VINICIUS SILVA DOS SANTOS

SELEÇÃO GENÔMICA AMPLA EM SUÍNOS USANDO O MODELO DE SOBREVIVÊNCIA DE COX

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

VIÇOSA
MINAS GERAIS - BRASIL
2013

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
S237s
2013

Santos, Vinicius Silva dos, 1987-
Seleção genômica ampla em suínos usando o modelo de sobrevivência de Cox / Vinicius Silva dos Santos. – Viçosa, MG, 2013.
xi, 75f. : il. ; 29 cm.

Inclui anexo.
Inclui apêndice.
Orientador: Sebastião Martins Filho.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Análise de regressão. 2. Modelos matemáticos. 3. Polimorfismo (Genética). 4. Genômica. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22.ed. 519.536

VINICIUS SILVA DOS SANTOS

SELEÇÃO GENÔMICA AMPLA EM SUÍNOS USANDO O MODELO DE SOBREVIVÊNCIA DE COX

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

APROVADA: 26 de fevereiro de 2013.

Fabyano Fonseca e Silva
(Coorientador)

Marcos Deon Vilela de Resende
(Coorientador)

Carlos Souza do Nascimento

Sebastião Martins Filho
(Orientador)

Aos meus pais, Dilma (in memorian) e Nazareno,

À minha irmã Camila,

Aos meus amigos e parentes.

AGRADECIMENTOS

À Deus, pelo seu amor, graça e misericórdia, conduzindo meus passos e ajudando-me a realizar mais este sonho. Sem Ele, nada posso fazer.

Aos meus pais, Dilma (*in memorian*) e Nazareno, pelo amor, dedicação, zelo e ensino dados a mim e minha irmã. Muito obrigado por dedicar suas vidas à nós. À minha irmã Camila, pelo amor, amizade, carinho e ajuda em todos os momentos que preciso. Vocês foram meu incentivo a continuar em meio às dificuldades e à saudade.

Aos meus amigos, parentes e irmãos na fé pelo apoio e incentivo em todos os momentos.

À Universidade Federal de Viçosa (UFV) e ao Programa de Pós–Graduação em Estatística Aplicada e Biometria (PPESTBIO), por me permitir a realização deste curso de mestrado.

À FAPEMIG pela bolsa de estudos concedida.

Ao orientador, professor Sebastião Martins Filho, pela instrução, confiança, incentivo e paciência durante a realização deste trabalho.

Ao professor e coorientador Fabyano Fonseca e Silva, pelos ensinamentos, ideias, prontidão em ajudar-me sempre que precisei e incentivo na conclusão deste trabalho.

Ao Doutor e coorientador Marcos Deon Vilela de Resende, pelos ensinamentos, apoio, disposição em ajudar-me e dedicação ao ensino e à pesquisa.

Aos membros da banca examinadora, Doutor Carlos Souza do Nascimento, professor Fabyano Fonseca e Silva e Doutor Marcos Deon Vilela de Resende, pela disponibilidade e pelas sugestões para o aprimoramento deste trabalho.

Ao Departamento de Zootecnia da Universidade Federal de Viçosa, pelos dados concedidos para realização deste trabalho.

À mestre Camila Azevedo pelo auxílio na implementação computacional.

À todos os professores do Programa de Pós-graduação em Estatística Aplicada e Biometria (PPESTBIO) e de outros programas de pós-graduação que lecionaram durante estes dois anos do curso;

Aos coordenadores do PPESTBIO pelo trabalho realizado para o crescimento do programa.

Aos secretários do PPESTBIO, Joel (início do curso) e Carla (final do curso), pela prontidão em ajudar e a sanar questões pendentes relacionadas ao curso.

Aos chefes do departamento de estatística da UFV pelo trabalho realizado.

À secretária do departamento de estatística, Anita, pela ajuda na utilização do espaço físico destinado ao programa e pela simpatia.

A todos os alunos do mestrado, pela troca de saberes e pela companhia nos momentos de estudos e descontração, em especial aos ingressantes das turmas 2010/II (Camila, Flávia, Gislane, Jaciane, Lidiane, Leilimar, Renata e Dirceu) e 2011/I (Maria de Fátima, Priscila, Bruno, Cássio, Diego e Wagner).

