LAÍS MAYARA AZEVEDO BARROSO

REGRESSÃO QUANTÍLICA: APLICAÇÕES EM SELEÇÃO GENÔMICA AMPLA

Tese 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 *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2018

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

T
B277r
2018

Barroso, Laís Mayara Azevedo, 1989-
Regressão quantílica : aplicações em seleção genômica ampla / Laís Mayara Azevedo Barroso. – Viçosa, MG, 2018.
x, 58f. : il. (algumas color.) ; 29 cm.

Orientador: Moysés Nascimento.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Biologia molecular. 2. Genômica - Seleção.
3. Biometria. 4. Regressão quantílica regularizada.
I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-graduação em Estatística Aplicada e Biometria.
II. Título.

CDD 22. ed. 572.8633

LAÍS MAYARA AZEVEDO BARROSO

REGRESSÃO QUANTÍLICA: APLICAÇÕES EM SELEÇÃO GENÔMICA AMPLA

Tese 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 *Doctor Scientiae*.

APROVADA: 02 de fevereiro de 2018.

Leonardo Siqueira Glória

Felipe Lopes da Silva

Camila Ferreira Azevedo

Fabyano Fonseca e Silva
Coorientador

Moysés Nascimento
Orientador

Aos meus pais, Adelson e Inêz,
e aos meus irmãos, Thiago, Lívia
e Maria Isabel por todo apoio,
carinho e dedicação.

AGRADECIMENTOS

Agradeço a Deus por ter me dado força para realização de mais este sonho, por ouvir minhas preces e por ter colocado anjos no meu caminho que muito me auxiliaram nesta etapa que chega ao fim.

Aos meus pais, Adelson e Inêz, por sonharem junto comigo e por serem minha base. Tudo o que sou hoje agradeço a vocês. Obrigada pelas orações, incentivos e apoio em todos os momentos. Vocês são meus heróis e exemplos.

Aos meus irmãos Thiago, Lívia e Maria Isabel por estarem sempre ao meu lado me apoiando e me fazendo dar muitas risadas. Por todos os momentos que vivemos que só nos ajudaram a crescer e aumentar ainda mais a nossa união.

Ao meu orientador e amigo Moysés Nascimento pelos conselhos, incentivo e confiança depositada na execução deste trabalho. Agradeço pela disponibilidade, atenção e amizade adquirida ao longo destes anos de convivência. Por me incentivar a cada dia mais e por ser um exemplo de profissional a ser seguido.

À professora, co-orientadora e amiga Ana Carolina Campana Nascimento pelas sugestões, incentivo, carinho e apoio.

Aos professores e co-orientadores Cosme Damião Cruz, Fabyano Fonseca e Silva e Marcos Deon Vilela de Resende pelos saberes transmitido, pela confiança, disponibilidade, incentivo e generosidade.

Ao professor Nick Serão pela oportunidade de estagiar em seu laboratório e pelos ensinamentos passados neste período.

Aos membros da banca examinadora, Prof. Doutor Fabyano Fonseca e Silva, Prof. Doutor Felipe Lopes da Silva, Profª. Doutora Camila Ferreira Azevedo, Prof. Doutor Leonardo Glória, Prof. Doutor Moyses Nascimento, pela disponibilidade e pelas valiosas sugestões para o enriquecimento deste trabalho.

Ao Leandro pelo companheirismo, carinho e apoio.

À amiga Camila pelas conversas e pelos conselhos. A Gabi por todos os momentos inesquecíveis, pela força e pela amizade. A Isabela pelas palavras de conforto e de motivação.

Aos meus amigos do Laboratório de Bioinformática e LICAE pelos ótimos momentos e pela valiosa amizade.

Aos amigos de Viçosa por sempre terem uma palavra de incentivo e apoiarem minhas decisões.

Às amigas de Bocaiúva pela amizade e por torcerem sempre por mim.

À República Sereníssimas, por ser minha família em Viçosa, por tornar essa caminhada mais tranquila e divertida. Serei eternamente grata a vocês.

Aos amigos de Raleigh, por tornarem meu intercambio tão especial.

Aos meus familiares pelo apoio.

Aos professores e funcionários do Departamento de Estatística da UFV, pela competência profissional e por todo apoio dado ao longo das minhas atividades acadêmicas.

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.

À todos que de uma maneira ou outra auxiliaram na concretização deste trabalho.

BIOGRAFIA

LAÍS MAYARA AZEVEDO BARROSO, filha de Maria Inez Azevedo Barroso e Antônio Adelson Barroso, nasceu em Bocaiúva, Minas Gerais, em 16 de março de 1989.

Em março de 2007, ingressou no curso de Licenciatura em Matemática na Universidade Federal de Viçosa, Viçosa - MG, graduando-se em janeiro de 2012.

Em março do mesmo ano, iniciou o curso de Mestrado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa da dissertação em 17 de fevereiro de 2014.

Em março de 2014, iniciou o curso de Doutorado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de tese em 02 de fevereiro de 2018.

SUMÁRIO

RESUMO viii

ABSTRACT x

INTRODUÇÃO GERAL 1

REVISÃO DE LITERATURA 4

1. Seleção genômica ampla (GWS) 4

2. Random Regression – Best Linear Unbiased Predictor (RR-BLUP) 4

3. Genomic - Best Linear Unbiased Predictor (G-BLUP) 5

4. Abordagens Bayesianas 6

4.1. Bayes A 7

4.2. Bayes B 7

4.3. Bayesian LASSO (BLASSO) 8

5. Regressão Quantílica (RQ) 9

6. Regressão Quantílica Regularizada (RQR) 13

7. Modelos Não Lineares para curvas de crescimento 14

8. Epistasia 15

9. Assimetria 16

REFERÊNCIAS BIBLIOGRÁFICAS 18

CAPÍTULO 1 23

Regularized quantile regression for SNP marker estimation of pig growth curves 23

Abstract 23

Background 24

Methods 25

Animals and genotyping data 25

Statistical analysis 25

Computational features 28

Results 28

Discussion 37
Conclusions 39
References 40
CAPÍTULO 2 44
Genomic prediction accuracies using regularized quantile regression methodology 44
Abstract 44
Background 44
Materials and Methods 46
Simulated genome (Population structure) 46
Simulated phenotypes 47
Statistical Analysis 48
Results and discussion 50
Conclusions 54
Reference 55
CONCLUSÕES GERAIS 58

RESUMO

BARROSO, Laís Mayara Azevedo, D.Sc., Universidade Federal de Viçosa, fevereiro de 2018. **Regressão Quantílica: aplicações em seleção genômica ampla.** Orientador: Moysés Nascimento. Coorientadores: Ana Carolina Campana Nascimento, Cosme Damião Cruz, Fabyano Fonseca e Silva, Marcos Deon Vilela de Resende e Nicola Vergara Lopes Serrão.

A principal contribuição da genética molecular no melhoramento é a utilização direta das informações de DNA no processo de identificação de indivíduos geneticamente superiores. Sob esse enfoque, idealizou-se a seleção genômica ampla (*Genome Wide Selection* – GWS), a qual consiste no uso de um grande número de marcadores SNPs (*Single Nucleotide Polymorphisms*) amplamente distribuídos no genoma para predizer o mérito genético de indivíduos. Diversas abordagens estatísticas foram propostas para a predição de valores genéticos permitindo estimar os efeitos dos marcadores com base apenas na média condicional da variável dependente. Uma metodologia ainda pouco explorada em GWS é a regressão quantilica (RQ). Diferentemente das outras metodologias, a RQ permite avaliar os fenótipos de interesse em diferentes níveis da distribuição. Desta forma, este trabalho tem como objetivo apresentar duas aplicações de GWS utilizando a RQ. Na primeira aplicação foi proposto e avaliado o uso da Regressão Quantílica Regularizada (RQR) para estimar os efeitos marcadores SNPs para curvas de crescimento em suínos. O modelo proposto permitiu a descoberta, em diferentes níveis de interesse (quantils), de marcadores relevantes para cada característica e suas respectivas posições cromossômicas. Além disso, RQR permitiu a construção de curvas de crescimento genômico, que identificaram indivíduos geneticamente superiores em relação à eficiência de crescimento. Na segunda aplicação utilizou-se a RQR para predizer valores genéticos de conjuntos de dados simulados com diferentes proporções de epistasia na variância genética e valores fenótipos com distribuições simétrica e assimétrica a direita. Neste trabalho verificou-se que a RQR teve, em geral, maiores acurácias do que as outras metodologias avaliadas quando a característica é de baixa herdabilidade. Além disso, quando tem-se 100% da variância genética como sendo epistática, a RQR foi, na maioria dos casos, melhor do que os métodos tradicionais. Desta forma, avaliando as duas aplicações apresentadas, tem-se que a RQR é uma alternativa interessante em estudos de GWS, uma vez que possibilita a descoberta do modelo que melhor representa a relação

entre as variáveis dependentes (fenótipos) e independentes (efeitos dos marcadores) aumentando o desempenho preditivo do modelo.

ABSTRACT

BARROSO, Laís Mayara Azevedo, D.Sc., Universidade Federal de Viçosa, February, 2018. **Quantile regression: applications in genome-wide selection.** Adviser: Moysés Nascimento. Co-advisers: Ana Carolina Campana Nascimento, Cosme Damião Cruz, Fabyano Fonseca e Silva, Marcos Deon Vilela de Resende and Nicola Vergara Lopes Serrão.

The main contribution of molecular genetics in breeding is the direct use of DNA information in the process of identifying genetically superior individuals. Under this approach, Genome Wide Selection (GWS) was idealized and consists of the use of a large number of single nucleotide polymorphisms (SNPs) widely distributed in the genome to predict the genetic merit of individuals. Several statistical approaches have been proposed for the prediction of genetic values, however they allow estimating the effects of the markers based only on the conditional mean of the dependent variable. A methodology not yet explored in GWS is quantile regression (QR). Differently from the other methodologies, the QR allows to evaluate the phenotypes of interest in different levels of the distribution. In this way, this work aims to present two applications of GWS using QR. In the first application, the Regulated Quantile Regression (RQR) was proposed and evaluated to estimate the marker effects SNPs for growth curves in pigs. The proposed model allowed the discovery, in different levels of interest (quantiles), of more relevant markers for each trait and their respective chromosomal positions. In addition, RQR allowed the construction of genomic growth curves, which identified genetically superior individuals in relation to growth efficiency. In the second application, the RQR was used to predict genetic values of simulated datasets with different proportions of epistasis in genetic variance and phenotype values with symmetric and positive asymmetric distributions. In this work it was verified that the RQR had, in general, greater accuracies than the other methodologies evaluated when the trait is low heritability. Furthermore, when 100% of the genetic variance is epistatic, RQR was, in most cases, better than traditional methods. Thus, RQR is an interesting alternative in GWS studies, since RQR allows the discovery of the model that best represents the relationship between the dependent (phenotype) and independent (markers effects) increasing the predictive performance of the model.

INTRODUÇÃO GERAL

Recentemente, devido ao avanço na genética molecular com o desenvolvimento de novas classes de marcadores, dentre os quais se destacam os SNPs (*Single Nucleotide Polymorphisms*), mapas densos estão disponíveis para diversas espécies de plantas e animais. Diante da abundância destes marcadores, Meuwissen et al. (2001) idealizaram a Seleção Genômica Ampla (*Genome Wide Selection* - GWS), a qual consiste em incorporar informações genômicas diretamente na predição do mérito genético individual para características de interesse econômico.

Uma das principais vantagens da GWS está relacionada com o intervalo de seleção, uma vez que a GWS viabiliza a seleção precoce aumentando assim o ganho genético por unidade de tempo. Essa vantagem merece destaque pois algumas características são medidas somente quando o indivíduo já está na fase adulta ou após o abate do mesmo. Além disso, com a GWS é possível utilizar um maior número de informações (fenotípica, genotípica e genealógica) para corrigir os dados e fazer análise genômica, fato que aumenta a acurácia (Resende et al. 2014).

Geralmente o número de marcadores é muito maior do que o número de indivíduos genotipados e fenotipados e tais marcadores são altamente correlacionados, necessitando da utilização de métodos estatísticos apropriados para análise dos dados no âmbito da genômica (Gianola et al., 2003).

Diversas abordagens estatísticas foram propostas para a predição de valores genéticos, como regressão aleatória (G-BLUP e RR-BLUP) e abordagens bayesianas (BLASSO, BayesA, Bayes B, ...). Embora essas metodologias sejam amplamente utilizadas, elas permitem estimar os efeitos dos marcadores com base na média condicional da variável dependente, ou seja, E(Y|X). Mais especificamente, caso as metodologias citadas acima para GWS sejam aplicadas, a relação funcional entre o fenótipo e os SNPs é explicada por meio de um comportamento médio, o que possibilita apenas o pesquisador selecionar indivíduos em termos médios da população.

Uma metodologia ainda pouco explorada em GWS é a regressão quantílica (RQ). Diferentemente das outras metodologias, a RQ, proposta por Koenker e Basset (1978) permite avaliar os fenótipos de interesse em diferentes níveis da distribuição. A obtenção destes efeitos em quantis específicos permite ao pesquisador selecionar indivíduos de acordo com o critério desejado, podendo obter assim um estudo mais informativo a respeito da variável analisada.

O estudo de diferentes níveis de distribuição da variável de interesse usando RQ foi realizado com sucesso na medicina por Beyerlein et al. (2011), que usaram RQ para análise GWAS (*Genome-Wide Association Study*) em genética humana, onde enfatizaram as vantagens estatísticas e biológicas ao estimar os efeitos dos marcadores em diferentes quantis da distribuição fenotípica. Nascimento et al. (2017) propuseram e avaliaram a utilização da RQ regularizada (RQR) na GWS, para tanto os autores simularam cenários com diferentes graus de assimetria nos fenótipos e com uso da RQR houve um aumento da performace preditiva do modelo.

Diante do exposto, o objetivo deste trabalho é apresentar a metodologia da Regressão quantílica e duas de suas aplicações. Desta forma, este trabalho está dividido em três partes: Revisão de Literatura, Capítulo 1 e Capítulo 2.

Na Revisão de Literatura, foi abordado conceitos, modelos e metodologias que serão utilizados posteriormente nos capítulos 1 e 2. Inicialmente é feita uma revisão sobre GWS e alguns modelos presentes na literatura. Posteriormente é discutida a metodologia proposta neste estudo, apresentando a RQ e a RQR. Logo após é feita uma breve revisão de modelos não lineares para curvas de crescimento que é o tema do primeiro capítulo desta tese. Finalmente, apresenta-se os conceitos de assimetria e epistasia que serão abordados no segundo capítulo.

No primeiro capítulo, buscou-se propor e avaliar RQR para estimar os efeitos marcadores SNPs para curvas de crescimento em suínos, bem como identificar as regiões cromossômicas dos marcadores mais relevantes e estimar a trajetória de peso individual genético ao longo do tempo (curva de crescimento genômico) em diferentes quantis. Com o objetivo de estimar os efeitos dos marcadores SNPs nas estimativas dos parâmetros das curvas de crescimento, utilizou-se uma abordagem em duas etapas. Na primeira etapa, os modelos não-lineares foram ajustados aos dados de peso-idade de cada animal. Na segunda, os modelos de regressão genômica, RQR, foram ajustados considerando as estimativas de parâmetros do passo anterior como variável dependente.

O segundo capítulo utilizou-se RQR para predizer valores genéticos em um conjunto de dados simulados com diferentes proporções de epistasia na variância genética quantitativa e fenótipos com distribuições simétrica e assimétrica a direita. O uso de RQR para obter a estimativa dos efeitos de marcadores permite um estudo mais informativo sobre os fenótipos, e com isso é possível trabalhar em qualquer quantil de interesse. Neste capítulo, também comparou-se o desempenho da RQR com outras metodologias presentes na literatura. Além disso, buscou-se verificar se é possível, com

a utilização da RQR, capturar os efeitos epistáticos, mesmo que estas interações não estejam explicitas no modelo.

Finalmente, apresentam-se as considerações finais do trabalho.

REVISÃO DE LITERATURA

1. Seleção genômica ampla (GWS)

Devido à ampla disponibilidade de marcadores SNPs no genoma, Meuwissen et al. (2001) propuseram a GWS com o objetivo de utilizar informações diretas do DNA na seleção e predição do mérito genético, de forma a permitir alta eficiência seletiva e grande rapidez na obtenção de ganhos genéticos com a seleção de baixo custo (Resende et al., 2014).

A GWS baseia-se na utilização de um grande número de marcadores SNPs (polimorfismo de um único nucleotídeo) que são amplamente distribuídos ao longo do genoma, possibilitando obter o mérito genético individual para características de interesse econômico. Desta forma, a GWS busca a determinação do valor genético genômico estimado (GEBV) de um indivíduo a partir de metodologias específicas para predição do mérito genético e seleção, podendo também utilizar o GEBV para uma seleção posterior, incluindo indivíduos que não foram fenotipados.

