÇÕES FINAIS

Este trabalho foi uma proposta da aplicação de Regressão Quantílica Regularizada no contexto da seleção genômica ampla em suínos. O método proposto permite lidar com problemas de multicolinearidade e dimensionalidade, além de permitir estimar o efeito em diferentes níveis da característica de interesse.

Com a aplicação dos modelos de regressão quantílica para estimar o mérito genético dos indivíduos e, posteriormente o cálculo da acurácia, percebeu-se que para os quantis $\tau = 0.15, 0.45\ e\ 0.50$ e os respectivos valores de utilizados $\lambda = 2{,}5, 14{,}5\ e\ 11{,}5$ os valores de acurácia foram superiores ou iguais aqueles encontrados pelo BLASSO. Em relação à classificação dos animais, os métodos classificaram os animais de forma semelhante para as variáveis ETUC e EBACON e de forma diferente para o RCARC. Além disso, o padrão dos efeitos de marcadores encontrados para as variáveis pelo método BLASSO foi diferente ao encontrado pela RQR.

Logo, a utilização de modelos de regressão quantílica em estudos de seleção genômica é uma abordagem interessante e promissora.

Cabe destacar, que a utilização de valores idênticos de λ nos ajustes via RQR e BLASSO, não é uma boa estratégia para a RQR uma vez que a força de regularização (*shrinkage*) é diferente nos dois métodos. Dessa forma, a escolha da metodologia deve ser baseada entre a comparação do método BLASSO com diversos ajustes obtidos para diferentes quantis e valores de λ.

Apêndice A: Demonstração de que a mediana minimiza a média da distância absoluta

Demonstração : Suponha que F seja uma função de distribuição acumulada e f uma função de densidade de probabilidade. Assim, tem-se:

$$E|Y-m| = \int_{-\infty}^{+\infty} |y-m| f(y)\, dy = \int_{-\infty}^{m} |y-m| f(y)\, dy + \int_{m}^{+\infty} |y-m| f(y)\, dy$$

$$= \int_{-\infty}^{m} (m-y) f(y) dy + \int_{m}^{+\infty} (y-m) f(y) dy$$

Sabe-se que para obter o mínimo de uma função é necessário que sua derivada parcial seja igual a zero. Logo,

$$\frac{\partial}{\partial m} \int_{-\infty}^{+\infty} |y-m| f(y) dy = \frac{\partial}{\partial m} \left[\int_{-\infty}^{m} (m-y) f(y) dy + \int_{m}^{+\infty} (y-m) f(y) dy \right]$$

$$= \frac{\partial}{\partial m} \int_{-\infty}^{m} (m-y) f(y) dy + \frac{\partial}{\partial m} \int_{m}^{+\infty} (y-m) f(y) dy$$

$$= \int_{-\infty}^{m} \frac{\partial}{\partial m} (m-y) f(y) dy + \int_{m}^{+\infty} \frac{\partial}{\partial m} (y-m) f(y) dy$$

$$= \int_{-\infty}^{m} f(y) dy + \int_{m}^{+\infty} -f(y) dy$$

$$= \int_{-\infty}^{m} f(y) dy - \int_{m}^{+\infty} f(y) dy$$

$$= F(m) - \big(1 - F(m)\big) = 2F(m) - 1 \qquad (1)$$

Igualando (1) a zero, tem-se:

$$2F(m) - 1 = 0 \Rightarrow 2F(m) = 1 \Rightarrow F(m) = \frac{1}{2}$$

Portanto, como $F(m) = \frac{1}{2}$ tem-se que o valor de m que minimiza a média da distância absoluta é a mediana

Apêndice B: Detalhes do algoritmo de minimização da soma dos erros ponderados

O algoritmo de minimização da soma dos erros ponderados é exemplificado conforme descrito em Hao e Naiman (2007).

A figura (1B) apresenta 4 pares de pontos hipotéticos (-3,-1), (1,3), (2,-2) e (3,2) e as seis retas que ligam cada um destes pares de pontos.