À todos que de alguma forma contribuíram para a conclusão deste trabalho. Meu muito obrigado.

BIOGRAFIA

VINICIUS SILVA DOS SANTOS, filho de Dilma Silva dos Santos e de Antonio Nazareno dos Santos, nasceu em Belém, Pará, em 29 de dezembro de 1987.

Em maio de 2006, ingressou no curso de Bacharelado em Estatística na Universidade Federal do Pará, Belém-PA, graduando-se em outubro de 2010.

Em fevereiro de 2011, ingressou no curso de Mestrado do Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em 26 de fevereiro de 2013.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

SANTOS, Vinicius Silva dos, M.Sc., Universidade Federal de Viçosa, fevereiro de 2013. **Seleção genômica ampla em suínos usando o modelo de sobrevivência de Cox**. Orientador: Sebastião Martins Filho. Coorientadores: Fabyano Fonseca e Silva e Marcos Deon Vilela de Resende.

A seleção genômica ampla (GWS) surgiu em 2001 com o objetivo de aumentar a eficiência e acelerar o ganho de seleção no melhoramento genético baseando-se exclusivamente em marcadores após terem seus efeitos genéticos estimados a partir de dados fenotípicos. No contexto de análise de sobrevivência, o modelo de riscos proporcionais de Cox com efeito aleatório foi comparado ao modelo linear misto, ambos usando a matriz de parentesco baseada em marcadores em substituição à baseada em pedigree, método esse denominado GBLUP. A aplicação foi feita aos dados reais de uma população F_2 de suínos em que a variável resposta foi o tempo em dias, do nascimento até o abate do animal e as covariáveis: marcadores SNPs (238), sexo e lote de manejo. Os dados foram previamente corrigidos para seus efeitos fixos e a acurácia do método foi calculada com base na correlação dos postos dos valores genéticos genômicos preditos em ambos os modelos com os valores fenotípicos corrigidos. A análise foi repetida considerando menor número de marcadores SNPs que apresentassem maiores efeitos em módulo. Os resultados demonstraram concordância na predição dos valores genéticos genômicos e na estimação dos efeitos de marcadores para ambos os modelos na situação de dados não censurados e normalidade. No entanto, ao considerar a censura, o modelo de Cox com efeito aleatório normal foi o mais apropriado, uma vez que não houve concordância na predição dos valores genéticos genômicos e na estimação dos efeitos de marcadores com o modelo linear misto com dados imputados. A seleção de marcas permitiu um aumento nas correlações entre os postos dos valores genéticos genômicos preditos pelo modelo linear e pelo modelo de fragilidade de Cox com os valores fenotípicos corrigidos, sendo que para a característica analisada, 120 marcadores foram suficientes para maximizar a capacidade preditiva.

ABSTRACT

SANTOS, Vinicius Silva dos, M.Sc., Universidade Federal de Viçosa, february, 2013. **Genomic Wide Selection (GWS) in pigs using the survival model of Cox**. Adviser: Sebastião Martins Filho. Co-Advisers: Fabyano Fonseca e Silva and Marcos Deon Vilela de Resende.

The genomic wide selection (GWS) emerged in 2001 with the goal of increasing efficiency and accelerating the selection gain in genetic improvement based exclusively on markers after their genetic effects estimated from phenotypic data. In the context of survival analysis, Cox's proportional risk model with random effects was compared to the mixed linear model, both using parenthood matrices based on markers in substitution to basing on pedigree, this method being named GBLUP. The application was made on real data from an F2 population of pigs in which the dependent variable was the time in days, from birth to slaughter of the animal and the covariables: SNP markers (238), sex and handled lot. The data was previously corrected for fixed effects and the accuracy of the method was calculated based on the correlation of the ranks of genomic genetic values predicted in both models with the phenotypic values corrected. The analysis was repeated considering the least number of SNP markers that presented the greatest effect in module. The results showed agreement in the prediction of genomic genetic values and estimation of the effects of markers for both models in the situation of uncensored data and normality. However, when considering censored data, the Cox model with normal random effect was more appropriate, since there was no agreement in the prediction of genomic genetic values and estimation of the effects of markers with the mixed linear model with imputed data. The selection of markers allowed an increase in correlations between the positions of genomic genetic values predicted by the linear model and the Cox frailty model with phenotypic values corrected, being that for the characteristic being analyzed, 120 markers were sufficient to increase the predictive power.