Geralmente o número de marcadores é muito maior do que o número de indivíduos que foram genotipados e fenotipados, tornando os marcadores altamente correlacionados, o que requer métodos estatísticos adequados que permitem propriedades de estimabilidade e regularização (Gianola et al., 2003). Para contornar essa limitação, várias abordagens estatísticas foram propostas para obtenção de valores genéticos genômicos, como regressão aleatória (RR-BLUP e G-BLUP), abordagens bayesianas (BLASSO, Bayes A, Bayes B, ...) e regressão de redução de dimensionalidade (PLS, componentes principais).

Nas próximas seções serão apresentados os principais modelos utilizados em estudos de Seleção Genômica Ampla.

2. *Random Regression – Best Linear Unbiased Predictor (RR-BLUP)*

A Regressão Aleatória - Melhor Preditor Linear não viesado (RR-BLUP) usa preditores do tipo BLUP e assume que os efeitos dos marcadores são covariáveis de efeitos aleatórios (Resende et al., 2014).

O modelo linear misto aleatório é dado por:

$$y = Xb + Wm_a + e,$$

em que: y é o vetor de fenótipos; b é o vetor de efeitos fixos; m_a é o vetor de efeitos aleatórios dos marcadores; e é o vetor de resíduos aleatórios; X e W são as matrizes de incidência para b e m_a. A matriz de incidência W contém os valores -1, 0 e 1 para o número de um dos alelos do marcador em um indivíduo diploide.

A predição dos efeitos genéticos dos marcadores através do RR-BLUP baseia-se nas equações dos modelos genômicos mistos apresentados abaixo:

$$\begin{bmatrix} X'X & X'W \\ W'X & W'W + I\lambda_a \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{m}_a \end{bmatrix} = \begin{bmatrix} X'y \\ W'y \end{bmatrix}$$

em que, $m_a \sim N(0, I\sigma_a^2 / n_q)$, $\lambda_a = \dfrac{\sigma_e^2}{(\sigma_a^2 / n_q)}$ é denominado parâmetro de encolhimento sendo σ_a^2 a variância genética aditiva, σ_e^2 a variância residual e n_q é o número de locus e é dado por $n_q = 2\sum_i^n p_i(1-p_i)$ (Resende et al., 2014). O parâmetro de penalidade λ_a promove o encolhimento homogêneo para todos os marcadores. Os componentes de variância (σ_e^2 e σ_a^2) são estimados através do método da Máxima Verossimilhança Restrita (REML).

3. *Genomic - Best Linear Unbiased Predictor* (G-BLUP)

O modelo linear misto a nível individual para valores genéticos aditivos individuais (u_a) é dado por:

$$y = Xb + Zu_a + e,$$

em que: y é o vetor de fenotípicos; b é o vetor de efeitos fixos com matriz de incidência X; u_a é o vetor de efeitos genéticos aditivos de indivíduos com matriz de incidência Z, sendo a estrutura de variância dada por $u_a \sim N(0, G_a\sigma_a^2)$ onde σ_a^2 é variância aditiva e G_a (N × N) é a matriz de parentesco genômica aditiva; e é o vetor residual do modelo com $e \sim N(0, I\sigma_e^2)$ sendo que σ_e^2 é a variância residual e I a matriz de identidade.

Utilizando as equações de modelo misto, podemos predizer por meio do método G-BLUP conforme a seguir:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G_a^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u}_a \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}.$$

Para que as equações acima tenham solução, G_a deve ser uma matriz positiva definida e, da mesma forma que no RR-BLUP os componentes de variância (σ_e^2 e σ_a^2) são estimados por meio do método da Máxima Verossimilhança Restrita (REML).

A matriz de parentesco genômico (G_a) descreve o parentesco entre indivíduos e pode ser construída a partir do conjunto de marcadores pelo qual os indivíduos foram genotipados e é dada por:

$$G_a = \frac{WW'}{\sum_{i=1}^{n} 2p_i(1-p_i)}$$

em que W é uma matriz de dimensão *n x m*, sendo que *n* corresponde ao número de indivíduos e *m* o número de marcadores e p_i são as frequências alélicas.

A análise G-BLUP é favorável em termos computacionais, porque resulta em um número menor de equações a serem resolvidas (Resende et al., 2014).

4. Abordagens Bayesianas

Os métodos bayesianos teoricamente proporcionam uma maior acurácia (Meuwissen et al., 2001; Bolormaa et al., 2013; Meuwissen et al., 2016), uma vez que forçam muitos efeitos dos segmentos cromossômicos a valores próximos de 0 (Bayes A) ou 0 (Bayes B) e os efeitos estimados são ditados pelas distribuições a *priori* dos efeitos dos marcadores (Resende et al., 2014).

A predição de valores genéticos genômicos, utilizando informações fenotípicas e genotípicas para cada indivíduo, pode ser estimada usando o seguinte modelo:

$$y = Xb + Wm + e,$$

em que: y é o vetor de fenótipos; X é um vetor de mesma dimensão de y com todos os elementos iguais a 1; b é a média da característica de interesse; m é o vetor de efeitos genéticos aditivos dos marcadores; W é a matriz de incidência que relaciona os efeitos aditivos dos marcadores aos fenótipos; e é o vetor de erro aleatório do modelo com $e \sim N(0, \sigma_e^2)$ e σ_e^2 é a variância do erro.

4.1. Bayes A

Meuwissen et al. (2001) propuseram uma metodologia baseada na abordagem bayesiana em que diferentes componentes de variância são assumidos para cada marcador. O método Bayes A, por meio de suas distribuições a *priori*, permite um fator de encolhimento para cada marcador.

Este método pressupõe poucos genes de grandes efeitos e muitos genes de pequenos efeitos, ou seja, encolhimento específico para cada marcador. A distribuição *priori* de cada efeito de marcador (dada a sua variância) segue uma distribuição normal, $m_j \left| \sigma^2_{m_j} \right. \sim N(0, \sigma^2_{m_j})$, e as variâncias dos efeitos marcadores assumem distribuição qui-quadrado invertida escalada com graus de liberdade v_m e parâmetro de escala S^2_m, ou seja, $\sigma^2_{m_j} \sim \chi^{-2}\left(v_m, S^2_m\right)$ (Azevedo, 2015).

A escolha de uma distribuição de qui-quadrado invertida para componentes de variância é conveniente uma vez que tal distribuição combinada com a informação dos dados também resulta em uma distribuição qui-quadrado invertida escalada (Meuwissen et al., 2001).

O uso dessas duas distribuições (normal e qui-quadrada invertida) leva à distribuição t univariada para os efeitos dos marcadores com média zero (Sorensen e Gianola, 2002), da seguinte forma:

$$m_j \left| v_m, S^2_m \sim t\left(0, v_m, S^2_m\right). \right.$$

Este método pode ser implementado através da amostragem de Gibbs para obter essa informação combinada (Resende et al., 2014). A amostragem de Gibbs baseada nas distribuições condicionais *a posteriori* de todos os efeitos é utilizada para obter amostras que serão usadas para estimar os efeitos e a variância.

4.2. Bayes B

Embora útil, o método Bayes A não permite a seleção de variáveis (SV). A SV é interessante uma vez que muitos marcadores não possuem efeitos genéticos ou não estão em desequilíbrio de ligação com os QTLs. Desta forma, permite que alguns efeitos de marcadores sejam iguais a zero, levando a uma condição estatística mais favorável, ou seja, quanto menor o número de efeitos de marcadores a serem estimados, mais preciso é o processo de estimação considerando o mesmo número de observações.

O método Bayes B assume uma mistura de distribuições para o efeito do marcador, ou seja, enquanto uma fração π dos marcadores assume distribuição normal (os mesmos pressupostos do método Bayes A), a fração $1-\pi$ dos marcadores não tem efeito. No método de Bayes B, o valor de π é definido subjetivamente. Especificamente, o método Bayes B assume as seguintes distribuições *priori* para cada efeito marcador:

$m_j \left| \pi, \sigma^2_{m_j} \right. \sim \pi\, N(0, \sigma^2_{m_j}) + (1-\pi) N(0, \sigma^2_{m_j} = 0)$, sendo $\sigma^2_{m_j}$ a variância específica do j[th] marcador.

As variâncias dos efeitos dos marcadores são assumidas $\sigma^2_{m_j} = 0$ com probabilidade $1-\pi$ e distribuição qui-quadrado invertida escalada com v_m graus de liberdade e parâmetro de escala S^2_m, isto é, $\sigma^2_{m_j} \sim \chi^{-2}\left(v_m, S^2_m\right)$ com probabilidade π.

As distribuições condicionais completas para o método Bayes B não possuem distribuições de probabilidade conhecidas. Assim, é necessário usar o algoritmo Metropolis-Hastings (Gelman et al., 2004) para gerar amostras sequenciais como meio de obter uma distribuição a partir da qual não há amostragem direta. Mais detalhes sobre o método Bayes B podem ser encontrados em Azevedo (2015). As distribuições condicionais completas para os parâmetros do modelo Bayes B foram apresentadas em detalhes por Zeng et al. (2013).

4.3. Bayesian LASSO (BLASSO)

O uso da versão bayesiana da regressão LASSO (BLASSO) para seleção genômica foi proposta por de los Campos et al. (2009). O BLASSO inclui um termo de variância comum para modelar os resíduos e os efeitos genéticos dos marcadores. Na formulação do BLASSO são consideradas as seguintes distribuições a *priori* (Park e Casella, 2008; de los Campos et al., 2009):

- $m_j \left| \sigma^2_e, \tau^2_j \right. \sim N(0, \tau^2_j \sigma^2_e)$
- $\tau^2_j \left| \lambda^2 \right. \sim Exp\,(\lambda^2)$
- $\lambda^2 \sim Gamma\,(\alpha_1, \alpha_2)$

em que: λ é o parâmetro de encolhimento e α_1 e α_2 são os hiperparâmetros da distribuição gama. O parâmetro λ pode ser estimado a partir dos dados por métodos MCMC.

A distribuição a *priori* utilizada no LASSO Bayesiano mostra maior massa de densidade em zero e caudas mais robustas, exercendo maior encurtamento sobre coeficientes de regressão próximos de 0 e menor encurtamento sobre coeficientes de regressão distantes de zero. Assim, as médias a *posteriori* são estimadas, produzindo valores muito pequenos, mas não zero como no LASSO original.

Como definido anteriormente, a junção de uma distribuição normal e de uma exponencial leva a uma distribuição exponencial dupla para os efeitos dos marcadores (Park e Casella, 2008), da seguinte forma:

$$m_j \left| \lambda^2 \sim DuplaExp\left(0, \frac{\sigma_e}{\lambda}\right)\right.$$

O Lasso Bayesiano é vantajoso em comparação com Bayes B e Bayes A (Meuwissen et al., 2001) uma vez que minimiza a influência da distribuição a *priori*, ou seja, proporciona uma melhor aprendizagem a partir de dados (Gianola et al., 2009; Gianola, 2013). Os detalhes das distribuições condicionais completas para os parâmetros do modelo BLASSO foram apresentados em de los Campos et al. (2009).

5. Regressão Quantílica (RQ)

Nesta seção serão apresentados os conceitos e fundamentos da Regressão Quantílica (RQ). Mais detalhes a respeito da metodologia pode ser encontrado em Barroso (2014) e Hao e Naiman (2007).

A RQ, diferentemente da abordagem tradicional de Modelos de Regressão, possibilita obter informações da variável de interesse em diversos níveis da distribuição, tornando o estudo sobre o fenômeno mais completo e informativo. A RQ pode ser usada para fornecer uma análise estatística mais completa das relações entre variáveis aleatórias. Em geral, os quantils escolhidos dependem inteiramente da finalidade do estudo, ou seja, podemos estudar toda a distribuição ou apenas algumas partes definindo quantils específicos.

Os modelos de regressão tradicionais, apesar da facilidade de interpretação e implementação, apresenta algumas limitações. Em primeiro lugar, os pressupostos do modelo nem sempre são válidos para conjuntos de dados reais. A pressuposição de homocedasticidade (Figura 1) geralmente falha. Desta forma, quando excluímos o pressuposto de homocedasticidade, embora os estimadores de mínimos quadrados sejam ainda não tendenciosos os mesmos não possuem variância mínima (Montgomery et al.,

2012). Assim, os estimadores obtidos por meio do método dos mínimos quadrados não serão BLUE (*Best Linear Unbiased Estimator*).

Quando a distribuição apresenta caudas pesadas, a média condicional pode se tornar uma medida inadequada e enganosa de localização central, pois é fortemente influenciada por *outliers*.

Além disso, o cálculo dos p-valores se baseiam no pressuposto de normalidade e caso exista violação desta condição pode ocorrer viés nos p-valores, acarretando em testes de hipóteses inválidos (Hao e Naiman, 2007).

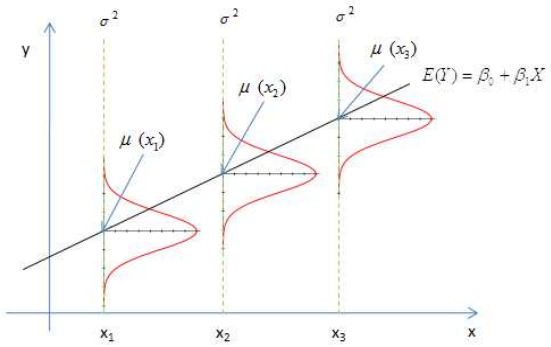


Figura 1. Representação do modelo de regressão linear simples com erros homocedásticos.

Para contornar a limitação referente à distribuição normal dos erros, pode-se utilizar o método de minimização dos erros absolutos, uma vez que este é robusto na presença de *outliers* e assimetria e descreve melhor uma medida de posição central da distribuição condicional da variável resposta, uma vez que estima o valor mediano da distribuição.

A RQ, proposta por Koenker e Bassett (1978) se baseia no método dos erros absolutos ponderados. Entretanto, nesta metodologia não se considera apenas o valor mediano, e sim é realizada uma ponderação na minimização dos erros para se estimar os diversos quantis de interesse.

A utilização de diversos quantis possibilita a obtenção de maiores informações de localização do que quando se utiliza apenas o centro da distribuição. Desta forma é possível examinar uma localização na cauda inferior (por exemplo, o quantil 0,1) ou na cauda superior (por exemplo, o quantil 0,9) quando o pesquisador necessita de informações sobre subpopulações específicas (Hao e Naiman, 2007). Como por exemplo,

nos estudos envolvendo problemas econômicos (Silva e Porto Junior, 2006), medicina humana (Beyerlein et al., 2011) e problemas sociais (Hao e Naiman, 2007).

Além disso, de acordo com Koenker (2005), modelos de RQ são capazes de incorporar uma possível heterocedasticidade, que seria detectada a partir da variação das estimativas dos coeficientes dos parâmetros para diferentes quantis ($\tau's$).

O modelo da Regressão Quantílica, que descreve a relação funcional entre a variável dependente e a variável independente, pode ser descrito como:

$$y = \beta_0(\tau) + \beta_1(\tau)X + e(\tau), \quad (1)$$

em que $\beta_0(\tau)$ é a constante da regressão; $\beta_1(\tau)$ é o coeficiente da regressão; $e(\tau)$ são os erros aleatórios independentes e identicamente distribuídos com quantil de ordem τ igual a zero; X é variável independente e τ refere-se ao quantil assumido $(\tau \in (0,1))$.

Nos modelos de regressão linear simples (RLS) estima-se apenas uma reta para explicar todo o conjunto de dados. Entretanto, a linha da regressão não captura mudanças na distribuição da variável dependente. Já no modelo da Regressão Quantílica (1) podem ser estimadas retas para cada quantil de interesse, desta maneira se torna mais adequado à interpretação dos resultados para o conjunto de dados com presença de assimetria, pois através dela é possível traçar a relação em regiões centrais, através da mediana, e nas caudas da distribuição condicional de acordo com o interesse.

Pode-se observar que a RLS fornece apenas informações em termos médios da variável dependente (Figura 2A). Enquanto que a RQ possibilita observar a relação funcional em diferentes níveis da variável dependente (Figura 2B).