Suponha que deseja-se encontrar a reta ajustada pela regressão mediana $(\tau = 0.5)$, que passa por um par de pontos do conjunto de forma que, metade dos pontos estejam abaixo desta reta e a outra metade acima. Deste modo tem-se que minimizar a seguinte equação:

$$\sum_{i=1}^{n} d_\tau\,(y_i, \hat{y}_i) = \sum_{y_i \geq \hat{y}_i} 0{,}5|y_i - \beta_0(0{,}5) + \beta_1(0{,}5)x_{i1}|$$
$$+\sum_{y_i < \hat{y}_i}(1 - 0{,}5)|y_i - \beta_0(0{,}5) + \beta_1(0{,}5)x_{i1}| \qquad (1)$$

A linha destacada (vermelha) representa a reta ajustada pela regressão mediana $(\tau = 0.5)$, uma vez que apresenta a menor soma das distâncias ponderadas para $(\tau = 0.5)$ e justamente um ponto abaixo e acima da reta (Figura 1B).

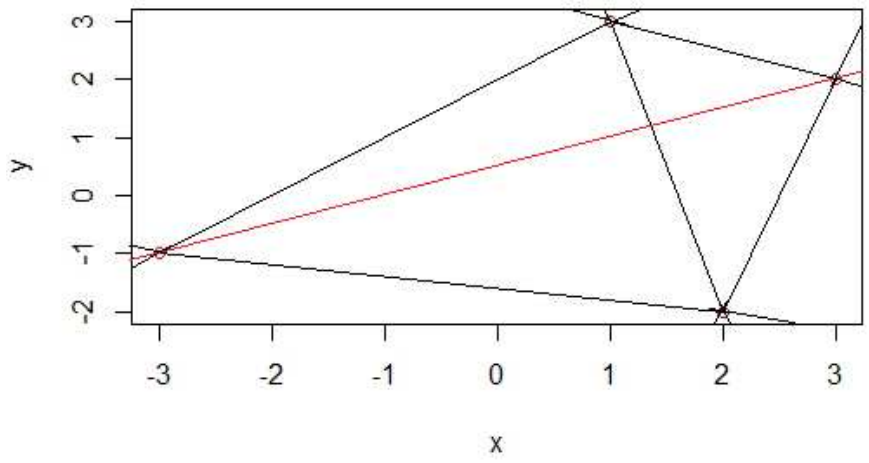


Figura 1B. Gráfico do plano (x,y).

Cada reta no plano (x, y) é representada por uma equação $y = \beta_0 + \beta_1 x$ para algum ponto (β_0, β_1). Assim é possível determinar uma relação entre os pontos do plano (x, y) e as retas do plano (β_0, β_1), já que para o plano (x, y) tem-se a reta $y = \beta_0 + \beta_1 x$, e de modo que (β_0, β_1) encontram-se na reta $\beta_1 = \left(\frac{y_i}{x_i}\right) - \left(\frac{1}{x_i}\right)\beta_0$. Segundo Edgeworth (1888) esta relação é conhecida como dualidade ponto/reta (Figura 2B) (HAO E NAIMAN, 2007).

A figura 2B apresenta o plano (β_0, β_1) que contém um ponto correspondente a cada reta do plano (x, y). Em particular, o ponto destacado corresponde à reta da regressão mediana da Figura 1B.

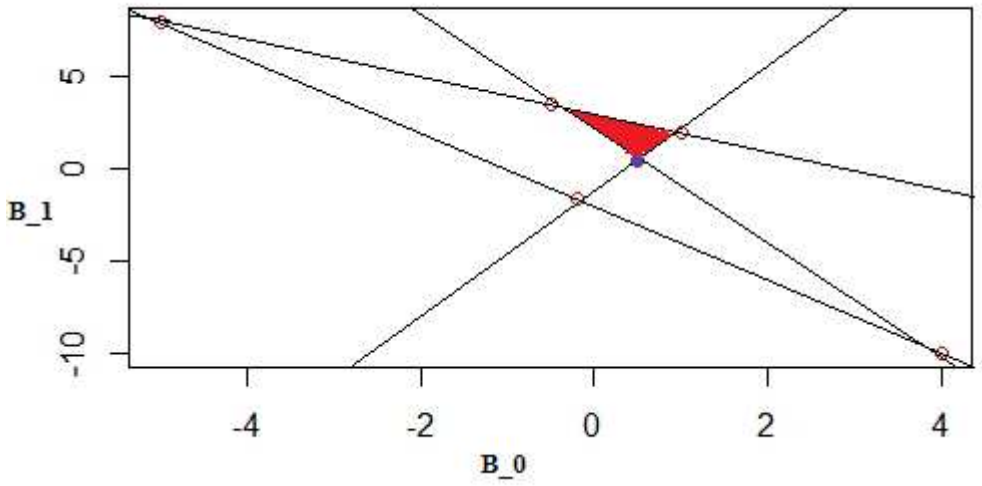


Figura 2B. Ilustração da dualidade ponto/reta.