1. INTRODUÇÃO

O Brasil ocupa a quarta posição no *ranking* dos maiores produtores e exportadores mundiais de carne suína, ficando atrás apenas da China, União Européia e Estados Unidos. Nos últimos dez anos, a participação do país subiu de 4% para 11% e deve atingir a 21% em 2019, sendo elevado ao posto de segundo maior exportador (Anuário Brasileiro de Aves e Suínos, 2011). Por meio do melhoramento genético é possível identificar indivíduos geneticamente superiores e criar novas combinações genotípicas por meio do cruzamento entre esses indivíduos com o intuito de aumentar a produção para atender ao crescente mercado consumidor.

No melhoramento genético animal e de plantas, a seleção tradicional dos melhores indivíduos é feita com base nos valores genéticos preditos que são obtidos pelo procedimento conhecido como REML/BLUP (estimação dos componentes de variância por máxima verossimilhança restrita - REML e predição dos efeitos aleatórios pelo BLUP - melhor preditor linear não viesado) com base nas informações fenotípicas e da genealogia. Entretanto, para muitas características, essa seleção demanda tempo e custo elevado (Resende et al., 2008; Goddard & Hayes, 2007).

Uma forma de acelerar esse processo de seleção promovendo ganhos genéticos com maior precisão e rapidez foi proposta por Lande & Thompson (1990) por meio da seleção assistida por marcadores (MAS), a qual utiliza conjuntamente dados fenotípicos e genotípicos (marcadores moleculares). No entanto, a MAS apresenta algumas limitações, tais como (Resende et al., 2008): as associações entre marcadores e locos controladores de características quantitativas (*Quantitative Trait Loci* - QTL) apresentam utilidade somente dentro de cada família mapeada e a seleção de marcadores ligados a QTLs está sujeita aos erros tipo II. Com isso, Meuwissen et al. (2001) propuseram a Seleção Genômica Ampla (GWS), baseada na análise de centenas ou milhares de marcadores amplamente distribuídos no genoma, a qual pode ser aplicada a todas as famílias em avaliação e não está sujeita aos erros tipo II oriundos da seleção marcador-QTL.

O uso de grande quantidade de marcadores se deu pelo desenvolvimento e baixo custo de novas classes desses, com destaque para os marcadores tipo SNPs (*Single Nucleotide Polymorphisms* - polimorfismos de base única), os quais se baseiam na detecção de polimorfismos (diferenças que ocorrem na sequência de DNA presente em mais de 1% dos indivíduos da população) resultantes da troca de uma única base no genoma (Resende et al., 2012; Simko et al., 2012; Resende et al., 2008).

Dado um grande número de marcadores moleculares gerados, a probabilidade de alguns desses estarem em LD (*linkage disequilibrium* - desequilíbrio de ligação) com o QTL é muito alta. Entende-se por LD o grau de associação não aleatória entre dois genes ou entre um QTL e marcador (Resende et al., 2012; Resende Jr. et al., 2010).

Pelo fato da GWS se basear na predição dos efeitos de QTL em LD com o marcador, surgem basicamente dois problemas na estimação desses efeitos. O primeiro está relacionado à multicolinearidade oriunda do desequilíbrio de ligação entre os marcadores. O segundo refere-se ao grande número de efeitos a serem estimados, uma vez que o número de marcadores (variáveis explicativas) pode ser igual ou maior que o número de observações (Resende et al., 2008).

Devido o número de marcadores ser maior que o número de indivíduos, métodos de estimação como quadrados mínimos não podem ser empregados, visto que falta graus de liberdade para estimar os efeitos de todos os marcadores. Como alternativa, tem sido proposto utilizar a seleção de variáveis ou procedimentos de estimação "*shrinkage*" (encurtamento dos coeficientes de regressão), ou ainda uma combinação de ambos (De Los Campos et al., 2012; Resende et al., 2008). Outra abordagem que pode ser empregada é a metodologia de modelos mistos, em que podem ser citados os métodos RR-BLUP (regressão aleatória do tipo BLUP) e GBLUP (BLUP tradicional baseado na matriz de parentesco dos marcadores).