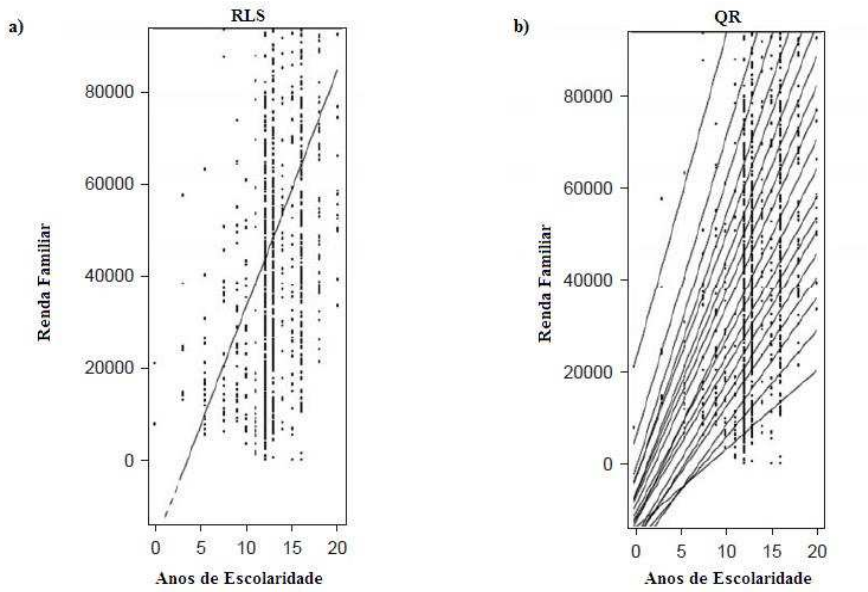


Figura 2. Ajuste de um modelo linear e diversos ajustes da regressão quantílica.
Fonte: Hao e Naiman (2007).

A principal diferença entre as estimações da RQ e da RLS é que a distância dos pontos observados a reta estimada na RQ é medida minimizando a média ponderada da soma das distâncias verticais, sendo que para pontos abaixo da linha atribui-se peso $1-\tau$, e para pontos acima da linha peso τ, conforme apresentado em Hao e Naiman (2007).

Cada escolha do valor do peso τ resultará numa função ajustada do quantil condicional. Assim, tem-se que um quantil τ pode ser visto como o ponto em que minimiza a distância média ponderada, com pesos dependendo da localização do ponto, se o ponto está acima ou abaixo da reta ajustada.

O objetivo então é encontrar os valores de $\beta_i(\tau)$ que minimizem a equação:

$$\sum_{i=1}^{n} d_\tau(y_i, \hat{y}_i) = \tau \sum_{y_i \geq \beta_{0i}(\tau)+\beta_{1i}(\tau)X_i} \left|y_i - \beta_{0i}(\tau) - \beta_{1i}(\tau)X_i\right| + (1-\tau) \sum_{y_i < \beta_{0i}(\tau)+\beta_{1i}(\tau)X_i} \left|y_i - \beta_{0i}(\tau) - {}_i\beta_{1i}(\tau)X_i\right| \qquad (2)$$

em que d_τ é a distância entre y_i e $\hat{y}_i$; $\beta_{0i}(\tau)$ é a constante da regressão; $\beta_{1i}(\tau)$ é o coeficiente da regressão; X é variável independente e τ refere-se ao quantil assumido $(\tau \in (0,1))$.

Ao minimizar a Equação 2 tem-se como resultados a reta da regressão do quantil de interesse.

Um algoritmo que permite minimizar a Equação 2, estimando os coeficientes da RQ $\left(\hat{\beta}_0, \hat{\beta}_1\right)$, é baseado em algoritmos de problemas de programação linear (Hao e Naiman, 2007). O método utilizado para solução de problemas de programação linear é o Método Simplex. O principal objetivo neste método é obter o ponto que corresponde ao menor valor da equação 2.

Deste modo é necessário encontrar os estimadores $\hat{\beta}_{0i}(\tau)$ e $\hat{\beta}_{1i}(\tau)$ que minimizem a soma dos erros absolutos ponderados da equação 2, ou seja, deve-se minimizar a soma dos resíduos $y_i - \hat{y}_i$ de maneira que resíduos com valores positivos recebem peso τ e resíduos negativos recebem peso $1-\tau$.

6. Regressão Quantílica Regularizada (RQR)

Como mencionado anteriormente, RQ é baseada em quantils condicionais, Q(Y|X), isto é, é possível ajustar modelos para toda a distribuição de probabilidade da característica avaliada, permitindo um estudo mais informativo da relação entre variáveis.

De acordo com Nascimento et al. (2017), a RQ facilita a seleção de uma função quantilíca que "melhor" represente a relação entre as variáveis dependente (fenotípos) e independente (marcadores) para resolver o problema da assimetria.

Quando combinamos as propriedades dos métodos tradicionais de GWS (por exemplo, a teoria de estimação de encolhimento) com características desejáveis da RQ, geramos a regressão quantílica regularizada (RQR), que é um método de predição novo e poderoso que pode resolver problemas relacionados à dimensionalidade, multicolinearidade e distribuição fenotípica assimétrica.

Li e Zhu (2008) propuseram a RQR que usa a soma dos valores absolutos dos coeficientes como penalidade.

Os métodos de penalização consistem em encolher as estimativas de coeficientes em relação a zero relacionado com às estimativas de mínimos quadrados. Esses métodos levam a estimativas estáveis, permitindo a estimativa dos parâmetros no caso n >> N, onde n é o número covariáveis e N é o número de observações, e quando há multicolinearidade entre as variáveis. Assim, esta propriedade é muito importante para o caso de alta dimensionalidade.

O RQR consiste em obter os efeitos marcador $\left(\beta_j\right)$ que solucionam o seguinte problema de otimização:

$$\hat{\beta} = \text{argmin}_{\beta} \left\{ \sum_{i=1}^{n} \rho_{\tau} \left[y_i - \left(\mu + \sum_{j=1}^{p} x_{ij} \beta_j \right) \right] + \lambda \sum_{i=1}^{p} \left| \beta_j \right| \right\} \quad (3)$$

em que: $\sum_{i=1}^{p} \left| \beta_j \right|$ é a soma dos valores absolutos dos coeficientes de regressão, λ é o parâmetro que controla a força da regularização, e ρ_{τ} é a função que pondera as observações para a estimativa da função quantilíca (Koenker e Bassett, 1978) chamada função de *check* e é definida como:

$$\rho_{\tau} \left[y_i - \left(\mu + \sum_{j=1}^{p} x_{ij} \beta_j \right) \right] = \begin{cases} \tau \left[y_i - \left(\mu + \sum_{j=1}^{p} x_{ij} \beta_j \right) \right], & \text{se} \quad y_i - \left(\mu + \sum_{j=1}^{p} x_{ij} \beta_j \right) > 0, \\ -(1-\tau) \left[y_i - \left(\mu + \sum_{j=1}^{p} x_{ij} \beta_j \right) \right], & \text{caso contrário.} \end{cases}$$

em que $\tau \in (0,1)$ indica o quantil de interesse. Através do coeficiente de encolhimento, esta metodologia combina a seleção de variáveis e a regularização através do encurtamento dos coeficientes de regressão

7. Modelos Não Lineares para curvas de crescimento

Para descrever a relação entre o peso e a idade dos animais é utilizada as curvas de crescimento. Em geral, o estudo das curvas de crescimento é realizado ajustando modelos não-lineares utilizando o peso como variável dependente e idade como variável independente. Esses modelos são usados porque são flexíveis e apresentam um número reduzido de parâmetros com interpretações biológicas, como por exemplo, a taxa de maturidade e peso do adulto (Silva et al., 2017).

Um resumo das equações comumente usadas é apresentado na Tabela 1.

Tabela 1. Um resumo das equações de crescimento.

Referência	Nome	Equação
Gompertz (1825)	Gompertz	$y = \beta_1 e^{\left(-\beta_2 e^{(-\beta_3 x)}\right)} + \varepsilon$
Ratkowsky (1983)	Logistic	$y = \frac{\beta_1}{1 + \beta_2 e^{(-\beta_3 x)}} + \varepsilon$
von Bertalanffy (1957)	von Bertalanffy	$y = \beta_1 \left(1 - \beta_2 e^{(-\beta_3 x)}\right)^3 + \varepsilon$

em que β_1 é o peso assintótico; β_2 é um parâmetro de localização, sem interpretação biológica; β_3 é o parâmetro da taxa de maturidade; y é a observação da variável resposta; x é a idade e ε é o termo de erro aleatório.

8. Epistasia

A partição clássica da variância genética para características quantitativas é dada em termos de componentes aditivos, dominantes e epistasia. Embora a variância aditiva determine correlações entre pais e resposta à seleção, o estudo de componentes de epistasia (Mackay, 2014) e dominância (Azevedo, 2015) também é muito importante.Diversos estudos destacaram a presença da epistasia na arquitetura genética das características quantitativas (Phillips, 2008; Mackay, 2014; Huang e Mackay, 2016).

Bateson (1909) definiu a epistasia quando o efeito do alelo em um locus, ao se expressar, esconde ou mascara o efeito de outro alelo em um segundo locus. Esta definição é conhecida na literatura como epistasia biológica. A epistasia estatística foi proposta por Fisher (1918) e refere-se a qualquer desvio da combinação aditiva de dois loci em relação à sua contribuição para um fenótipo quantitativo.

Usando a fórmula clássica para fenótipos, é possível representar um modelo de dois locus como:

$$F = \mu + G_A + G_B + G_{A*B} + \varepsilon,$$

em que: F é o valor fenotípico, μ é a média da população, G_A e G_B são os valores genotípicos nos locus A e B, respectivamente, G_{A*B} representa a interação entre locus A e B, ε é o valor ambiental.

A variância epistática pode ser decomposta em seus componentes aditivo x aditivo, aditivo x dominante e dominante x dominante (Cockerhan, 1954; Kempthorne, 1954).

É possível expandir a equação fenotípica, considerando apenas efeitos aditivos, para:

$$F = \mu + A_A + A_B + A_A * A_B + \varepsilon,$$

em que A_A é o efeito aditivo para locus A; A_B é o efeito aditivo para locus B. O valor epistático ($A_A * A_B$) descreve a interação entre locus.

Considerando um locus bialélico (A_{11} , A_{12} e A_{22} , por exemplo), sob epistasia aditiva-aditiva ($A_A * A_B$), o valor aditivo do locus A muda dependendo do genótipo do locus B e vice-versa (Wolf et al., 2000).

Tabela 2. Genótipos que representam a epistasia aditiva-aditiva.

Genótipo	A_{11}	A_{12}	A_{22}
B_{11}	1	0	-1
B_{12}	0	0	0
B_{22}	-1	0	1

Neste tipo de epistasia, o genótipo 22 do locus A resulta em valores menores quando combinados com o genótipo 11 do locus B e valores maiores com o genótipo 22 do locus B (Tabela 2 e Figura 3).

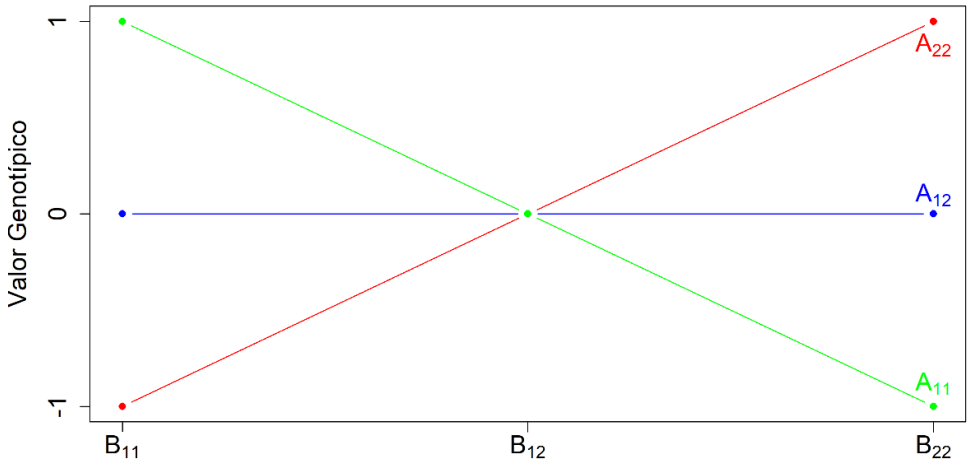


Figura 3. Representação gráfica da epistasia aditiva-aditiva.

A presença de epistasia em animais e plantas está em vários padrões clássicos de genótipo-fenótipo, como a cor da pele em vários animais; tipo de crista em galinhas; cor de semente no trigo; forma da asa e velocidade do vôo em *Drosophila melanogaster*; peso corporal e tamanho da ninhada em camundongos; "rat-tail" fenótipo no gado.

9. Assimetria

A assimetria descreve qual lado de uma distribuição tem uma cauda mais longa. Se a cauda longa estiver à direita, então a assimetria é considerada a direita ou positiva (Figura 4, b); se a cauda longa estiver à esquerda, então a assimetria é considerada a

esquerda ou negativa (Figura 4, c).

Matematicamente, a assimetria é dada pela razão do momento amostral de ordem 3 pelo de ordem 2, na potência de 3/2:

$$As = \frac{m_3}{(m_2)^{3/2}},$$

em que $m_r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^r}{n}$, r é a ordem do momento, $\overline{X}$ é a média da amostra, n é o tamanho amostral.

Distribuições simétricas, como a apresentada na Figura 4 (a), possuem $As = 0$.

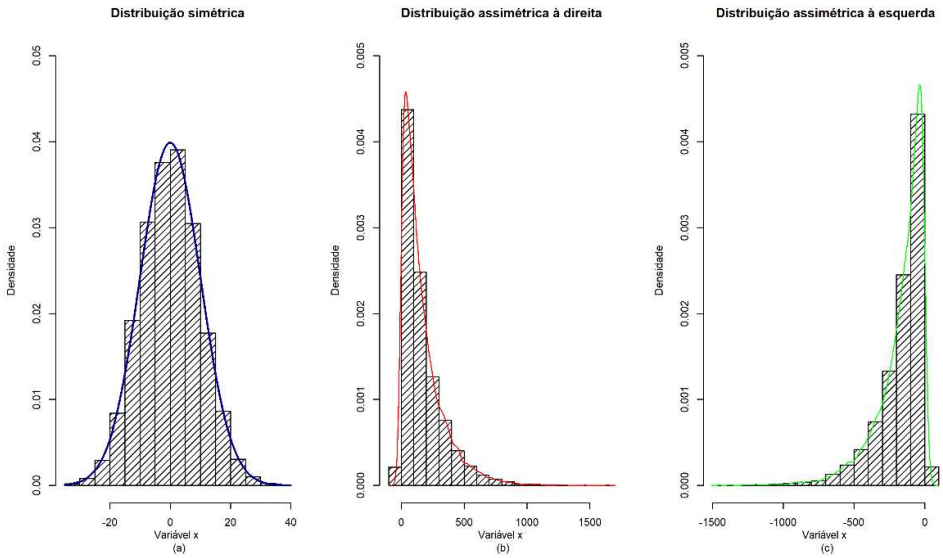


Figura 4. Possíveis distribuições para o fenótipo. (a) Distribuição simétrica. (b) Distribuição assimétrica a direita. (c) Distribuição assimétrica a esquerda.

Existem distribuições assimétricas para várias características de plantas e animais. Na criação de gado leiteiro, número de tetas, taxas de distocia, contagem de células somáticas do leite (SCC) são exemplos de fenótipos assimétricos muito importantes. A produção de ovos e a idade até a produção do primeiro ovo são exemplos de fenótipos assimétricos em frango. A ordem de parto (Varona et al., 2008) e as concentrações de hormônios (Campos et al., 2015) não são normalmente distribuídos em suínos.

De acordo com Roy (2000), há muitas situações que podem acarretar a assimetria da variável, uma delas é quando há epistasia. A Figura 5 mostra a relação entre assimetria e epistasia. A assimetria é positiva quando há epistasia complementar (Figura 5 - b) e negativa para epistasia duplicada (Figura 5 - c) (Snape e Riggs, 1975).

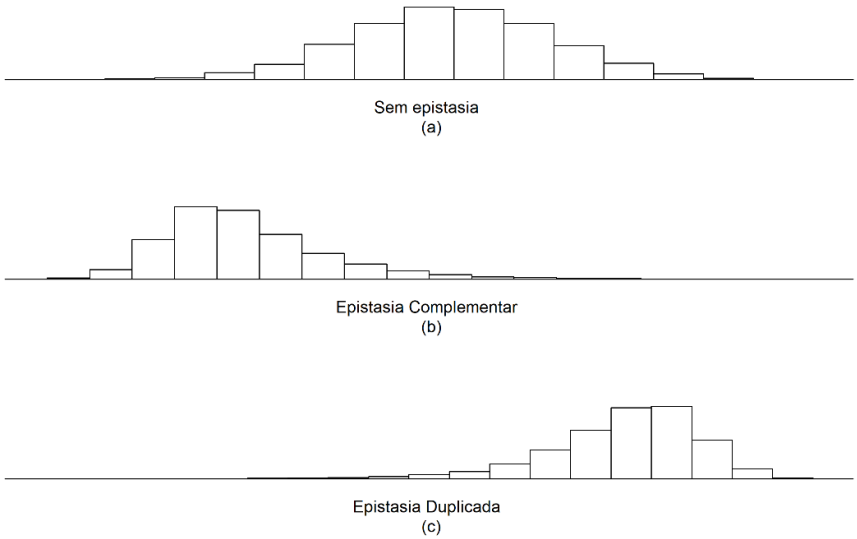


Figura 5. Distribuições para situações em que não existe epistasia (a), epistasia complementar (b) e epistasia duplicada (c).