As quatro linhas que são apresentadas na Figura 2B correspondem aos quatro pontos da Figura 1B. Estas linhas dividem o plano (β_0, β_1) em regiões poligonais. Um exemplo dessa região é sombreada na Figura 2B. Os pontos que formam essas regiões poligonais correspondem a uma conjunto de retas no plano (x, y), os quais dividem o conjunto de dados em dois conjuntos da mesma forma. Ademais, como utilizou-se apenas retas a função de (β_0, β_1) que minimiza a equação (1) é linear em cada região. Considerando uma nova dimensão definida como os valores para todos os pares de betas (β_0, β_1) e de observações da equação (1) tem-se que esta função é convexa com um

gráfico que forma uma superfície poliédrica e o par de betas que minimiza a função é a reta mediana, ilustrada na Figura 3B.

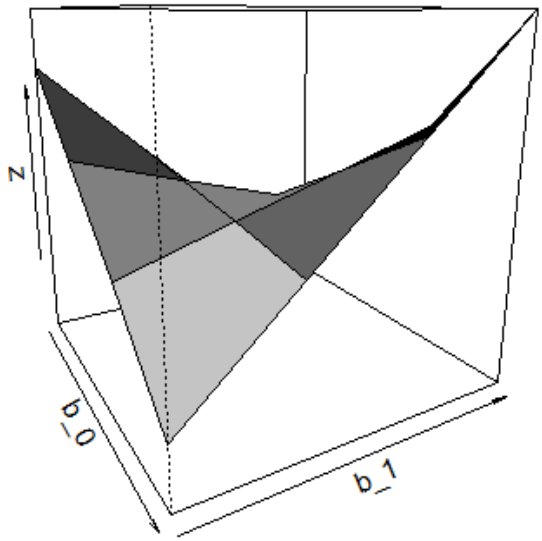


Figura 3B. Superfície Poliédrica.

Um algoritmo que possibilita minimizar a soma dos erros absolutos ponderados, estimando os coeficientes da regressão mediana $(\widehat{\beta_0}, \widehat{\beta_1})$ é baseado em algoritmos que resolvem problemas de programação linear (HAO e NAIMAN, 2007). O método simplex é utilizado para solução de problemas de programação linear . De acordo com o método, a partir de qualquer um dos pontos (β_0, β_1) correspondente aos vértices da região poligonal (Figura 3B), a minimização é obtida por movimentos iterativos de vértice a vértice ao longo das arestas da superfície poliédrica, escolhendo o caminho onde a distância é mínima. O objetivo é encontrar ponto que corresponde ao menor valor da equação (1).

Este resultado pode ser generalizado para permitir estimar o τ-ésimo quantil de regressão (KOENKER e D'OREY, 1987). Assim, defini-se o τ-ésimo estimador da regressão quantilica $\widehat{\beta_0}(\tau)$, $\widehat{\beta_1}(\tau)$ ..., $\widehat{\beta_p}(\tau)$ como os valores que minimizam a soma dos erros absolutos ponderados da equação (6). Em outras palavras, busca-se minimizar a soma de resíduos $y_i - \hat{y}_i$, de modo que resíduos positivos recebem um peso τ e resíduos negativos recebem um peso $1 - \tau$. Dessa forma, o τ-ésimo estimador de regressão quantílica $\widehat{\beta_0}(\tau)$, $\widehat{\beta_1}(\tau)$ ..., $\widehat{\beta_p}(\tau)$ são escolhidos para minimizar a equação 6.

Apêndice C: Gráficos das acurácias para as variáveis RCARC, EBACON e ETUC considerando os quantis ($\tau = 0,05$ a $\tau = 0,95$).