Além dos métodos de regressão penalizada e da metodologia de modelos mistos, outros métodos tais como de estimação bayesiana (Bayes A, Bayes B, Bayes Cπ, Bayes Dπ, BLASSO) e de redução dimensional (Regressão via Quadrados Mínimos Parciais - PLSR e Regressão via Componentes Principais - PCR) tem sido utilizados no estudo da GWS. No melhoramento de suínos, podem ser citados os

trabalhos de Rocha (2011) em que os métodos BLASSO (LASSO Bayseiano) e RR-BLUP/GWS foram empregados e Azevedo (2012), onde foram comparados os métodos de redução dimensional (PLSR, PCR e Regressão via Componentes Independentes – ICR, sendo este último até então não utilizado no estudo da GWS) e RR-BLUP/GWS – Regressão Aleatória tipo BLUP.

Vale ressaltar que todos esses métodos citados tem sido amplamente empregados somente para o caso de modelos lineares. No entanto, De Los Campos et al. (2010) afirmam que a GWS pode ser aplicada também à dados discretos com base em modelos lineares generalizados ou à dados censurados, por meio do ajuste de modelos de sobrevivência, tais como o modelo de regressão de Cox.

Em programas de melhoramento genético animal há interesse em selecionar indivíduos que apresentem ganho de peso em menos tempo possível, sendo a variável resposta o tempo observado até a ocorrência do evento de interesse (falha) e denominado tempo de falha (Colosimo & Giolo, 2006). Como essa variável é discreta e alguns indivíduos não atingem o peso desejado no período avaliado, modelos usados em análise de sobrevivência para dados censurados têm sido empregados. Um dos modelos mais utilizados em análise de sobrevivência é o modelo de Cox (1972) ou modelo de riscos proporcionais de Cox. Com a introdução de um efeito aleatório, tem-se o modelo misto de Cox.

Ainda não foram encontrados na literatura trabalhos considerando modelos de sobrevivência, tais como o modelo de Cox no estudo da seleção genômica ampla para a idade ao abate em suínos. Assim, o objetivo geral deste trabalho foi empregar a metodologia GBLUP com base no modelo de Cox aos dados reais de uma população F_2 de suínos. Os objetivos específicos foram: comparar a metodologia usando o modelo linear misto e o modelo de sobrevivência de Cox; estimar valores genéticos genômicos e de marcadores SNPs por meio desses dois modelos; avaliar diferentes subconjuntos de marcadores e selecionar aquele com maior acurácia.

2. REVISÃO DE LITERATURA

2.1 Análise de Sobrevivência e Censura

Segundo Ducrocq (1997), a análise de sobrevivência consiste em examinar o intervalo de tempo para a ocorrência de um determinado evento de interesse, sendo esse tempo denominado **tempo de falha**. O termo *sobrevivência* refere-se basicamente a situações médicas em que o objetivo é estudar o tempo de cura ou recidiva da doença de um paciente. No entanto, a análise de sobrevivência também pode ser aplicada a estudos de seleção animal, em que, por exemplo, animais que apresentem ganho de peso em tempos não muito longos, dado as mesmas condições ambientais, serão selecionados (Colosimo & Giolo, 2006; Giolo et al., 2003).

Uma característica desses estudos longitudinais é que nem todos os indivíduos terão experimentado o evento até o final do estudo ou ainda poderá ocorrer a perda de acompanhamento desse indivíduo durante o estudo, devido a diversos fatores não relacionados ao experimento. Para esses indivíduos não se tem uma resposta completa, ou seja, não se sabe quando o evento ocorreu ou até mesmo se ele ocorreu, o que caracteriza observações incompletas ou parciais, denominadas censuras (Colosimo & Giolo, 2006).

Existem três tipos de censura: censura à direita, à esquerda e intervalar, e três mecanismos de censura: censura tipo I, tipo II e aleatória. Geralmente a censura é representada por uma variável indicadora de falha, onde 1 indica que a observação falhou, ou seja, ocorreu o evento de interesse e 0 que foi censurada (Gouvêa, 2010).