REFERÊNCIAS BIBLIOGRÁFICAS

AZEVEDO, C.F. Ridge, lasso and bayesian additive-dominance genomic models and new estimators for the experimental accuracy of genome selection. Tese de doutorado. Universidade Federal de Viçosa. 2015.

BARROSO, L.M.A. Regressão quantílica na avaliação da adaptabilidade e estabilidade fenotípica. Dissertação de Mestrado. Universidade Federal de Viçosa. 2014.

BATESON, W. **Mendel's Principles of Heredity**. Cambridge Univ. Press, Cambridge, 1909.

BEYERLEIN, A.; VON KRIES, R.; NESS, A.R.; ONG, K.K. Genetic Markers of Obesity Risk: Stronger Associations with Body Composition in Overweight Compared to Normal-Weight Children. **PLoS ONE**, v.6, e19057, 2011.

BOLORMAA, S.; PRYCE, J.E.; KEMPER, K.; SAVIN, K.; HAYES, B.J.; BARENDSE, W; ZHANG, Y.; REICH, C. M.; MASON, B. A.; BUNCH, R. J.; HARRISON, B. E.; REVERTER, A.; HERD, R. M.; TIER, B.; GRASER, H.-U.; GODDARD, M. E.

Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in *Bos taurus*, *Bos indicus*, and composite beef cattle. **J. Anim. Sci.** v. 91, p. 3088–3104, 2013.

CAMPOS, C.F.; LOPES, M.S.; SILVA, F.F.; VERONEZE, R.; KNOL, E.F.; LOPES, P.S.; GUIMARÃES, S.E. Genomic selection for boar taint compounds and carcass traits in a commercial pig population**. Livestock Science**. v.174, p. 10-17, 2015.

COCKERHAM, C.C. An Extension of the Concept of Partitioning Hereditary Variance for Analysis of Covariances among Relatives When Epistasis Is Present. **Genetics**, v. 39, p. 859-882, 1954.

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 Markers and Pedigree. **Genetics**, v. 182, p. 375-385, 2009.

FISHER, R.A. The Correlation between Relatives on the Supposition of Mendelian Inheritance. **Trans. R. Soc. Edin**, v. 52, p. 399-433, 1918.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, D. B. **Bayesian Data Analysis**. Chapman & Hall, London. 2004.

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.

GIANOLA, D.; DE LOS CAMPOS, G.; HILL, W. G.; MANFREDI, E.; FERNANDO, R. Additive genetic variability and the Bayesian alphabet. **Genetics**, v. 183, p.347-363, 2009.

GIANOLA, D. Priors in whole-genome regression: the bayesian alphabet returns. **Genetics**, v.194, p.573-96, 2013.

GOMPERTZ, B.; PHILOS, T.; On the nature of the function expressive of the law of

human mortality, and on a new mode of determining the value of life contingencies. **Royal Society of London**, v.115, p.513 585, 1825.

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

HUANG, W; MACKAY, T.F.C. The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis. **PLoS Genet,** 12(11), e1006421, 2016.

KEMPTHORNE, O. The correlation between relatives in a random mating population. **Proceedings R Soc Biol Sci**, v. 143, p. 103-113, 1954.

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

KOENKER, R. **Quantile Regression**. 1.ed. Cambridge University Press, v.1. 349p, 2005.

LI, Y.; ZHU, J. L1-Norm Quantile Regression. **Journal of Computational and Graphical Statistics**, v. 17, p.1-23, 2008.

MACKAY, T.F.C. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. **Nat Rev Genet.**, v. 15, p. 22–33, 2014.

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

MEUWISSEN, T. H. E.; HAYES B. J.; GODDARD M. E. Genomic selection: A paradigm shift in animal breeding. **Animal Frontiers**, v. 6, p. 6–14, 2016.

MONTGOMERY, D.C.; PECK, E.A.; VINING, G.G. **Introduction to linear regression analysis**. 5. ed. New York: John Wiley & Sons, 645p., 2012.

NASCIMENTO, M.; SILVA, F. F.; RESENDE, M.D.V.; CRUZ, C. D.; NASCIMENTO, A. C. C.; VIANA, J.M.S.; AZEVEDO, C.F.; BARROSO, L.M.A. Regularized quantile regression applied to genome-enabled prediction of quantitative traits. **Genetics and Molecular Research**, v.16, 2017.

PARK, T.; CASELLA, G. The Bayesian LASSO. **Journal of the American Statistical Association**, v.103, p.681-686, 2008.

PHILLIPS, P.C. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. **Nat Rev Genet**. v.9, p. 855–867, 2008.

RATKOWSKY, D. A. **Nonlinear regression modeling: a unified practical approach.** Marcel Dekker, New York. 1983.

RESENDE, M. D. V. de; SILVA, F. F. e; AZEVEDO, C. F. **Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão aleatória, seleção genômica, QTL- GWAS, estatística espacial e temporal, competição, sobrevivência**. Viçosa, MG: UFV, 2014. 881 p.

ROY, D. Plant Breeding: Analysis and Exploitation of Variation. Alpha Science International, 701 p. 2000.

SILVA, E. N da; PORTO JÚNIOR, S. S. Sistema financeiro e crescimento econômico: Uma aplicação de regressão Quantílica. **Econ. Aplic**., v. 10, p. 425-442, 2006.

SILVA, F.F.; ZAMBRANO, M.F.B.; VARONA, L.; GLORIA, L.S.; LOPES, P.S.; SILVA, M.V.G.B.; ARBEX, W.; LÁZARO, S.F.; RESENDE, M.D.V.; GUIMARÃES, S.E.F. Genome association study through nonlinear mixed models revealed new candidate genes for pig growth curves. **Sci. Agric**. v.74, n.1, p.1-7, 2017.

SNAPE, J. W.; RIGGS, T.J. Genetical consequences of single seed descent in the breeding of self pollinating crops. **Heredity,** v.35, 211-219, 1975.

SORENSEN, D.; GIANOLA, D. **Likelihood, Bayesian and MCMC methods in quantitative genetics**. New York: Springer Verlag, 2002, 740 p.

VARONA, L.; IBAÑEZ-ESCRICHE, N.; QUINTANILLA, R.; NOGUERA, J.L.; CASELLAS, J.. Bayesian analysis of quantitative traits using skewed distributions. **Genet. Res**.v. 90, p. 179-190, 2008.

VON BERTALANFFY, L. Quantitative laws for metabolismo and growth. **Q. Rev. Biol**. v. 32, p. 217-231, 1957.

WOLF, J. B.; BRODIE, E. D. I.; WADE, M. J. **Epistasis and the Evolutionary Process***,* Oxford, New York, 2000.

ZENG, J.; TOOSI, A.; FERNANDO, R. L.; DEKKERS, J. C. M.; GARRICK, D. J. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. **Genetics Selection Evolution**, 45:11, 2013.

CAPÍTULO 1

Published as original paper to Journal of Animal Science and Biotechnology

Reference: Barroso, L.M.A.; Nascimento, M.; Nascimento, A.C.C.; Silva, F.F.; Serão, N.V.L.; Cruz, C.D.; Resende, M.D.V.; Silva, F.L.; Azevedo, C.F.; Lopes, P.S.; Guimarães, S.E.F. Regularized quantile regression for SNP marker estimation of pig growth curves. *Journal of Animal Science and Biotechnology* (2017) 8:59
DOI: 10.1186/s40104-017-0187-z

Regularized quantile regression for SNP marker estimation of pig growth curves

Abstract: Genomic growth curves are generally defined only in terms of population mean; an alternative approach that has not yet been exploited in genomic analyses of growth curves is the Quantile Regression (QR). This methodology allows for the estimation of marker effects at different levels of the variable of interest. We aimed to propose and evaluate a regularized quantile regression for SNP marker effect estimation of pig growth curves, as well as to identify the chromosome regions of the most relevant markers and to estimate the genetic individual weight trajectory over time (genomic growth curve) under different quantiles (levels). The regularized quantile regression (RQR) enabled the discovery, at different levels of interest (quantiles), of the most relevant markers allowing for the identification of QTL regions. We found the same relevant markers simultaneously affecting different growth curve parameters (mature weight and maturity rate): two (ALGA0096701 and ALGA0029483) for RQR(0.2), one (ALGA0096701) for RQR(0.5), and one (ALGA0003761) for RQR(0.8). Three average genomic growth curves were obtained and the behavior was explained by the curve in quantile 0.2, which differed from the others. RQR allowed for the construction of genomic growth curves, which is the key to identifying and selecting the most desirable animals for breeding purposes. Furthermore, the proposed model enabled us to find, at different levels of interest (quantiles), the most relevant markers for each trait (growth curve parameter estimates) and their respective chromosomal positions (identification of new QTL regions for growth curves in pigs). These markers can be exploited under the context of marker assisted selection while aiming to change the shape of pig growth curves.

Keywords: Genome association, growth curve, pig, QTL, regularized quantile regression.

Background

In general, the study of growth curves is carried out by fitting nonlinear models to weight (dependent variable) and age (independent variable) data. These models are used because they are flexible and have parameters with biological interpretations, such as maturity rate and adult weight.

With the goal of estimating SNP marker effects on parameter estimates of growth curves, Pong-Wong and Hadjipavlou [1] proposed a two-step approach. In the first step, nonlinear models were fitted to the weight-age data of each animal. In the second step, genomic regression models were fitted while considering the parameter estimates from the previous step as the dependent variable. Such an approach allows for the estimation of marker effects based only on the conditional mean of the dependent variable. Specifically, genomic growth curves are defined only in terms of population mean, i.e., the identification of genetically superior individuals in relation to the growth efficiency is based on population mean distribution (quantile 0.5 of a normal distribution of the sampled data).

An alternative approach for the second step that has not yet been exploited in genomic analyses of growth curves is the Quantile Regression (QR) [2]. This methodology allows for the estimation of marker effects at different levels (quantiles) of the variable of interest. Obtaining these effects in specific quantiles allows for a more informative study on the chromosomal regions affecting the growth curve trajectory.

In general, the larger number of markers and the dependence between them due to linkage disequilibrium leads to multicolinearity estimation problems. Thus, methods such as shrinkage estimation, which highlight the high dimensionality and multicollinearity issues, are required. Under a QR framework, this method is named regularized quantile regression (RQR), since the shrinkage (or penalty) parameter regularizes the variance of the markers' effects, thus performing a direct variable selection framework.

We aimed to propose and evaluate a regularized quantile regression for SNP marker effect estimation of pig growth curves, as well as to identify the chromosome regions of the most relevant markers and to estimate the genetic individual weight trajectory over time (genomic growth curve) under different quantiles (levels).

Methods

Animals and genotyping data

Phenotypic data was obtained from the Pig Breeding Farm of the Department of Animal Science of the Federal University of Viçosa, Minas Gerais, and refer to the weights at birth, 21, 42, 63, 77, 105 and 150 days of age. These weights were measured in 345 animals from a F2 outbred population (Brazilian Piau X commercial). More details about this population are found by Azevedo et al. [3] and Band et al. [4].

DNA was extracted at the Animal Biotechnology Lab from Animal Science Department of Federal University of Viçosa. The low-density customized SNPChip with 384 markers was based on the Illumina Porcine SNP60 BeadChip (San Diego, CA, USA, [5]). The number of SNP markers was distributed as follows in the pig chromosomes: (*Sus scrofa*; SSC): SSC1 (n = 56), SSC4 (n = 54), SSC7 (n = 59), SSC8 (n = 31), SSC17 (n = 25), and SSCX (n = 12), totaling 237 SNPs. These markers were selected according to QTL positions that were previously identified in this population by using meta-analyses [6] and fine mapping [7, 8]. Thus, although a small number of markers have been used, the customized SNPchip based on previously identified QTL positions ensures appropriate coverage of the relevant genome regions in this population.

Statistical analysis

Initially, the logistic nonlinear regression model [9] was fitted to the individual weight-age data:

$$w_{ij} = \frac{\alpha_{1i}}{1+\exp[(\alpha_{2i} - t_j)/\alpha_{3i}]} + e_{ij}, \qquad (1)$$

where w_{ij} is the weight of the animal i at age $t_j (0, 21, 42, 63, 77, 105 \text{and} 150)$; α_{1i}, α_{2i} and α_{3i} are the parameters. If $\alpha_{3i} > 0$ then α_{1i} is the horizontal asymptote as $t_j \to \infty$ (mature weight) and 0 is the horizontal asymptote as $t_j \to -\infty$. If $\alpha_{3i} < 0$ these roles are reversed. The parameter α_{2i} is the t_j value at which the response is $\alpha_{1i}/2$. It is the inflection point of the curve. The scale parameter α_{3i} (growth scale) represents the distance on the t-axis between this inflection point and the point where the response is $\alpha_{1i}/(1+e^{-1}) \approx 0.73\alpha_{1i}$; e_{ij} is the independent and normally distributed residual term,

$e_{ij} \sim N(0, \sigma_e^2)$. In this parameterization, the growth scale parameter is the reciprocal of growth rate on the model presented by Ratkowsky [10].

After obtaining parameter estimates of the logistic model, they were used as dependent variables in a linear model to carry out fixed effect corrections (sex, lot, and halothane gene). The corrected variables were identified based on the residual of the fitted linear model plus the overall mean. Subsequently, the corrected variables ($\hat{\alpha}_{1i}^*, \hat{\alpha}_{2i}^*$ and $\hat{\alpha}_{3i}^*$) were used as dependent variables in a multiple regression model while using SNP markers as the independent variables. This procedure is known in the literature as a two-step approach: in the first step, a growth curve is fitted to the data of each animal, and in the second step, the parameter estimates from the previous step are used as phenotypic values [1, 11].

In the second step, the following genomic model proposed by Meuwissen et al. [12] was fitted separately for each trait (parameter estimates from previous step):

$$y_i = \left[\mu + \sum_{k=1}^{237} x_{ik} \beta_k \right] + \varepsilon_i, \qquad (2)$$

in which y_i is the corrected phenotype $\hat{\alpha}_{1i}^*, \hat{\alpha}_{2i}^*$ and $\hat{\alpha}_{3i}^*$ from the first step; μ is the general mean; x_{ik} is the SNP marker, encoded as 2 (AA), 1 (Aa), or 0 (aa); β_k is the effect of the marker k; and ε_i corresponds to the residual term, $\varepsilon_i \sim N(0, \sigma_e^2)$.

To obtain the markers' effects at different levels of the variables (traits defined by $\hat{\alpha}_1^*, \hat{\alpha}_2^*$ and $\hat{\alpha}_3^*$), the regularized quantile regression [13] was used. This method consists of obtaining the marker effects (β_k) that solve the following optimization problem:

$$\hat{\beta}_s = \text{argmin}_{\beta} \left\{ \sum_{i=1}^{345} \rho_{\tau s} \left[\hat{\alpha}_{si}^* - \left(\mu + \sum_{k=1}^{237} x_{ik} \beta_{sk} \right) \right] + \lambda_s \sum_{k=1}^{237} |\beta_{sk}| \right\},$$

where s = 1, 2, and 3 (respectively for each assumed trait, $\hat{\alpha}_1^*, \hat{\alpha}_2^*$ and $\hat{\alpha}_3^*$); $\sum_{k=1}^{237} |\beta_{sk}|$ is the sum of the absolute values of the regression coefficients; λ_s is the regularization parameter for each trait; and $\tau \in (0,1)$ indicates the quantile of interest. This parameter (λ_s) is required to avoid multicollinearity problems that are a result of the larger number of highly dependent markers associated with linkage disequilibrium. It leads to the formulation of the RQR.