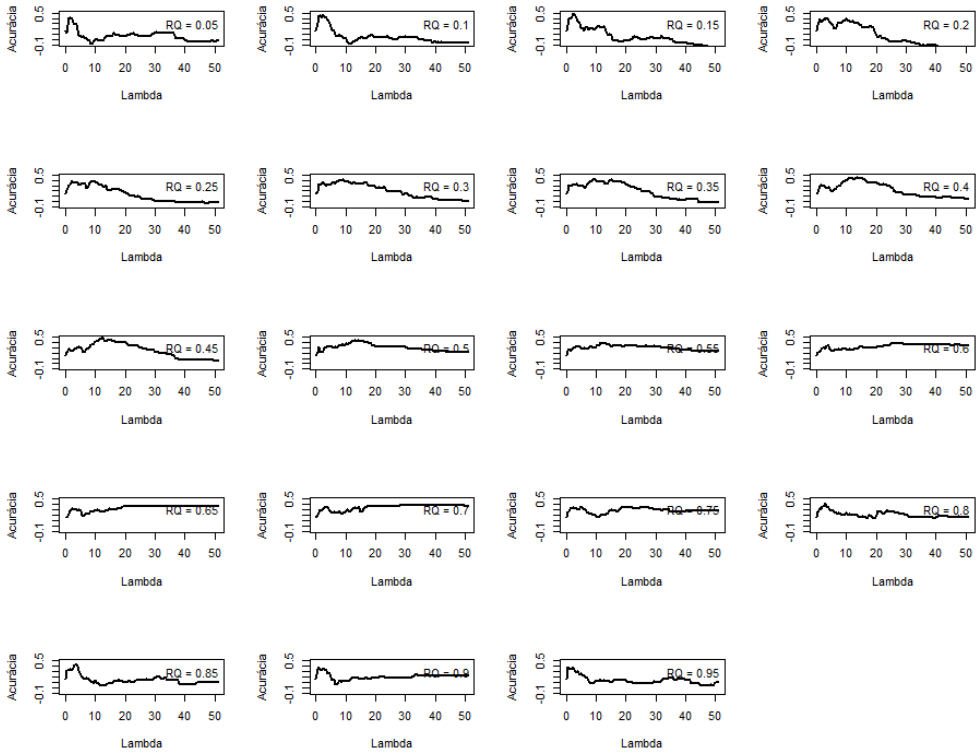


Figura 1C: Acurácia relativa a variável RCARC.