Censuras do tipo I ocorrem quando o estudo é finalizado após um período de tempo pré-especificado e alguns indivíduos ainda não apresentaram o evento de interesse. Censuras do tipo II são aquelas em que o estudo é finalizado após ocorrer o evento de interesse em um número pré-estabelecido de indivíduos. E a censura aleatória ocorre quando indivíduos são retirados do estudo sem ter ocorrido a falha ou também a ocorrência de um evento que não seja o de interesse (Colosimo & Giolo, 2006).

O fato de o tempo de falha ser maior que o tempo pré-estabelecido no estudo caracteriza censura à direita, pois o tempo de falha está à direita do tempo registrado.

Ocorre nos três mecanismos acima citados, sendo essa a mais comum em análise de sobrevivência. Já a censura à esquerda ocorre quando o tempo de falha é menor que o tempo registrado. Na censura intervalar não se conhece o tempo exato da ocorrência do evento de interesse, somente o intervalo em que ocorreu.

A censura pode ainda ser classificada em informativa e não informativa. Segundo Carvalho et al. (2011), a censura é informativa quando o indivíduo deixa de participar do estudo por motivo relacionado ao evento de interesse, e é não informativa quando ocorre ao acaso, ou seja, o motivo que o indivíduo deixou de participar do estudo independe do evento estudado. Os tipos e mecanismos de censura citados anteriormente são considerados censura não informativa. Esta suposição leva à simplificações na análise estatística (Ducrocq, 1997).

Em análise de sobrevivência, as observações são representadas pelo par (t_i, δ_i), em que t_i é o tempo de falha ou de censura e δ_i é a variável indicadora em que 1 representa o i-ésimo tempo de falha e 0 o i-ésimo tempo de censura. Neste trabalho será considerada a censura à direita, em que $t = \min(T, C)$ e $\delta = 1$ se $T \leq C$ e $\delta = 0$ se $T > C$, em que *T* é uma variável aleatória representando o tempo de falha e *C* uma outra variável aleatória independente de T, representando o tempo de censura, ou seja, para o individuo $i(i = 1, \ldots, n)$, de uma amostra de tamanho *n*, o par $(C_i, 0)$ representa uma observação censurada e $(T_i, 1)$ uma observação de falha. E para cada indivíduo *i*, considera-se o vetor $\boldsymbol{x}_i$ de covariáveis.

A seguir serão apresentadas as funções de sobrevivência *S(t)* e de taxa de falha ou função de risco *h(t)* comumente usadas para especificar o tempo de falha, representado pela variável aleatória contínua não-negativa T. Serão apresentadas também algumas relações entre essas funções, com base em Lee & Wang (2003).

2.2 Função de Sobrevivência

Seja *T* a variável aleatória contínua, que representa o tempo até a ocorrência de um evento de interesse, com função densidade de probabilidade $f(t)$, a função distribuição acumulada de *T* é dada por (Mood et al., 1974):

$$F(t) = P(T \leq t) = \int_0^t f(v)\,dv,$$

e representa a probabilidade do evento ocorrer até certo tempo *t*, enquanto que a função de sobrevivência, denotada por *S(t)*, é definida como a probabilidade de uma observação não falhar até certo tempo *t*, ou seja, a probabilidade de um indivíduo sobreviver por mais que um certo tempo *t*. Em termos probabilísticos, isto é escrito da forma:

$$S(t) = P(T > t) \tag{2.1}$$

A função de sobrevivência também pode ser escrita com base na função distribuição acumulada *F(t)*, em que

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t), \tag{2.2}$$

ou seja, a função de sobrevivência pode ser definida também como o complemento da função distribuição acumulada.

Segundo Lee & Wang (2003) e Pascoa (2008), a função de sobrevivência é monótona decrescente e contínua no tempo, além que, para $t = 0$, $S(t) = 1$, ou seja, a probabilidade de um indivíduo sobreviver ao tempo zero é 1 e $\lim_{t\to\infty} S(t) = 0$, ou seja, a probabilidade de um indivíduo sobreviver no tempo infinito é zero.