The parameter $\rho_{\tau s}(.)$ is denoted as a check function [2] and is defined by:

$$\rho_{\tau s}\left[\hat{\alpha}_{si}^{*}-\left(\mu+\sum_{k=1}^{237}x_{ik}\beta_{sk}\right)\right]=\begin{cases}\tau\cdot\left[\hat{\alpha}_{si}^{*}-\left(\mu+\sum_{k=1}^{237}x_{ik}\beta_{sk}\right)\right], & \text{if } \hat{\alpha}_{si}^{*}-\mu+\sum_{k=1}^{237}x_{ik}\beta_{sk}>0,\\ -(1-\tau)\cdot\left[\hat{\alpha}_{si}^{*}-\left(\mu+\sum_{k=1}^{237}x_{ik}\beta_{sk}\right)\right], & \text{otherwise.}\end{cases}$$

in which $\tau \in (0,1)$ indicates the quantile of interest. Thus, the values of $\beta_{sk}(\tau)$ represent the markers' effects in the τ^{th} quantile of interest for s^{th} trait.

In this study, for each trait ($\hat{\alpha}_1^*, \hat{\alpha}_2^*$ and $\hat{\alpha}_3^*$), the quantiles $\tau = 0.2, 0.5$ and 0.8 were used to generate results at three distinct levels that may characterize the low, average, and high distribution of the phenotypic values under study ($\hat{\alpha}_{1i}^*, \hat{\alpha}_{2i}^*$ and $\hat{\alpha}_{3i}^*$). Furthermore, these quantiles were chosen to minimize the residual term in previous studies (pilot analysis) by using the same datasets.

In order to verify whether marker effects differ between the quantile levels of the traits ($\hat{\alpha}_1^*, \hat{\alpha}_2^*$ and $\hat{\alpha}_3^*$), the 2.5% most relevant SNPs (highest absolute values) and their p values, based on bootstrapped standard error values, were presented. In addition, these SNPs were used to identify possible QTL regions affecting growth traits in pigs.

The Genomic Estimated Breeding Values (GEBV) from RQR were obtained through $GEBV(\tau) = \hat{u} = \sum_k x_{ik}\hat{\beta}_k(\tau)$, in which τ represents the quantile of interest. Subsequently, the genomic growth curves were obtained for each animal based on GEBV ($\hat{\mathbf{u}}$) according to the following expression:

$$\hat{y}_{ij} = \frac{\hat{\mu}_{\hat{\alpha}_1^*} + \hat{u}_{\hat{\alpha}_{1i}^*}}{\left\{1+\exp\left[\left(\hat{\mu}_{\hat{\alpha}_2^*} + \hat{u}_{\hat{\alpha}_{2i}^*}\right) - \left(\hat{\mu}_{\hat{\alpha}_3^*} + \hat{u}_{\hat{\alpha}_{3i}^*}\right)t_{ij}\right]\right\}}, \qquad (3)$$

in which $\hat{y}_{ij}$ is the predicted breeding value for each animal i for the weight at each age (t_{ij}) (j = 0 to 150 d); $\hat{\mu}_{\hat{\alpha}_1^*}, \hat{\mu}_{\hat{\alpha}_2^*}$ and $\hat{\mu}_{\hat{\alpha}_3^*}$ are the means of each trait (parameter estimates for the logistic model); and $\hat{u}_{\hat{\alpha}_{1i}^*}, \hat{u}_{\hat{\alpha}_{2i}^*}$ and $\hat{u}_{\hat{\alpha}_{3i}^*}$ are the GEBV of these traits.

Finally, the genetic parameters for the interpretable traits derived from the logistic model (α_1 and α_3) as well as the original traits associated with slaughter weight (SW) and average daily gain (ADG) were estimated by using the following multi-trait model:

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X_1} & \mathbf{0}\\ \mathbf{0} & \mathbf{X_2}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix} + \begin{bmatrix}\mathbf{Z_1} & \mathbf{0}\\ \mathbf{0} & \mathbf{Z_2}\end{bmatrix}\begin{bmatrix}\mathbf{g_1}\\ \mathbf{g_2}\end{bmatrix} + \begin{bmatrix}\mathbf{e_1}\\ \mathbf{e_2}\end{bmatrix}, \qquad (4)$$

where $\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}$ is the vector of response variables of traits I and II (α_1 and α_3 with SW and ADG), $\mathbf{X_1}$ and $\mathbf{X}_2$ are the fixed-effects design matrix (Sex, Batch, and Halothane presence), $\mathbf{Z_1}$ and $\mathbf{Z}_2$ are the random-effects design matrix, and $\begin{bmatrix} \mathbf{e_1} \\ \mathbf{e}_2 \end{bmatrix}$ is the vector of random residuals of the two traits. It is assumed that $\begin{bmatrix} \mathbf{g_1} \\ \mathbf{g}_2 \end{bmatrix} \sim N(\mathbf{0}, \mathbf{G} \otimes \mathbf{H})$, where $\mathbf{H} = \begin{bmatrix} \sigma^2_{g_1} & \sigma_{g_{12}} \\ \sigma_{g_{21}} & \sigma^2_{g_2} \end{bmatrix}$ is the additive genetic variance and covariance matrix of the two traits, and $\begin{bmatrix} \mathbf{e_1} \\ \mathbf{e}_2 \end{bmatrix} \sim N(\mathbf{0}, \mathbf{I} \otimes \mathbf{R})$, where $\mathbf{R} = \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_{12}} \\ \sigma_{e_{21}} & \sigma^2_{e_2} \end{bmatrix}$ is the residual variance and covariance matrix of the two traits. Finally, **G** is the additive relationship matrix constructed by using 501 pigs and **I** is the identity matrix.

Computational features

Fitting of the models was carried out by using the *nls* (to fit the logistic nonlinear model in the first step) and *rq* (to fit the regularized quartile regression in the second step) functions of the *stats* and *quantreg* packages [14] of R software [15], respectively. The Mixed Model Analyses were performed in ASReml 3.0 [16].

To obtain the shrinkage parameter values (λ), a grid of λ values between 0 and 50 was utilized, varying in 0.5 increments. The predictive capacity, defined as the correlation between the estimated and observed values (curve parameters that were obtained from fitting the Logistic model to the weight-age data), was used as a criterion to define the optimal value λ .

The computational codes that were implemented in the R software are found on the website of the Statistics Department of the Federal University of Viçosa (2017): http://www.det.ufv.br/~moyses/links.php.

Results

The summary containing the descriptive statistics of the adjusted phenotypic data is presented in Table 1.

Table 1. Means, standard deviations and ranges for weights at seven different ages of F2 outbred population.

Age, d	N	Mean weight ± SD, kg	Min, Kg	Max, kg
0	345	1.20 ± 0.27	0.53	2.13
21	345	4.90 ± 1.00	2.56	8.00
42	345	8.36 ± 1.81	2.66	12.90
63	345	16.29 ± 3.38	7.43	26.53
77	345	21.44 ± 4.39	9.30	34.50

The summary containing the correlation and descriptive statistics of the adjusted phenotypic data ($\hat{\alpha}_{1i}^{*}$, $\hat{\alpha}_{2i}^{*}$ and $\hat{\alpha}_{3i}^{*}$) is presented in Table 2.

Table 2. Correlation and descriptive statistics among the adjusted phenotypic data ($\hat{\alpha}_{1i}^{*}$, $\hat{\alpha}_{2i}^{*}$ and $\hat{\alpha}_{3i}^{*}$).

	Correlation			Descriptive statistics		
	$\hat{\alpha}_{1i}^{*}$	$\hat{\alpha}_{2i}^{*}$	$\hat{\alpha}_{3i}^{*}$	Mean ± SD	Min	Max
$\hat{\alpha}_{1i}^{*}$	1.00	0.82	0.63	89.43 ± 22.32	35.70	149.85
$\hat{\alpha}_{2i}^{*}$	0.82	1.00	0.83	113.18 ± 17.97	72.83	166.43
$\hat{\alpha}_{3i}^{*}$	0.63	0.83	1.00	32.03 ± 4.24	22.76	47.29

Considering the aforementioned grid (0 to 50, by 0.5), the shrinkage parameter value that showed the best results in terms of predictive capacity was λ = 0.5. Specifically, the predictive capacity ranged between 0.6219 and 0.8252 (Table 3).

Table 3. Predictive capacity obtained by means of RQR, considering estimates of the

nonlinear regression parameters.

Quantile	Trait		
	$\alpha_1(\lambda=0.5)$	$\alpha_2(\lambda=0.5)$	$\alpha_3(\lambda=0.5)$
0.2	0.7143	0.6938	0.6219
0.5	0.8252	0.7889	0.7904
0.8	0.7678	0.7663	0.7636

The mean and standard error for marker effects ($\hat{\beta}_k$'s) and R^1 goodness of fit measure for each quantile adjusted model are present in Table 4. The goodness of fit ranged between 0.67 and 0.75 (Table 4).

Table 4. Mean, standard error for marker effects and Pseudo R^2 for each quantile adjusted model.

Model	Trait	Mean (Standard error)	Pseudo R^2*
RQR (0.2)	$\hat{\alpha}_1$	0.43(0.37)	0.71
	$\hat{\alpha}_2$	0.44(0.45)	0.69
	$\hat{\alpha}_3$	0.11(0.14)	0.70
RQR (0.5)	$\hat{\alpha}_1$	0.28(0.44)	0.68
	$\hat{\alpha}_2$	0.42(0.40)	0.67
	$\hat{\alpha}_3$	0.10(0.12)	0.68
RQR (0.8)	$\hat{\alpha}_1$	0.48(0.44)	0.75
	$\hat{\alpha}_2$	0.52(0.49)	0.74
	$\hat{\alpha}_3$	0.13(0.09)	0.75

*Pseudo R^2 [17].

In order to verify whether the most relevant SNPs for the three approaches (RQR (0.2), RQR (0.5), and RQR (0.8)) were the same, the 2.5% most relevant SNPs for each phenotype ($\hat{\alpha}_{1i}^{*}$, $\hat{\alpha}_{2i}^{*}$ and $\hat{\alpha}_{3i}^{*}$) were reported (Table 5).

Table 5 describes the most relevant markers considering the fitting through RQR (0.2). For the mature weight (α_1), the markers are located on chromosomes SSC1, SSC4, SSC7, SSC8, and SSC17 (Table 5). The position of the marker ALGA0096701 on chromosome 17 (55.81 cM) is in accordance with the results of Pierzchala et al. [18], in which the authors found QTL for the slaughter weight at the position 51.1 cM with the cross between Meishan, Pietrain, and European Wild Boar. For birth weight (α_2), the marker ALGA0044519 stands out, which is found in the SSC7 at the position 115.23 cM, next to the QTL for the birth weight found by Guo et al. [19] at the position 120.9 cM for crosses of Large white and Meishan. In terms of growth rate (α_3), the marker that presented with the highest effect is found on chromosome 8. The position of the marker ALGA0049546 at SSC8 (60.04 cM) is close to the position 62.2 cM, as reported by Casas-Carrillo et al. [20] for average daily gain when using families from outbred lines that were selected for high (fast) and low (slow) growth rates.

Table 5. Absolute values of the estimated effects of the 2.5% most relevant SNP by RQR

Phenotype	Quantile	SNP marker	Estimated effect (abs)	P Value*	Chromossome (SSC)	Position (cM)
	0.20	ALGA0096701	18.93	0.099	17	55.81
	0.20	ALGA0026109	15.29	0.019	4	75.57
	0.20	ALGA0024036	14.98	0.007	4	20.55
	0.20	ALGA0038840	14.50	0.041	7	15.18
	0.20	ALGA0029474	14.15	0.060	4	122.99
	0.20	ALGA0029483	14.07	0.042	4	123.28
	0.50	ALGA0047992	30.89	0.008	8	30.17
Mature	0.50	ALGA0047995	29.47	0.006	8	30.31
Weight,	0.50	ALGA0096701	21.81	0.058	17	55.81

α_1	0.50	ALGA0003761	17.22	0.098	1	50.37
	0.50	ALGA0044299	15.65	0.153	7	110.66
	0.50	ALGA0096707	15.57	0.144	17	55.84
	0.80	ALGA0007216	22.14	0.001	1	160.61
	0.80	ALGA0003761	19.86	0.018	1	50.37
	0.80	ALGA0096701	19.71	0.005	17	55.81
	0.80	ALGA0042986	15.88	0.014	7	90.01
	0.80	ALGA0029474	15.57	0.042	4	122.99
	0.80	ALGA0042863	15.57	0.009	7	86.24
	0.20	ALGA0048131	13.55	0.027	8	35.02
	0.20	ALGA0044519	13.12	0.020	7	115.23
	0.20	ALGA0096701	12.98	0.011	17	55.81
	0.20	ALGA0029483	12.50	0.029	4	123.28
	0.20	ALGA0026109	11.23	0.033	4	75.57
	0.20	ALGA0003761	10.85	0.095	1	50.37
	0.50	ALGA0026100	19.87	0.009	4	75.53
Birth	0.50	ALGA0047995	18.71	0.027	8	30.31
Weight,	0.50	ALGA0048131	18.47	0.029	8	35.02
α_2	0.50	ALGA0047992	16.36	0.062	8	30.17
	0.50	ALGA0039880	14.78	0.047	7	30.13
	0.50	ALGA0021973	14.36	0.015	4	0.28
	0.80	ALGA0048131	17.66	0.007	8	35.02
	0.80	ALGA0005071	17.64	0.002	1	80.44
	0.80	ALGA0042986	16.32	0.005	7	90.01
	0.80	ALGA0029483	15.21	0.010	4	123.28

	0.80	ALGA0003761	15.08	0.025	1	50.37
	0.80	ALGA0026769	14.49	0.073	4	90.18
	0.20	ALGA0049546	3.91	0.015	8	60.04
	0.20	ALGA0029483	3.77	0.005	4	123.28
	0.20	ALGA0096701	3.42	0.011	17	55.81
	0.20	ALGA0021973	3.31	0.014	4	0.28
	0.20	ALGA0048854	3.29	0.031	8	50.17
	0.20	ALGA0048131	3.27	0.035	8	35.02
	0.50	ALGA0048131	5.89	0.004	8	35.02
Growth	0.50	ALGA0021973	4.37	0.023	4	0.28
Rate,	0.50	ALGA0048854	4.13	0.058	8	50.17
α_3	0.50	ALGA0096701	3.66	0.075	17	55.81
	0.50	ALGA0027642	3.62	0.054	4	102.39
	0.50	ALGA0027644	3.36	0.087	4	102.41
	0.80	ALGA0003761	4.43	0.008	1	50.37
	0.80	ALGA0048131	3.74	0.018	8	35.02
	0.80	ALGA0024881	3.61	0.005	4	40.50
	0.80	ALGA0044299	3.34	0.052	7	110.66
	0.80	ALGA0026769	3.07	0.105	4	90.18
	0.80	ALGA0048133	3.01	0.034	8	35.04

*P value calculated using the bootstrap standard error.

Considering the RQR (0.5) in Table 5, the most important markers for α_1, α_2 and α_3 are located on chromosomes SSC4 and SSC8 (Table 5; RQR (0.5)). For α_1, , the marker ALGA0047992 stands out, which is found on SSC8 at the position 30.17 cM, which is close to the QTL for slaughter weight found by Beeckmann et al. [21], and at the position 33.9 cM on chromosome 8 in pigs obtained from crosses between Meishan,

Pietrain, and European Wild Boar. For the birth weight trait (α_2), the marker with the greatest estimated effect was ALGA0026100. The position of this marker at SSC4 (75.53 cM) is close to the position at 74.4 cM reported by Walling et al. [22] for body weight at birth. For α_3, the position of marker ALGA0048131 on SSC8 (35.02 cM) was close to the position 33.1 cM reported by Beeckmann et al. [21] who used data from an experimental cross between Meishan, Pietrain, and European Wild Boar for average daily gain (Table 5; RQR (0.5)).

Considering the RQR (0.8) in Table 5, the most significant SNPs for α_1, α_2 and α_3 are located on chromosomes SSC1 and SSC8 (Table 5; RQR (0.8)). Regarding the mature weight trait (α_1), the marker with the highest absolute value pertaining to the estimates of the parameter effect is ALGA0007216. This marker is located on chromosome 1 (160.61 cM). Chen et al. [23] used a pig population comprised of Yorkshires and Meishans to find significant QTLs for slaughter weight at the position 122.4 cM of SSC1, i.e., close to the position 160.61 cM of the ALGA0007216 marker (Table 5; RQR (0.8)).

Another interesting result that was observed through RQR is the simultaneous existence of important markers for different traits (Table 5). This fact is important for breeding, since pleiotropy is the main factor in genetic correlation. Specifically, for RQR (0.5) (Table 5), two markers (ALGA0047992 and ALGA0047995) were simultaneously important for the mature weight (α_1) and birth weight (α_2) traits. In addition, three SNPs for RQR (0.2) (ALGA0096701, ALGA0026109, and ALGA0029483) and one for RQR (0.8) (ALGA0042986) were simultaneously relevant for α_1 and α_2.