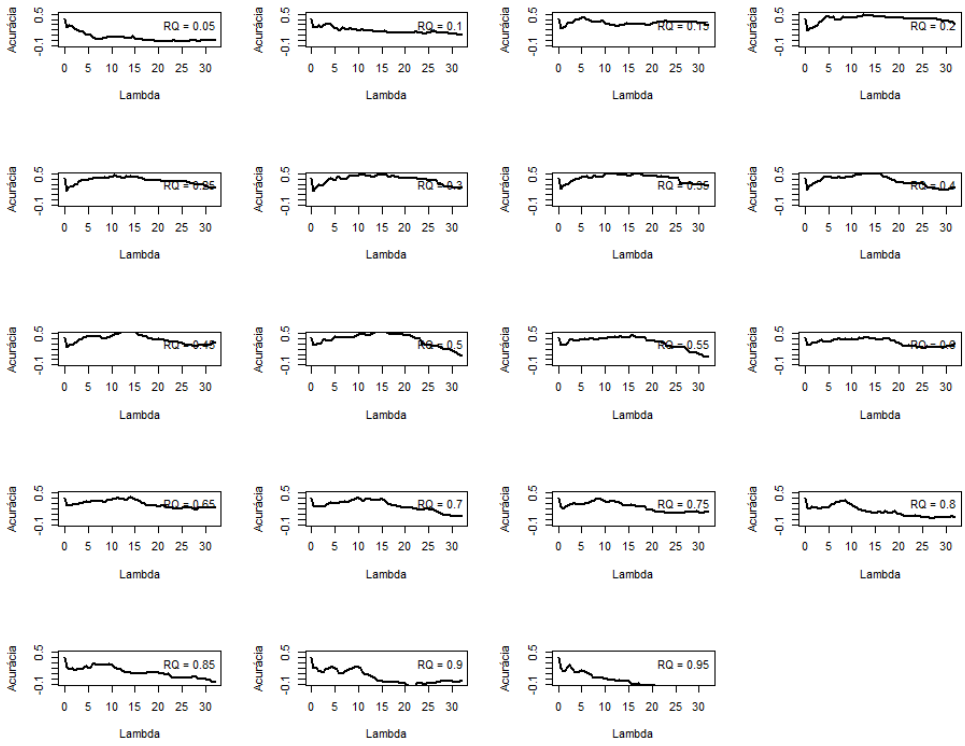


Figura 2C: Acurácia relativa a variável EBACON.

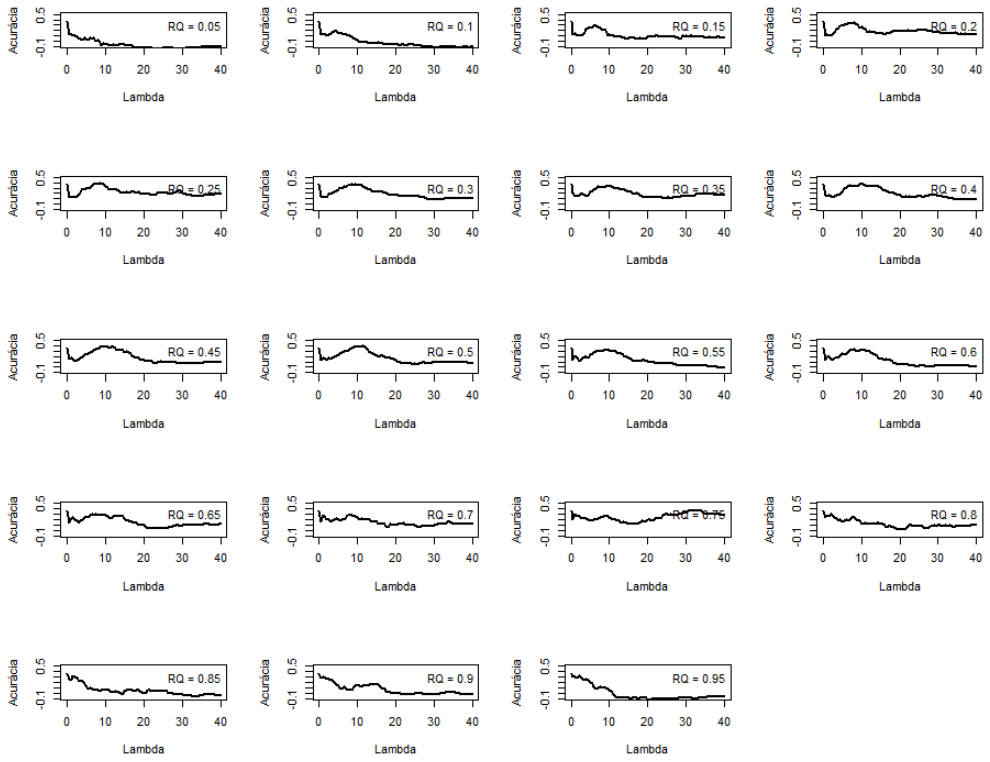


Figura 3C: Acurácia relativa a variável ETUC.

Apêndice D: Rotinas computacionais utilizadas para a análise.

As rotinas computacionais dos métodos descritos neste trabalho foram implementadas no software livre R (R Development Core Team, 2015) e estão descritas a seguir.

1. Leitura do conjunto de dados, Teste D'Agostino-Person e correção das variáveis.

```
dados<-read.table("dados_suínos.txt",h=T)
mapa<-read.table("mapa1.txt",h=T)
library(BGLR)
library(quantreg)
library(gap)
############ Teste D' Agostino###########
library(fBasics)
normalidade<-dagoTest(dados$fenótipo)
 # Os fenótipos utilizados foram: RCARC, EBACON e ETUC.
```