2.3 Função de Taxa Falha ou Função de Risco

A expressão (2.3) representa a taxa de falha que é obtida da probabilidade condicional da ocorrência de um evento no intervalo $[t, t+\Delta t)$ dado que o evento ainda não ocorreu no tempo *t*. Esta probabilidade condicional é dividida pelo intervalo de tempo Δt. A taxa de falha é o limite desta razão quando Δt tende a zero (Duchateau & Janssen, 2008):

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,|\, T \geq t)}{\Delta t} \tag{2.3}$$

Segundo Colosimo & Giolo (2006), a função distribuição da taxa de falha tem bastante utilidade na análise de sobrevivência, pois descreve a distribuição do tempo de vida de indivíduos, ou seja, como o risco instantâneo muda com o tempo.

Outra função útil em análise de sobrevivência é a função distribuição acumulada da taxa de falha, dada por:

$$H(t) = \int_0^t h(v)\,dv$$

2.4 Algumas Relações entre as Funções

As funções apresentadas anteriormente são matematicamente equivalentes. Dado o conhecimento de qualquer uma delas, as outras podem ser obtidas sem perda de generalidade. A seguir serão mostradas algumas dessas relações.

A função de taxa de falha *h(t)* pode também ser definida com base nas funções distribuição acumulada *F(t)* e função densidade de probabilidade *f(t)* pela seguinte expressão:

$$h(t) = \frac{f(t)}{1 - F(t)}$$

Da equação (2.2), tem-se que a função de sobrevivência é o complemento da função distribuição acumulada. Logo, a expressão acima pode ser reescrita da seguinte forma:

$$h(t) = \frac{f(t)}{S(t)} \tag{2.4}$$

Uma vez que a função densidade de probabilidade é obtida pela derivação da função distribuição acumulada, logo,

$$f(t) = \frac{d}{dt}F(t) = \frac{d}{dt}\left[1 - S(t)\right] = -S'(t) \tag{2.5}$$

Substituindo (2.5) em (2.4), temos que

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\left(\ln S(t)\right) \tag{2.6}$$

Integrando (2.6) de zero a *t* e usando a propriedade de que *S(0)*=1, temos que a função distribuição acumulada da taxa de falha é dada por:

$$H(t) = \int_0^t h(v)\,dv = \int_0^t -\frac{d}{dv}\left(\ln S(v)\right)dv = -\ln S(t)$$

Da expressão acima pode-se obter o seguinte:

$$S(t) = \exp\left\{-H(t)\right\} = \exp\left\{-\int_0^t h(v)\,dv\right\} \tag{2.7}$$

E ainda, das expressões (2.4) e (2.7) obtemos

$$f(t) = h(t)\exp\left\{-H(t)\right\}.$$

Para melhor entendimento, Lee & Wang (2003) apresentam o seguinte exemplo: suponha que o tempo de sobrevivência de uma população apresenta a seguinte função densidade de probabilidade

$$f(t) = e^{-t}, \quad t \geq 0.$$

Usando a definição de função distribuição acumulada,

$$F(t) = \int_0^t f(v)\,dv = \int_0^t e^{-v}dv = -e^{-v}\Big|_0^t = 1 - e^{-t}.$$

De (2.2) obtemos a seguinte função de sobrevivência

$$S(t) = e^{t},$$

e a função de taxa de falha é então obtida de (2.4):

$$h(t) = \frac{e^{-t}}{e^{-t}} = 1.$$

2.5 Modelo de Regressão de Cox

O modelo de regressão de Cox (1972) ou simplesmente modelo de Cox consiste em estimar os efeitos de covariáveis presentes no modelo em que a variável resposta é o tempo até a ocorrência de um evento de interesse. O modelo de Cox ajusta a função de taxa de falha considerando uma função base de risco $h_0(t)$ e

incluindo um vetor de covariáveis $\boldsymbol{x}$, com os componentes $\boldsymbol{x} = (x_1, x_2, ..., x_p)'$, de forma que:

$$h(t \mid \boldsymbol{x}) = h_0(t) g(\boldsymbol{x}'\boldsymbol{\beta}), \tag{2.8}$$