Considering the traits α_1 (mature weight) and α_3 (growth rate), two (ALGA0096701 and ALGA0029483), one (ALGA0096701), and one (ALGA0003761) markers were simultaneously important for the methodologies RQR (0.2), RQR (0.5), and RQR (0.8), respectively. For the traits α_2 (birth weight) and α_3 (growth rate), three markers in the RQR (0.2) methodology (ALGA0048131, ALGA0096701, and ALGA0029483), two in the RQR (0.5) methodology (ALGA0048131 and ALGA0021973), and three in the RQR (0.8) methodology (ALGA0003761, ALGA0026769, and ALGA0048131) were simultaneously relevant for these two traits (Table 5).

The three genomic growth curves ($\tau = 0.2, 0.5, 0.8$) that were obtained based on all of the data are shown in Fig. 1B. The estimated curve based on the three quantiles

showed a similar pattern until 100 d. After that, differences in the estimated growth curves increased with time (Fig. 1B). This result was expected given the increase in the heterogeneity of variances that were presented at the final evaluated times, 100 and 150 d (Fig. 1A).

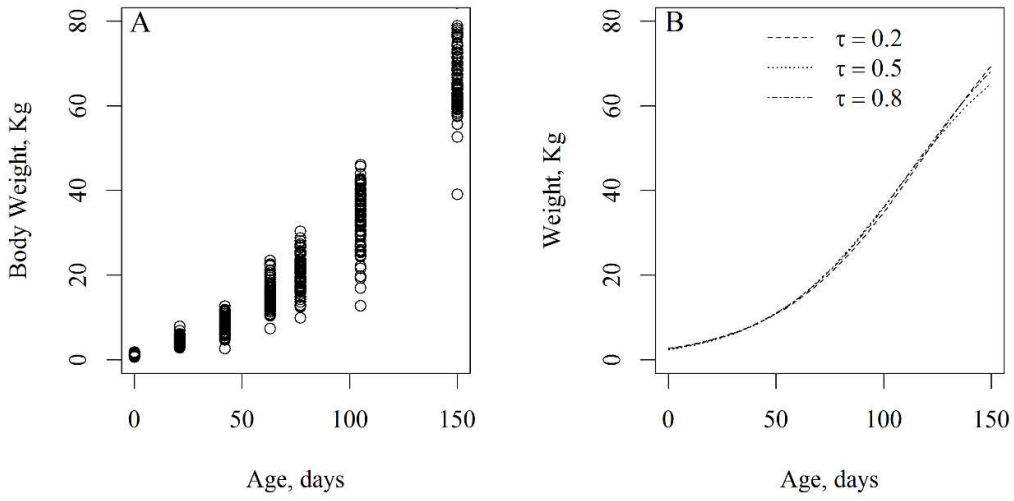


Figure 1. (A) Body weight (BW) data of animals over time. Each dot in the figures represents the BW of an animal, and (B) Genomic growth curves for each Regularized Quantile Regression (RQR), for quantiles 0.2, 0.5, and 0.8.

The genomic growth curves for each RQR, for quantiles 0.2, 0.5, and 0.8 and their confidence intervals showed significant differences (based on non-overlapping confidence intervals) only in terms of mature weight (Fig. 2A). These differences are highlighted in Fig. 2B.

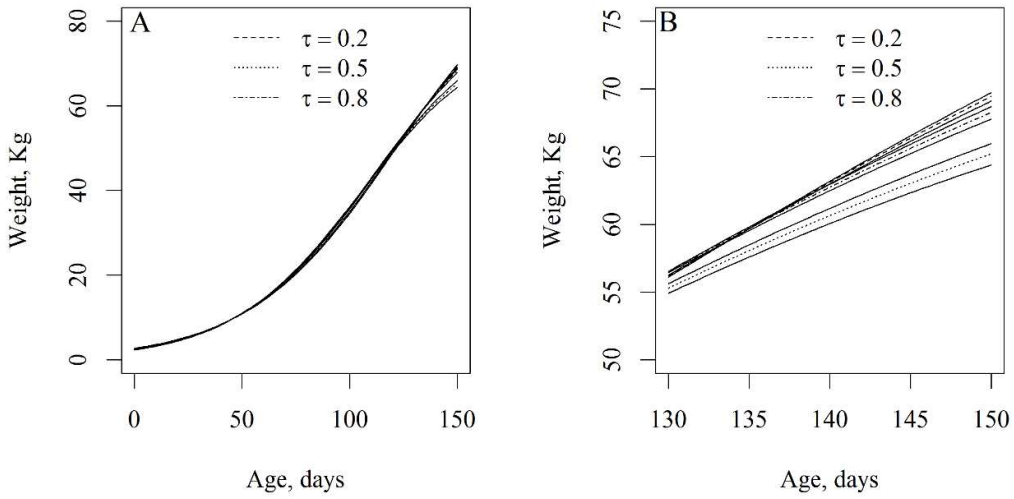


Figure 2. (A) Genomic growth curves for each Regularized Quantile Regression (RQR), for quantiles 0.2, 0.5, and 0.8 and their confidence intervals. (B) Genomic growth curves for each RQR, for quantiles 0.2, 0.5, and 0.8 and their confidence intervals. The genomic growth curves are highlighted within the distribution ranges of Age (130 to 150 d) and Weight (50 to 75 Kg).

Estimates of genetic parameters (heritability, and genetic and phenotypic correlations) are presented in Table 6. Estimates of heritability for growth curve parameters were moderate, with 0.447±0.200 and 0.4991±0.164, for parameters α_1 and α_3, respectively. The original traits (SW and ADG) had low heritability estimates, with 0.214±0.127 and 0.094±0.087, for SW and ADG, respectively.

Table 6. Genetic parameters[1] (standard error) for growth curve parameters, ADG, and SW.

Traits[2]	α_1	α_3	ADG	SW
α_1	0.447 (0.200)	0.809 (0.191)	-0.613 (0.390)	0.404 (0.113)
α_3	0.759 (0.030)	0.491 (0.164)	-0.681 (0.229)	-
ADG	0.047 (0.090)	-0.451 (0.06)	0.214 (0.127)	0.892 (0.677)

SW	0.662 (0.051)	-0.191 (0.080)	0.687 (0.039)	0.094 (0.087)

[1]Heritability, and genetic and phenotypic correlations presented on the diagonal, lower off-diagonal, and upper off-diagonal, respectively.

[2] α_1 = asymptotic weight (mature body weight); α_3 = inflection point; ADG = average daily gain; SW= slaughter weight.

Estimates of genetic and phenotypic correlations are presented in the off-diagonals (Table 6). Between the interpretable growth curve parameters (α_1 and α_3) with the original correspondent traits (SW and ADG), correlations were, respectively, highly positive and negative, with a positive genetic correlation estimated for parameters α_1 and SW (0.404±0.113) and a negative genetic correlation estimated for α_3 with ADG (-0.681±0.229). Phenotypic correlations between interpretable growth curve parameters with slaughter weight (SW) and average daily gain (ADG) traits were also moderately positive and negative, with 0.662±0.051 for α_1 with SW and -0.451±0.06 for α_3 with ADG.

Discussion

In this study, we aimed to propose and evaluate a regularized quantile regression (RQR) for SNP marker effect estimation on pig growth curves and to estimate the genetic weight trajectory over time (genomic growth curve) under different quantiles (levels). In order to do so, a real data set consisting of 345 animals from an F2 outbred population with information on 237 SNP markers, randomly distributed over six chromosomes, was used. The phenotypic data refers to the weight at birth, 21, 42, 63, 77, 105, and 150 days of age. To estimate SNP marker effects for growth curves, we used a two-step approach [1]. In the first step, we fitted logistic nonlinear models to the data of each animal, and in the second step, genomic regression models were fitted while considering the estimated parameters from the previous step as the phenotypic values. We obtained the three genomic growth curves for the three evaluated quantiles ($\tau = 0.2, 0.5, 0.8$). Finally, the genetic parameters for the interpretable traits of the logistic model (α_1 and α_3) and the original traits, slaughter weight and average daily gain, were estimated.

Quantile regression (QR) can be used to provide a more complete statistical analysis of the stochastic relationships among random variables. In general, the chosen

quantiles depend entirely on the purpose of the study, i.e., we can study all distributions or only some parts by defining specific quantiles. In this study, with the aim of representing three distinct levels that characterize low, average, and high distributions of the phenotypic values (estimated parameters while considering a logistic nonlinear model), we choose, $\tau = 0.2, 0.5, 0.8$.

The use of RQR to estimate SNP marker effects and obtain the estimated genomic growth curve was efficient since it was possible to construct genomic growth curves and find the most relevant markers, which thus allows for the identification of QTL regions at different levels of interest. Besides that, R^1 goodness of fit measures ranging from 0.67 to 0.75 indicating that the model fits well for the observations.

Unlike traditional methods that are based on conditional expectations, E(Y|X), RQR allows us to fit regression models on different parts of the distribution of the variable response, therefore enabling a more complete understanding of the phenomenon under study [2, 24]. Besides, the heterogeneous variance over time (Fig. 1A) indicates that there is not a single rate of change that characterizes changes in the probability distribution, therefore indicating that RQR is a good tool to deal with those situations. Also, the predictive capacity that was obtained by means of RQR (Table 3) was better than that obtained by Silva et al. [25].

The advantages of RQR, such as studying different parts of the distribution of the variable response, can be combined with those from the two-step approach. Specifically, the two-step approach enables us to obtain the genomic values for each observed time (t_j), as well as to estimate the weight for any other time of interest within the measured range before this weight is attained [25].

Based on the results, it is possible to note that RQR allows for the identification of markers close to QTLs at different distribution levels of the phenotypic values of interest. The regions indicated by RQR coincide with the results of several studies in which the authors found QTL for the traits that were evaluated in this study.

The use of quantile regression to estimate genomic curves based on three contrasting quantiles in our population was efficient when it came to producing distinct growth curves. Specifically, we can see in Fig. 2B that the final BW of the genomic growth curves was statistically different; in other words, the growth behavior over time changed in terms of mature weight. In fact, this result shows that RQR is a statistical

method that could be effectively used to estimate more than a single mean behavior, thereby providing a more complete picture of the relationships between variables.

The genetic correlations between α_1 and α_3 with BW and ADG had, respectively, a high positive and negative genetic correlation, which indicates that α_1 and α_3 have the potential to be used as selection tools to improve SW and ADG. Additionally, the high genetic correlation between α_1 with α_3 and SW with ADG enable us to understand causes of SNPs' pleiotropic effects. These results are in agreement with Silva et al. [25], who found significant genetic correlation between the interpretable traits of logistic model ($r_{\alpha_1,\alpha_3} = -0.69$) in the same populations that were used in this study. The difference between the signals of genetic correlation estimates observed in the present study is due to the different Logistic model parameterizations. Specifically, our approach uses the parameterization presented in Pinheiro and Bates [9], where the growth scale parameter (α_3) is the reciprocal of growth rate [10, 25].

The study of different distribution levels of the variable of interest using QR has been successfully performed in medicine by Beyerlein et al. [26], who used QR in GWAS (Genome-Wide Association Study) analysis in human genetics where they emphasized statistical and biological advantages when estimating marker effects in different quantiles of the phenotypic distribution. Sun et al. [27] proposed to use QR to identify hypermethylated CpG islands (CGIs) that can be associated with breast and ovarian cancer. They concluded that the quantile level between 80% and 90% is the best strategy to identify methylated and unmethylated CGIs. Moreover, regularized quantile regression has already been successfully evaluated for analyzing ultra-high dimension data [28]. These authors demonstrated that QR greatly enhances existing tools for large dimensional data analysis, since it revealed a substantial reduction in model complexity when compared with alternative methods.

However, even though the use of RQR is promising and efficient, more studies are needed to address the choice of the shrinkage parameter value, which is always critical to find as it can be defined by using a grid of values, cross-validation, or by using a Bayesian approach. Another issue about the use of RQR is the choice of the quantile. There are a lot of quantiles that can be used; therefore, finding the best one to explain the functional relationship is a challenge.

Conclusions

The proposed model enabled the discovery, at different levels of interest (quantiles), of the most relevant markers for each trait (growth curve parameter estimates) and their respective chromosomal positions (identification of new QTL regions for growth curves in pigs). Furthermore, RQR enabled the construction of genomic growth curves, which identified genetically superior individuals in relation to growth efficiency.

References

1. Pong-Wong R, Hadjipavlou GA. A two-step approach combining the Gompertz growth with genomic selection for longitudinal data. BMC Proceedings. 2010; 4:S4.

2. Koenker R, Basset G. Regression Quantiles. Econometrica. 1978; 46:33–50.

3. Azevedo CF, Nascimento M, Silva FF, Resende MDV, Lopes PS, Guimarães SEF. Comparison of dimensionality reduction methods to predict genomic breeding values for carcass traits in pigs. Genetics and Molecular Research. 2015; 14:12217-27.

4. Band GO, Guimarães SEF, Lopes PS, Peixoto JO, Faria DA, Pires AV, et al. Relationship between the Porcine Stress Syndrome gene and carcass and performance traits in F2 pigs resulting from divergent crosses. Genet Mol Biol. 2005; 28:92-96.

5. Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, et al. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS One. 2009 4: e6524.

6. Silva KM, Knol EF, Merks JWM, Guimarães SEF, Bastiaansen JWM, Van Arendonk JAM, et al. Meta-analysis of results from quantitative trait loci mapping studies on pig chromosome 4. Animal Genetics. 2011; 42:280-92.

7. Hidalgo AM, Lopes PS, Paixão DM, Silva FF, Bastiaansen JWM, Paiva SR, et al. Fine mapping and single nucleotide polymorphism effects estimation on pig chromosomes 1, 4, 7, 8, 17 and X. Genetics and Molecular Biology. 2013; 36:

511-19.

8. Verardo L, Silva FF, Varona L, Resende MDV, Bastiaansen JWM, Lopes PS, et al. Bayesian GWAS and network analysis revealed new candidate genes for number of teats in pigs. Journal of Applied Genetics. 2015; 56:123-32.

9. Pinheiro JC, Bates DM. Mixed-Effects Models in S and S-PLUS. Springer, New York; 2000.

10. Ratkowsky DA. Nonlinear Regression Modeling. Marcel Dekker, New York;1983.

11. Varona L, Moreno C, Garcia-Cortés LA, Yague G, Altarriba J. Two-step vs. joint analysis of Von Bertalanffy function. Journal of Animal Breeding and Genetics. 1999; 116:331-38.

12. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome wide dense marker maps. Genetics. 2001; 157:1819-29.

13. Li Y, Zhu J. L1-Norm Quantile Regression. Journal of Computational and Graphical Statistics. 2008; 17:1-23.

14. Koenker R. quantreg: Quantile Regression. R package version 5.29. 2016. https://cran.r-project.org/web/packages/quantreg/index.html. Accessed 19 Oct 2016.

15. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2014. https://www.r-project.org. Accessed 19 Oct 2016.

16. Gilmour AR, Gogel BJ, Cullis BR, Thompson R. ASReml User Guide Release 3.0 VSN International Ltd, Hemel Hempstead, HP1 1ES, UK. 2009.

17. Koenker R, Machado JAF. Goodness-of-Fit and Related Inference Processes for Quantile Regression. Journal of the American Statistical Association. 1999; 94: 1296-1310.

18. Pierzchala M, Cieslak D, Reiner G, Bartenschlager H, Moser G, Geldermann H. Linkage and QTL mapping for *Sus scrofa* chromosome 17. Journal of Animal Breeding and Genetics. 2003; 120:132- 37.

19. Guo YM, Lee GJ, Archibald AL, Haley CS. Quantitative trait loci for production traits in pigs: a combined analysis of two Meishan x Large White populations. Animal genetics. 2008; 39:486-95.

20. Casas-Carrillo E, Prill-Adams A, Price SG, Clutter AC, Kirkpatrick BW. Mapping genomic regions associated with growth rate in pigs. Journal of Animal Science. 1997; 75: 2047-53.

21. Beeckmann P, Mose G, Bartenschlager H, Reiner G, Geldermann H. Linkage and QTL mapping for *Sus scrofa* chromosome 8. Journal of Animal Breeding and Genetics. 2003;120:66–73.

22. Walling GA, Visscher PM, Andersson L, Rothschild MF, Wang L, Moser G, et al. Combined analyses of data from quantitative trait loci mapping studies. Chromosome 4 effects on porcine growth and fatness. Genetics. 2000; 155:1369-78.

23. Chen K, Hawken R, Flickinger GH, Rodriguez-Zas SL, Rund LA, Wheeler MB, et al. Association of the Porcine Transforming Growth Factor Beta Type I Receptor (TGFBR1) Gene with Growth and Carcass Traits. Animal biotechnology. 2012; 23: 43-63.