#Correção das variáveis para efeitos fixos de sexo, lote e presença do gene halotano

```
rm(list=ls(all=TRUE))
v<-matrix(NA,345,48)
for (i in 5:52){
  v[,i-4]<-dados[,i]
}
head(v)
# Correção da Matriz v:
sexo<-dados[,2]
lote<-dados[,3]
hal<-dados[,4]
```

```
vcor<-matrix(NA,345,48)
for (i in 1:48)
{
 vcor[,i]<-lm(v[,i]~sexo+lote+hal)$residuals
}
#############Variáveis que foram utilizadas############
RCARC<-vcor[,6]
EBACON<-vcor[,14]
ETUC<-vcor[,10]
############################################################
# Demais variáveis
PBR<-vcor[,27]
PCARC<-vcor[,1]
PCD<-vcor[,2]
TLD<-vcor[,3]
TLN<-vcor[,4]
IDA<-vcor[,5]
MBCC<-vcor[,7]
MLC<-vcor[,8]
PROLOM<-vcor[,15]
AOL<-vcor[,16]
CORAC<-vcor[,17]
PP<-vcor[,18]
PPL<-vcor[,19]
PCOPA<-vcor[,20]
PPA<-vcor[,21]
PC<-vcor[,22]
```

```
PL<-vcor[,23]

PB<-vcor[,24]

PCOST<-vcor[,25]

PF<-vcor[,26]

CR<-vcor[,28]

GPD<-vcor[,29]

CA<-vcor[,30]

NT<-vcor[,31]

PA<-vcor[,32]

PN<-vcor[,33]

ph45<-vcor[,34]

ph24<-vcor[,35]

L<-vcor[,36]

GOINTR<-vcor[,37]

PGOTEJ<-vcor[,38]

PCOZ<-vcor[,39]

MACIEZ<-vcor[,40]

C<-vcor[,41]

############################################################

# Validação

feno<-read.table("fenótipos.txt", h=T)

feno<-as.matrix(feno)

geno<-read.table("Marcadores.txt", h=T)

head(geno)

##### FENÓTIPOS ######

Populações individuais:

feno1=feno[1:115]
```

```
feno1=as.matrix(feno1)
feno2=feno[116:230]
feno2=as.matrix(feno2)
feno3=feno[231:345]
feno3=as.matrix(feno3)
# Populações agrupadas:
feno12=feno[1:230]
feno12=as.matrix(feno12)
feno13=rbind(feno1,feno3)
feno13=as.matrix(feno13)
feno23=feno[116:345]
feno23=as.matrix(feno23)
##### GENÓTIPOS #####
# Populações individuais:
geno1=geno[1:115,]
geno1=as.matrix(geno1)
geno2=geno[116:230,]
geno2=as.matrix(geno2)
geno3=geno[231:345,]
geno3=as.matrix(geno3)
# Populações agrupadas:
geno12=geno[1:230,]
geno12=as.matrix(geno12)
geno13=rbind(geno1,geno3)
geno23=geno[116:345,]
geno23=as.matrix(geno23)
```

```
######## Bayesian LASSO ########

#### Treinamento: 1 e 2 | Validação: 3 ####

BL1=BGLR(y=feno12,ETA=list(list(X=geno12,model='BL')),nIter=10000,
burnIn=2000,thin=10)

 gebv_val_BL1=geno3%*%BL1$ETA[[1]]$b          #valoresgenéticos
genômicos

lambda1<-(BL1)$ETA[[1]]$lambda

cor_BL1=cor(feno3,gebv_val_BL1)

cor_BL1

#### Treinamento: 1 e 3 | Validação: 2 ####

BL2=BGLR(y=feno13,ETA=list(list(X=geno13,model='BL')),nIter=10000
0,burnIn=20000,thin=10)

gebv_val_BL2=geno2%*%BL2$ETA[[1]]$b

lambda2<-(BL2)$ETA[[1]]$lambda

cor_BL2=cor(feno2,gebv_val_BL2)

cor_BL2

#### Treinamento: 2 e 3 | Validação: 1 ####

BL3=BGLR(y=feno23,ETA=list(list(X=geno23,model='BL')),nIter=10000
0,burnIn=20000,thin=10)

gebv_val_BL3=geno1%*%BL3$ETA[[1]]$b

lambda3<-(BL3)$ETA[[1]]$lambda

cor_BL3=cor(feno1,gebv_val_BL3)

cor_BL3

##################################################