em que $g(\boldsymbol{x}'\boldsymbol{\beta})$ é uma função não-negativa que deve ser especificada tal que $g(\mathbf{0}) = 1$ e $\boldsymbol{\beta}$ é o vetor de parâmetros associado às covariáveis. Note que o modelo acima é composto pelo produto de dois componentes, um não-paramétrico e outro paramétrico. O termo não-paramétrico $h_0(t)$ é uma função não-negativa do tempo, denominado função base de taxa de falha, pois quando $\boldsymbol{x} = \boldsymbol{0}$, $h(t \mid \boldsymbol{x}) = h_0(t)$. A presença desse termo não-paramétrico torna o modelo de Cox bastante flexível, apresentando, por exemplo, alguns modelos paramétricos (Weibull e exponencial) como casos particulares. Já o termo paramétrico $g(\boldsymbol{x}'\boldsymbol{\beta})$ é geralmente utilizado na seguinte forma:

$$g(\boldsymbol{x}'\boldsymbol{\beta}) = \exp\left\{\boldsymbol{x}'\boldsymbol{\beta}\right\} = \exp\left\{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right\}.$$

Vale observar que o intercepto, presente nos modelos paramétricos de sobrevivência não é incluído na equação acima. Isso ocorre devido à parte não-paramétrica do modelo que absorve essa constante. A equação (2.8) pode ser reescrita da seguinte forma:

$$h(t \mid \boldsymbol{x}) = h_0(t) \exp\left\{\boldsymbol{x}'\boldsymbol{\beta}\right\}. \tag{2.9}$$

Essa forma assume que as covariáveis têm um efeito multiplicativo na função de risco, logo, a razão das taxas de falha para dois indivíduos *a* e *b*, com covariáveis $\boldsymbol{x}_a = (x_{a1}, x_{a2}, ..., x_{ap})'$ e $\boldsymbol{x}_b = (x_{b1}, x_{b2}, ..., x_{bp})'$ é:

$$\frac{h_a(t \mid \boldsymbol{x}_a)}{h_b(t \mid \boldsymbol{x}_b)} = \frac{h_0(t) exp\{\boldsymbol{x}_a'\boldsymbol{\beta}\}}{h_0(t) exp\{\boldsymbol{x}_b'\boldsymbol{\beta}\}} = exp\left\{\left(\boldsymbol{x}_a' - \boldsymbol{x}_b'\right)\boldsymbol{\beta}\right\} = K,$$

ou seja, é constante no tempo. Por exemplo, se $K = 2$, pode-se afirmar que o indivíduo *a* tem uma taxa de falha igual a duas vezes a do indivíduo *b*, sendo esta razão a mesma para todo o período de acompanhamento. Por esse motivo, o modelo de Cox é também denominado modelo de riscos proporcionais, sendo a proporcionalidade dos riscos uma das principais suposições feitas para esse modelo.

Para estimar os parâmetros $\beta's$ do modelo de Cox (2.9), onde a função base de taxa de falha é arbitrária, métodos usuais como os da máxima verossimilhança não podem ser empregados devido à presença do componente não paramétrico $h_0(t)$ na função de verossimilhança $L(\beta)$. Sabe-se que, no contexto de análise de sobrevivência, existem observações censuradas e não censuradas, sendo que a contribuição de cada observação não censurada para $L(\beta)$ é a sua função densidade e para as observações censuradas o que se sabe apenas é que o tempo de falha é maior que o tempo observado na censura, e, portanto, que sua contribuição para $L(\beta)$ é a sua função de sobrevivência (Colosimo & Giolo, 2006), dada por (2.7). Assim, a função de verossimilhança é dada por:

$$L(\beta) = \prod_{i=1}^{n} \left[h\left(t_i \middle| \boldsymbol{x}_i\right) \right]^{\delta_i} S\left(t_i \middle| \boldsymbol{x}_i\right), \tag{2.10}$$

e conforme Colosimo & Giolo (2006), a equação acima pode ser expressa por:

$$L(\beta) = \prod_{i=1}^{n} \left[h_0\left(t_i\right) \exp\left\{ \boldsymbol{x}_i' \boldsymbol{\beta} \right\} \right]^{\delta_i} S_0\left(t_i\right)^{\exp\left\{ x_i' \boldsymbol{\beta} \right\}}. \tag{2.11}$$