24. Cade BS, Noon BR. A gentle introduction to quantile regression for ecologists. Frontiers in Ecology and the Environment. 2003; 1: 412-420.

25. Silva FF, Resende MDV, Rocha GS, Duarte DAS, Lopes PS, Brustolini OJB, et al. Genomic growth curves of an outbred pig population. Genetics and Molecular Biology. 2013; 36:520-27.

26. eyerlein A, Von Kries R, Ness AR, Ong KK. Genetic Markers of Obesity Risk: Stronger Associations with Body Composition in Overweight Compared to Normal-Weight Children. PLoS ONE. 2011; 6: e19057.

27. Sun S, Chen Z, Yan PS, Huang Y-W, Huang THM, Lin S. Identifying hypermethylated cpg islands using a quantile regression model. BMC Bioinform. 2011; 12:54.

28. Wang L, Wu Y, Li R. Quantile regression for analyzing heterogeneity in ultra-high dimension. Journal of the American Statistical Association. 2012;107:214–222.

CAPÍTULO 2

Genomic prediction accuracies using regularized quantile regression methodology

Abstract: We aimed to evaluate the use of regularized quantile regression (RQR) for genomic prediction analyses of traits with or without skewness and with different proportions of epistatic variance. Data were simulated for 2,500 individuals. The genome included 5,000 markers, 50 or 5,000 quantitative trait loci (QTLs) for a low heritable trait (heritability = 0.1) and a high heritable trait (heritability = 0.5) simulated for the scenarios: combinations of trait distributions (symmetric normal or positive skewed) and percentage of epistatic and additive genetic variances (100% epistatic, 50% epistatic and 50% additive, 100% additive). Estimation of marker effects was performed using 2,005 individuals (training). The remaining 495 individuals were used for validation. The accuracies were calculated by the correlation between genomic estimated breeding values and true breeding values. Analyses were performed in R software using the RQR considering three quantiles (0.10, 0.50 and 0.90), Bayesian LASSO (BLASSO), Bayes B and GBLUP methods. In general, accuracies of genomic prediction for the symmetric normal trait were variable and decreased as the proportion of epistatic variance increased. For the positively skewed distribution and low heritable trait, the RQR was better than the other methods. In the symmetrical scenarios and low heritability (0.1), the RQR presented better results compared with the other methods when the trait presented a proportion of epistatic variance. In all scenarios with 100% epistatic variance, the RQR presented better results than BLASSO, Bayes B and GBLUP. In conclusion, for the trait with low heritability, RQR showed, in general, greater accuracies than the other approaches. These results suggest that improved accuracy of genomic prediction for low heritable traits can be obtained using RQR.

Background

Several statistical approaches have been proposed for the prediction of genomic breeding values (GEBVs), such as random regression (G-BLUP and RR-BLUP), Bayesian approaches and dimension reduction regression. Although these methods are widely used, they allow estimating marker effects based only on the conditional mean of the dependent variable. This limitation may be a problem when the assumptions of the model are violated, e.g., when there is heterogeneity of error variances, or when the

phenotypes present a skewed distribution. For these situations, these kind of methods may not fit the best model for the data.

The quantile regression (QR), proposed by Koenker and Basset (1978), allows the estimation of marker effects at different levels (quantiles) of the dependent variable distribution. However, due the large number of markers and their interdependence, problems as multicollinearity and high dimensionality may occur. Li and Zhu (2008) proposed the regularized quantile regression (RQR), which uses shrinkage estimation, such as LASSO, for solving the high dimensionality and multicollinearity issues.

The use of RQR to obtain the estimations of marker effects enables a more informative study about the phenotypes, making it possible to work in any quantile of interest. The RQR has been applied with success in breeding studies, such as in Barroso et al. (2017) and in Nascimento et al. (2017). Barroso et al. (2017) proposed a RQR for estimation of single nucleotide polymorphism (SNP) marker effects of pig growth curves. Nascimento et al. (2017) proposed the RQR for genomic selection (GS) studies to improve GEBVs prediction using a simulated dataset with different degrees of skewness. However, those studies did not take into account skewed phenotypes and whether the quantitative genetic variation presented epistasis.

In general, the epistatic effects are not considered important and the estimations are based only in the additive part of the variation because the additive variance is responsible for the correlations among relatives and for the response to selection (Huang and Mackay, 2016). However, studies point out which proportion of the genetic variance in the populations is due to interactions between genes (Evans et al., 2006; Carlborg and Haley, 2004) and how important is epistasis in the genetic architecture of quantitative traits (Mackay, 2014; Huang and Mackay, 2016). The epistasis in animals and plants is present in several classic genotype-phenotype patterns, such as coat color in several animals; comb type in chickens; kernel color in wheat; wing shape and flight velocity in *Drosophila melanogaster*; body weight and litter size in mice; and “rat-tail” phenotype in cattle.

It is known that teat number, dystocia rates and milk somatic cell counts (SCC) are examples of important skewed phenotypes in dairy cattle. When the phenotypes present skewness, the general solution is to use a logarithm scale to turn skewed distributions symmetric. However, the use of this transformation does not help data to become normal and, in some cases, makes data more variable and skewed (Feng et al., 2014).

In this sense, we aimed to use the RQR in a simulated dataset to predict genetic

values of individuals in scenarios with different proportions of epistasis in the quantitative genetic variance and considering phenotypes with symmetric and positive skewed distributions. We also aimed to compare the performance of RQR with other methods present in literature. In addition, it was sought to verify if the RQR captures epistatic effects even if they are not explicit in the model.

Materials and Methods

Simulated genome (Population structure)

The genotypic dataset used in this study was simulated using QMSim software (Sargolzaei and Schenkel, 2009). For the population structure one replicate, heritability of 0.2 and phenotypic variance equals to 1.0 were used. In the historical population, 100 generations with a size of 0 were simulated. Then, 3,000 generations with an increase in population size from 0 to 100 were simulated to create an initial linkage disequilibrium (LD).

In the next simulation step, five generations were used to expand the population. A total of 200 male founders and 1,389 female founders were randomly selected from the last generation with random mating design, which allowed obtaining 2,500 animals in the last generation.

The population was simulated with 5,000 SNP markers and 50 or 5,000 quantitative trait loci (QTLs), totaling two scenarios. The parameters used in the simulation are shown in Table 1. All these parameters were chosen to generate a population with similar characteristics of the dairy cattle breed (Brito et al., 2011).

Table 1. Parameters of the simulation process.

Population Structure	
Step 1: Historical generations (HG)	
Number of generations(size) - phase 1	100(0)
Number of generations(size) - phase 2	3000(100)
Step 2: Expanded generations (EG)	
Number of founder males from HG	200
Number of founder females from HG	1389

Number of generations	5
Litter size	1
The proportion of male progeny	0.5 /fix
Mating design	Random
Sire replacement	0.5 0.2
Dam replacement	0.2 0.2
Genome	
Number of chromosomes (Total)	29
Total length	2333
Number of markers (Total)	5000
Marker positions	Random
Number of marker alleles in the first historical generation	All 2
Marker allele frequencies in the first historical generation	Equal
Number of QTL loci on the chromosome (Total)	50 / 5000
QTL position	Random
Number of QTL alleles in the first historical generation	All 2
QTL allele frequency in the first historical generation	Equal
QTL allele effect	Gamma distribution (Shape = 0.40*)

*****The scale parameter were determined internally with QMSim, obeying the simulated genetic variance (Melo et al., 2016).

Simulated phenotypes

The phenotypes were simulated using the R software (R Development 2018). The QTLs file simulated in QMSim software was used to calculate the QTLs effects. Two types of heritabilities – low (0.1) and high (0.5) – were considered and five replicates were generated.

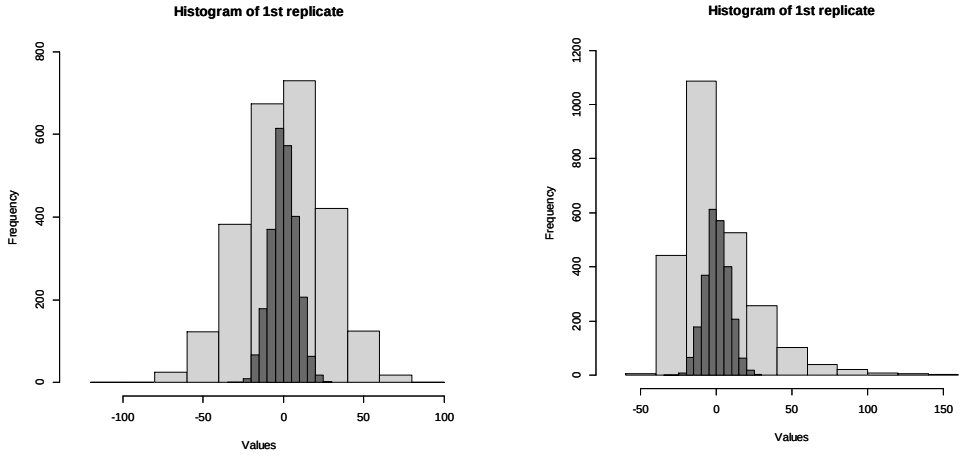
The quantitative genetic variance was split into three scenarios, 100% of the variance due to epistasis, 50% epistasis and 50% additive, and 100% additive. In the cases with both variances (additive and epistatic), the number of QTLs was divided into 40% for only additive effects, 40% for only epistatic effects and 20% having both effects.

Since this study considers a biallelic locus in a diploid genome, for insertion of epistatic effects, a two-allele system with additive-by-additive epistasis was used. Under

this system, the additive value for one locus changes depending on the other locus, and vice versa (Wolf et al., 2000; Mackay, 2014; Huang and Mackay, 2016).

The markers genetic effects were simulated using a gamma distribution with shape and scale parameters equal to 0.4 and 1.66, respectively. To control the percentage of quantitative genetic variance, when 50% was epistatic and 50% additive, the additive and epistatic effects were simulated and the variances of these effects were calculated. Then, a constant based on these two variances was multiplied by the epistatic effects. In this way, the variances are equal and each one assumes half of the heritability.

Since the objective was to evaluate skewed phenotypes as well as the proportion of epistasis in the genetic variance, two distributions were used for the error. For symmetrical phenotypes, errors were simulated using normal distribution with mean 0 and variance equals to the variance of the genetic effects, whereas for the skewed phenotypes, an exponential distribution with parameter equals to 1 divided by the square root of the variance of the genetic effects was used. Therefore, the phenotypes were obtained summing the genetic effect to the error. Figure 1 shows an example of phenotypes (gray color) and genomic values (black color) for a symmetrical and skewed distribution.



(a) (b)

Figure 1. Histograms of the first replicate of the scenario with 5,000 SNP markers and 50 QTLs, heritability equals to 0.1 and 50% additive variance and 50% epistatic variance. (a) error with symmetrical distribution. (b) error with exponential distribution.

Statistical Analysis

The dataset with 2,500 animals was split into two groups. The groups were formed based on the distance between the animals considering the genetic relationship matrix. The function *hclust* from R software was used to separate the animals. The training dataset consisted of the 2,005 most closely related animals, while the validation dataset contained 495 animals. Compared to approaches in which the separation of individuals is random, this way of splitting the data is more rigorous.

The training dataset was used with the proposal of estimating the SNP effects for the four methods analyzed in the study: BLASSO, Bayes B, GBLUP and RQR.

The RQR (Li and Zhu, 2008) was used to calculate the markers effects at different levels of the phenotypes. This method consists in obtaining the marker effect (β_k) that solve the following optimization problem:

$$\hat{\beta}_s(\tau) = \text{argmin}_{\beta}\left\{\sum_{i=1}^{2005}\rho_{\tau}\left(y_i - \left(\mu + \sum_{k=1}^{5000}x_{ik}\beta_k\right)\right) + \lambda\sum_{k=1}^{5000}|\beta_k|\right\},$$

in which $\tau = 0.1, 0.5$ and 0.9, $\sum_{k=1}^{5000}|\beta_k|$ is the sum of the absolute values of the regression coefficients, y_i is the phenotype, μ is the overall mean and λ is the regularization parameter.

The parameter $\rho_{\tau}(.)$ is denoted check function (Koenker and Bassett, 1978) and is defined by:

$$\rho_{\tau}\left(y_i - \left(\mu + \sum_{k=1}^{5000}x_{ik}\beta_k\right)\right) = \begin{cases} \tau\left(y_i - \left(\mu + \sum_{k=1}^{5000}x_{ik}\beta_k\right)\right), & \text{if } y_i - \left(\mu + \sum_{k=1}^{5000}x_{ik}\beta_k\right) > 0 \\ -(1-\tau)\left(y_i - \left(\mu + \sum_{k=1}^{5000}x_{ik}\beta_k\right)\right), & \text{otherwise.} \end{cases}$$

in which $\tau \in (0,1)$ indicates the quantile of interest. Thus, the values of β_k represent the effects of markers in the τ^{th} quantile of interest.

The quantiles $\tau = 0.1, 0.5$ and 0.9 were used to generate results at three distinct levels that may characterize the low, average, and high distribution of the phenotypic values under study. In RQR, to obtain the shrinkage parameter (λ) values, a grid of five λ values were used. The choice of these values was defined as varying between 1 and the mean of the BLASSO shrinkages.

With the estimated betas for each method, the GEBVs were obtained by $GEBV = \hat{u} = \sum_k x_{ik}\hat{\beta}_k$. The accuracies were calculated by the correlation between the true breeding value simulated and the GEBV.

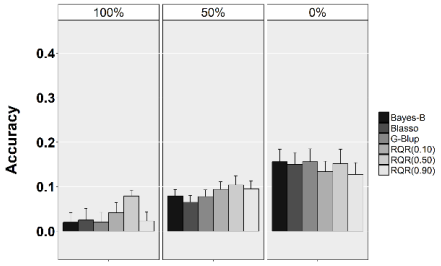
For the Bayesian methods, 50,000 iterations for the MCMC algorithms of the different models were used, with the first 5,000 iterations discarded as burn in. After every set of one iteration (thin), a sample was retained to calculate *a posteriori* statistics.

The *BGLR* function of BGLR package (Pérez and de los Campos, 2013) was used to estimate the effects for BLASSO and Bayes B methods. For GBLUP, the *kin.blup* function of rrBLUP package (Endelman, 2011) was applied. The RQR model fittings were performed using the *rq* function of the quantreg package (Koenker, 2015).

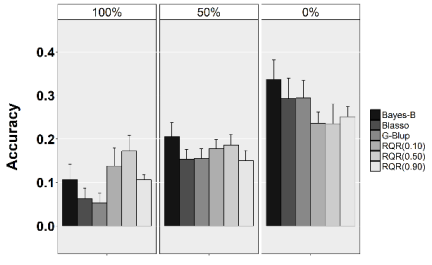
Results and discussion

For the scenario with 50 QTLs, symmetrical phenotypes and heritability equals to 0.1 (Figure 2 - a), it was observed that RQR performed better than the traditional methods when there was a proportion of epistatic variance in the genetic variance. In the case of 100% epistatic variance, the RQR(0.5) accuracy was greater than the other models. The RQR(0.5) accuracy was 0.079 (0.01), while for BLASSO and Bayes B the accuracies were 0.025 (0.03) and 0.020 (0.02), respectively. For 0% epistatic variance, the models showed similar performance. On the other hand, in the situation with 50% epistatic variance, the RQR(0.50), Bayes B and BLASSO presented accuracies of 0.094 (0.017), 0.079 (0.01) and 0.065 (0.01), respectively.

For heritability equals to 0.5, (Figure 2 - b), the RQR was superior to the other models only in the case with 100% epistasis. More specifically, RQR(0.5) accuracy was 60.7% greater than the Bayes B accuracy. In the 50% and the 0% situations, the Bayes B presented better results than RQR.



(a)

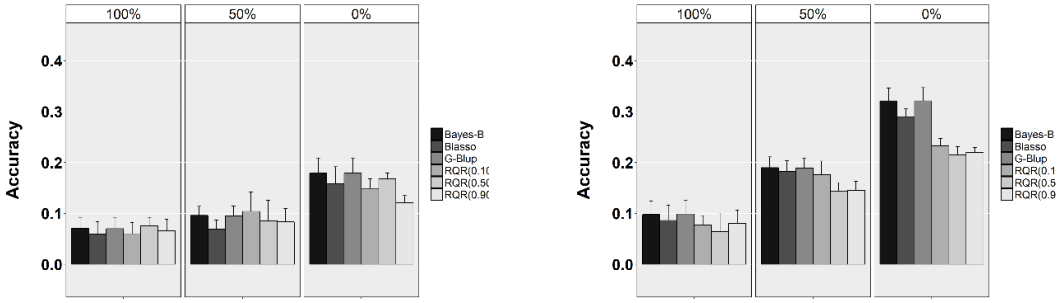


(b)

Figure 2. Accuracies and standard errors of the scenario with 5,000 SNPs, 50 QTLs and symmetrical phenotypes. The 100% corresponds to 0% additive variance and 100% epistatic variance; 50% represents 50% additive variance and 50% epistatic variance; and 0% symbolizes 100% additive variance and 0% epistatic variance. (a) heritability equals to 0.1, (b) heritability equals to 0.5.