# Média das capacidades preditivas - BLASSO

Cor_BL<-(cor_BL1+cor_BL2+cor_BL3)/3

Cor_BL       # Capacidade preditiva

##################################################
```

```
########Regressão Quantílica Regularizada#########
library(quantreg)
par(mfrow=c(5,4))
##Ajustando as RQ para diferentes quantis e lambdas
#lambda<-seq(0,53,.5) #para obter o melhor lambda para a RQ
#Tau<-seq(0.05,.95,.05)
acur<-matrix(NA,length(lambda),5)
result<-matrix(NA,length(Tau),3)
p<-0
for (k in Tau){
  p<-p+1
  j=0
  for (i in lambda)
    {
    j=j+1
    acur[j,1]<-k
    acur[j,2]<-i
    mod1<-rq(feno12~-1+geno12, tau=k, method="lasso", lambda=i)
    coef1<-mod1$coefficients
    pred1<-(geno3%*%coef1)
    cor1<-cor(pred1,feno3)
    write.table(pred1,"egbvbrq1.txt")
    write.table(coef1,"coefrq1.txt")
    write.table(cor1,"cor1.txt")
  acur[j,3]<-cor1/sqrt(0.20)
    mod2<-rq(feno13~-1+geno13, tau=k, method="lasso", lambda=i)
    coef2<-mod2$coefficients
```

```
    pred2<-(geno2%*%coef2)
    cor2<-cor(pred2,feno2)
     write.table(pred2,"egbvbrq2.txt")
    write.table(coef2,"coefrq2.txt")
    write.table(cor2,"cor2.txt")
   acur[j,4]<-cor2/sqrt(0.20)
    mod3<-rq(feno23~-1+geno23, tau=k, method="lasso", lambda=i)
    coef3<-mod3$coefficients
    #c<-coef3[order(coef3),]
    pred3<-(geno1%*%coef3)
    cor3<-cor(pred3,feno1)
   acur[j,5]<-cor3/sqrt(0.20)
   }
   print(acur)
   cormedia<-rowMeans(acur[,3:5])
   cormedia<-cbind(acur[,1:2],cormedia)
   plot(cormedia[,2], cormedia[,3],col="black",lty =1,lwd=2,type="l",
ylim=c(-0.1,0.5), ylab="Acurácia", xlab="Lambda")
   quant<-k
  legend("topright",paste("RQ =",quant),bty="n")
   max_rq<-max(cormedia[,3])
   lmax_rq<-cormedia[which(cormedia[,3]==max_rq),2]
   result[p,1]<-k
   result[p,2]<-max_rq
   result[p,3]<-lmax_rq
  }
  colnames(result) <- c("tau","acuracia","lambda")
```

```
result
######Coeficiente de correlação de postos de  Spearman####
blasso<-read.table("sperman ebacon.txt",h=T)
rq<-read.table("sperman ebacon sem blasso.txt",h=T)
cor(rank(rq[,1]),rank(blasso[,1]))
######################## Manhattan plot##############
b1<-read.table("betablasso3.txt",h=T)
b2<-read.table("betablasso2.txt",h=T)
b3<-read.table("betablasso1.txt",h=T)
blasso<-cbind(b1,b2,b3)
ahat_BL.rcarc<-rowMeans(blasso)
r1<-read.table("betarq3_0.1.txt",h=T)
r2<-read.table("betarq2_0.1.txt",h=T)
r3<-read.table("betarq1_0.1.txt",h=T)
rq<-cbind(r1,r2,r3)
ahat_rq.rcarc<-rowMeans(rq)
graf_rq.rcarc=cbind(mapa,ahat_BL.rcarc)
par(las=2, xpd=TRUE, cex.axis=0.9, cex=0.8,family="serif")
color=rep(c("blue","red"),5)
mhtprq.rcarc<-mhtplot(abs(graf_rq.rcarc[,1]),mht.control(logscale=
FALSE,colors=color,labels=c("1","4","7","8","17","x"),srt=0),xlab="Crom
ossomo",ylab="efeitoSNP",ylim=c(0,0.2),pch=12)
title("EBACON")
axis(2, at = seq(0, 0.2, by = 0.05))
```