For the scenario with 5,000 QTLs, symmetrical phenotypes, and heritability equals to 0.1 (Figure 3 – a), the results were similar for all variance proportions. Even when we had 100% epistatic variance, the accuracies ranged from 0.059 (0.024) to 0.076 (0.016). For 50% epistatic variance, the accuracies ranged from 0.069 (0.018) to 0.104 (0.038), and for 0% epistatic variance, they ranged from 0.121 (0.014) to 0.179 (0.03).

Figure 3 (b) shows that the traditional methods presented better results than RQR for all variance proportions. For 100% epistatic variance (or 0% additive variance), the GBLUP method showed the best accuracy, 0.099 (0.027). When we turn to 0% epistatic variance, the same result was achieved, but in this situation the difference between GBLUP and RQR(0.10) accuracies was higher, around 0.088.



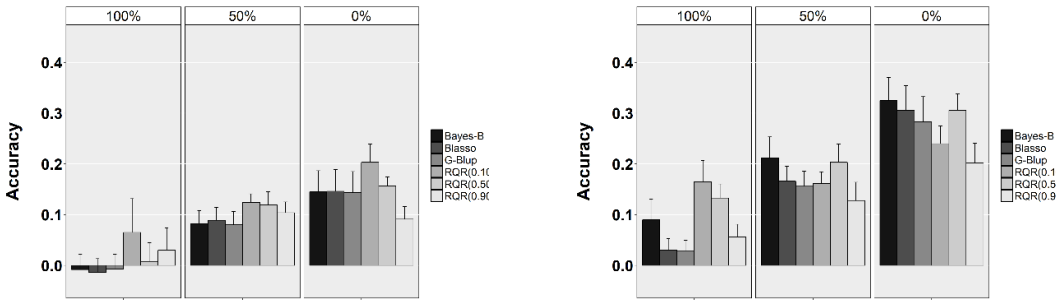
(a) (b)

Figure 3. Accuracies and standard errors of the scenario with 5,000 SNPs, 5,000 QTLs and symmetrical phenotypes. The 100% corresponds to 0% additive variance and 100% epistatic variance; 50% represents 50% additive variance and 50% epistatic variance; and 0% symbolizes 100% additive variance and 0% epistatic variance. (a) heritability equals to 0.1, (b) heritability equals to 0.5.

Figure 4 (a) presents the results for the scenario with 50 QTLs, skewed phenotypes, and low heritability (0.1). In this scenario, the RQR(0.1) was superior in all situations. We highlighted that in the case with 100% epistatic variance, BLASSO, Bayes B and

GBLUP methods presented negative accuracies and the RQR(0.1) accuracy was 0.065 (0.067). For 0% epistatic variance, the RQR(0.1) got almost 40% greater accuracy than BLASSO.

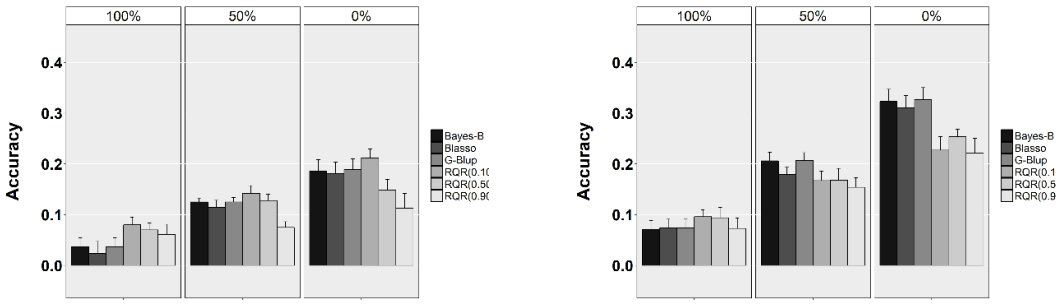
Figure 4 (b) shows the accuracies for heritability equals to 0.5. For this scenario, the RQR presented superior results only for the case with 100% epistatic variance. So, even when the phenotypes are skewed, for the cases with 50% and 100% additive variances, the Bayes B method showed better results than the RQR.



(a) (b)

Figure 4. Accuracies and standard errors of the scenario with 5,000 SNPs, 50 QTLs and skewed phenotypes. The 100% corresponds to 0% additive variance and 100% epistatic variance; 50% represents 50% additive variance and 50% epistatic variance; and 0% symbolizes 100% additive variance and 0% epistatic variance. (a) heritability equals to 0.1, (b) heritability equals to 0.5.

Figure 5 (a) shows the same results presented in Figure 4 (a), i.e., RQR(0.1) was the best method in terms of accuracy. For 100% epistatic variance, the accuracies ranged from 0.024 (0.023) for BLASSO method to 0.080 (0.015) for RQR(0.1). Figure 5 (b) shows that only for the situation with 100% epistatic variance, the RQR(0.1) presented superior accuracy than the traditional methods. For 0% epistasis, GBLUP method got 44% greater accuracy than RQR (0.1) method.



(a) (b)

Figure 5. Accuracies and standard errors of the scenario with 5,000 SNPs, 5,000 QTLs and skewed phenotypes. The 100% corresponds to 0% additive variance and 100% epistatic variance; 50% represents 50% additive variance and 50% epistatic variance; and 0% symbolizes 100% additive variance and 0% epistatic variance. (a) heritability equals to 0.1, (b) heritability equals to 0.5.

Considering the symmetrical scenarios (Figures 2 and 3), the RQR method performed better than the other methods when the simulated trait had low heritability (0.1) and presented epistatic variance. In general, for the scenarios with high heritability (0.5) and high number of QTLs, the traditional methods performed better than the RQR method. When the number of QTLs is low (50), the RQR method presented the best results only for 0% additive variance. In the skewed scenarios (Figures 4 and 5), RQR(0.1) was superior to the other methods when the simulated trait presented low heritability. Since the phenotypes were positively skewed, we expect that lower quantiles work better.

According to Briollais and Durrieu (2014), when the conditional distributions of the dependent variable are non-normal (for instance, positively skewed), the mean might not be the best measure. However, when we had high heritability (0.5), only for the case with 0% additive variance, the RQR presented better results. This situation can be explained by the shrinkage parameter. Since this parameter was defined as a grid of values based on the BLASSO method, these values cannot be the "best" ones for the quantile regression. This is a limitation of RQR analysis, i.e., the choice of the shrinkage value is always critical (Barroso et al., 2017), once this parameter exerts a greater shrinkage force in the RQR estimation process compared with BLASSO. Therefore, it is expected that a choice based on other methodologies can improve the RQR fit.

According to Roy (2000), one of the factors that may result in skewness of phenotypes is the epistasis. When there is complementary epistasis, the skewness is

positive and it is negative for duplicate epistasis (Snape and Riggs, 1975). The skewed distribution (positive in this case) of a trait in general suggests that the trait is under the control of non-additive gene action, especially epistasis (Babu et al., 2017). In the present study, when the phenotypes were skewed and the genetic variance has a proportion of epistasis, the RQR accuracy was better or close to the other methods. The RQR estimates can be used to construct predictions without assuming any parametric error distribution and when the variance is heterogeneous (Cade and Noon, 2003).

The RQR method efficiently predicted GEBVs considering a simulated trait with low heritability and the phenotype distributions presenting positive skewness. In these scenarios, RQR method showed better results than the traditional methods evaluated, even when the model was 100% additive. These results are reasonable because the RQR is an extension of a linear model for estimating functional relationships between variables in all portions of the distribution of a response variable (Nascimento et al., 2017). This also can be justified by Chicano et al. (2017), who reported that high epistasis results in a low coefficient of determination (R^2), consequently in a low heritability.

Many studies have confirmed the potential of QR. Briollais and Durrieu (2014) pointed out some aspects of the use of QR in genome-wide association studies (GWAS). Nascimento et al. (2017) used the RQR method as a novel computational tool for GS that can improve marker estimation and GEBVs prediction. Barroso et al. (2017) applied the RQR for SNP marker effect estimation of pig growth curves and identified important chromosome regions under different quantiles. Although considered an interesting method, the RQR has some limitations. As previously described, the shrinkage value and the quantile used are subjectively chosen. According to Nascimento et al. (2017), specific penalty values can be accessed exclusively for RQR by cross-validation (Silva et al., 2011) or via Bayesian inference (Alhamzawi et al., 2012). The use of these approaches may improve the performance of RQR. An infinite number of quantiles can exist in RQR, finding the “best” one to explain the functional relationship is still a challenge.

Although the QR approach seems to be interesting in situations considering high proportion of epistatic variance and traits with low heritability, further studies using simulated and real datasets with different genetic architectures, sizes (individuals and markers) and other proportions of epistasis (e.g., 15%, 25%, …) are needed to confirm the efficiency of QR compared to traditional GS approaches.

Conclusions

The use of RQR allows a more complete statistical analysis of the stochastic relationships among random variables. With RQR, it is possible to predict genetic values in different proportions of epistasis in the quantitative genetic variance, even if there are no interactions in the model. In general, RQR presented greater accuracies than the others methods evaluated in this study when the trait has low heritability. More specifically, when we have 100% of the genetic variance as epistatic variance, the RQR is, in most cases, better than the traditional methods. These results suggest that improved accuracy of genomic prediction for traits with low heritability and low additive variance can be obtained using RQR.

Reference

ALHAMZAWI, R.; YU, K.; BENOIT, D.F. Bayesian adaptive LASSO quantile regression. **Stat. Model**, v. 12, p. 279-297, 2012.

BABU, B.M.S.; JAGADEESH, B.N.; RAMESH, S.; KEERTHI, C.M.; SOWMYA, H.H. Third and Fourth Degree Statistics-Based Genetics of Quantitative Traits in Dolichos Bean (*Lablab purpureus L.*). **Int.J.Curr.Microbiol.App.Sci,** 6(10), p. 2551-2558, 2017.

BARROSO, L.M.A; NASCIMENTO, M.; NASCIMENTO, A.C.C.; SILVA, F.F.; SERAO, N.V.L.; CRUZ, C.D.; RESENDE, M.D.V.; SILVA, F.L.; AZEVEDO, C.F.; LOPES, P.S.; GUIMARÃES, S.E.F. Regularized quantile regression for SNP marker estimation of pig growth curves. **Journal of Animal Science and Biotechnology**, 8:59, 2017.

BRIOLLAIS, L.; DURRIEU, G. Application of quantile regression to recent genetic and -omic studies. **Hum. Genet**, v. 133, p. 951-966, 2014.

BRITO, F.V.; NETO, J.B.; SARGOLZAEI, M.; COBUCI, J.A.; SCHENKEL, F.S. Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. **BMC Genetics,** 12:80, 2011.

CADE, B.S.; NOON, B.R. A gentle introduction to quantile regression for ecologists. **Front Ecol Environ**, 1(8), p. 412–420, 2003.

CHICANO, F.; HU, B.; GARCÍA-SÁNCHEZ, P. **Evolutionary Computation in Combinatorial Optimization: 16th European Conference**, EvoCOP 2016, Porto, Portugal, March 30 -- Aril 1, 2016, Proceedings.

CARLBORG, O.; HALEY, C.S. Epistasis: too often neglected in complex trait studies? **Nat Rev Genet**, 5: 618–625, 2004.

EVANS, D.M.; MARCHINI, J.; MORRIS, A.P.; CARDON, L.R. Two-stage two-locus models in genome-wide association. **PLoS Genet**, 2: e157, 2006.

ENDELMAN, J. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. **Plant Genome**, v. 4, n. 3, p. 250-255, 2011.

FENG, C.; WANG, H.; LU, N.; CHEN, T.; HE, H.; LU, Y.; TU, X.M. Log-transformation and its implications for data analysis. **Shanghai Archives of Psychiatry**, v. 26, n.2, p. 105-109, 2014.

HUANG, W; MACKAY, T.F.C. The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis. **PLoS Genet,** 12(11), e1006421, 2016.

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

KOENKER R. quantreg: Quantile regression. R package version 4.91. http://CRAN.R-project.org/package=quantreg, 2015.

LI, Y.; ZHU, J. L1-Norm Quantile Regression. **Journal of Computational and Graphical Statistics**, v. 17, p.1-23, 2008.

NASCIMENTO, M.; SILVA, F. F.; RESENDE, M.D.V.; CRUZ, C. D.; NASCIMENTO, A. C. C.; VIANA, J.M.S.; AZEVEDO, C.F.; BARROSO, L.M.A. Regularized quantile regression applied to genome-enabled prediction of quantitative traits. **Genetics and Molecular Research**, v.16, 2017.

MACKAY, T.F.C. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. **Nat Rev Genet.**, v. 15, p. 22–33, 2014.

MELO, T.P.; TAKADA, L.; BALDI, F.; OLIVEIRA, H.N.; DIAS, M.M.; NEVES, H.H.; SCHENKEL, F.S.; ALBUQUERQUE, L.G.; CARVALHEIRO, R. Assessing the value of phenotypic information from non-genotyped animals for QTL mapping of complex traits in real and simulated populations. **BMC Genetics**,17: 89, 2016.

PÉREZ, P.; DE LOS CAMPOS, G. BGLR: a statistical package for whole genome regression and prediction. R package version 1.0.5, 2013.

R CORE TEAM. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018. https://www.r-project.org. Accessed 17 Jan 2017.

ROY, D. **Plant Breeding Analysis and Exploitation of Variation**, Alpha Science International Ltd, India, 701 p., 2000.

SARGOLZAEI, M.; SCHENKEL, F. QMSim: a large-scale genome simulator for livestock. **Bioinformatics**, v. 25, p. 680-681, 2009.

SILVA, F.F.; VARONA, L.; RESENDE, M.D.V.; BUENO FILHO, J.S.S.; ROSA, G.J.M.; VIANA, J.M.S. A note on accuracy of Bayesian LASSO regression in GWS. **Livest. Sci**., v. 142, p. 310-314, 2011.

SNAPE, J. W.; RIGGS, T.J. Genetical consequences of single seed descent in the breeding of self pollinating crops. **Heredity,** v.35, 211-219, 1975.

WOLF, J. B.; BRODIE, E. D. I.; WADE, M. J. **Epistasis and the Evolutionary Process**, Oxford, New York, 2000.

CONCLUSÕES GERAIS

Este trabalho abordou a utilização da Regressão Quantílica em seleção genômica ampla e comparou os resultados obtidos com aqueles provenientes de metodologias tradicionais, tais como o G-BLUP, BLASSO e Bayes B.

No Capítulo 1, o objetivo foi propor e avaliar a regressão quantílica regularizada (RQR) para obter a estimativa dos efeitos dos SNPs para curvas de crescimento de suínos, bem como identificar as regiões cromossômicas dos marcadores relevantes e estimar a trajetória de peso individual genético ao longo do tempo (curva de crescimento genômico) sob diferentes quantils (níveis). A RQR permitiu a descoberta, em diferentes níveis de interesse, dos marcadores mais relevantes, permitindo a identificação de QTLs.

Já no Capítulo 2, o objetivo foi avaliar o uso da RQR para análises de predição genômica em características com ou sem assimetria e com diferentes proporções de variância epistática. Os dados foram simulados para 2500 indivíduos. O genoma incluiu 5000 marcadores, 50 ou 5000 QTLs para uma característica de baixa herdabilidade ($h^2 = 0{,}1$) e característica de alta herdabilidade ($h^2 = 0{,}5$) simulada para os cenários: combinações de distribuição das características (simétrica normal ou assimétrica à direita) e porcentagem de epistasia na variância genética (100% epistática, 50% epistática e 50% aditiva, 100% aditiva). Em geral, a RQR apresentou maiores acurácias que as outras metodologias avaliadas (G-BLUP, BLASSO e Bayes B) quando a característica possui baixa herdabilidade. Especificamente, quando temos 100% da variância genética como variância epistática, a RQR é, na maioria dos casos, melhor do que os métodos tradicionais.

Os resultados obtidos indicam que a RQR é uma alternativa interessante em estudos de GWS, uma vez que possibilita a descoberta do modelo que melhor representa a relação entre as variáveis dependentes (fenótipos) e independentes (efeitos dos marcadores) aumentando o desempenho preditivo do modelo.