Note que a função de verossimilhança acima depende de $h_0\left(t\right)$. Diante disso, Cox em 1972, no mesmo artigo em que introduziu o modelo, propôs uma função de verossimilhança parcial baseada na probabilidade condicional de falha, ou seja, consiste em condicionar a obtenção da função de verossimilhança ao conhecimento anterior de falha e censura, visando eliminar a função base de risco $h_0(t)$. A vantagem desse método proposto por Cox é a facilidade em obter as estimativas para os parâmetros, não sendo necessário estimar $h_0(t)$ ou a função distribuição acumulada da taxa base de falha $H_0(t)$. No entanto, existem outros métodos, como o proposto por Breslow (1972), que estima $H_0(t)$ por uma função escada com saltos nos tempos de falha distintos, permitindo assim também estimar as funções de sobrevivência $S_0(t)$ e $S(t|\boldsymbol{x})$. A seguir será apresentado o método de máxima verossimilhança parcial proposto por Cox (1972) e formalizado pelo mesmo autor em um artigo posterior (Cox, 1975).

2.6 Método da Máxima Verossimilhança Parcial

Suponha que *m* dos tempos de falha de *n* indivíduos são não censurados e distintos, e *n* - *m* são censurados à direita. Sejam $t_1 < t_2 < \ldots < t_m$ os *m* tempos de falha ordenados com as seguintes covariáveis $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m$ e $R(t_i)$ o conjunto dos índices dos indivíduos sob risco no tempo t_i. A probabilidade condicional da *i*-ésima observação vir a falhar no tempo t_i, sabendo quais observações estão sob risco em t_i é (Colosimo & Giolo, 2006):

$$\mathrm{P}\left[\text{indivíduo falhar em } \mathrm{t}_i \,\middle|\, \text{uma falha em } \mathrm{t}_i \text{ e história até } \mathrm{t}_i\right] =$$

$$= \frac{\mathrm{P}\left[\text{indivíduo falhar em } \mathrm{t}_i \,\middle|\, \text{sobreviveu a } \mathrm{t}_i \text{ e história até } \mathrm{t}_i\right]}{\mathrm{P}\left[\text{uma falha em } \mathrm{t}_i \,\middle|\, \text{história até } \mathrm{t}_i\right]} =$$

$$= \frac{h_i(t|\boldsymbol{x}_i)}{\sum_{j\in R(t_i)} h_j(t|\boldsymbol{x}_j)} = \frac{h_0(t)\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)} h_0(t)\exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\}} = \frac{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}\exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\}} \qquad (2.12)$$

Ao utilizar a probabilidade condicional verifica-se que o componente não paramétrico $h_0(t)$ desaparece da equação (2.12). Desta forma, a função de verossimilhança é formada pelo produto de todos os termos representados por (2.12) associados aos diferentes tempos de falha. Assim,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}\exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\}} = \prod_{i=1}^{n}\left(\frac{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}\exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\}}\right)^{\delta_i}, \qquad (2.13)$$

em que δ_i é uma variável indicadora em que 1 representa o i-ésimo tempo de sobrevivência não censurado e 0 censurado. Os valores de $\boldsymbol{\beta}$ que maximizam (2.13) são obtidos pela resolução do sistema de equações dado por $\mathrm{U}(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$, em que $l(\boldsymbol{\beta}) = \log\left(L(\boldsymbol{\beta})\right)$, ou seja,

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n}\delta_i\left[x_i - \frac{\sum_{j\in R(t_i)} x_j \exp\{\boldsymbol{x}_j'\hat{\boldsymbol{\beta}}\}}{\sum_{j\in R(t_i)}\exp\{\boldsymbol{x}_j'\hat{\boldsymbol{\beta}}\}}\right] = 0. \qquad (2.14)$$

Pelo fato das equações encontradas em (2.13) não apresentarem forma fechada, geralmente utiliza-se o método iterativo de Newton-Raphson como procedimento de estimação. No caso de empates nos valores observados, a equação (2.13) deve ser modificada. Mais detalhes em Colosimo & Giolo (2006).

2.7 Modelo Misto de Cox

Os modelos de riscos proporcionais são geralmente usados para avaliar indivíduos não relacionados. No entanto, é comum considerar em análise genética que os tempos de sobrevivência são relacionados. Em animais, por exemplo, é natural supor que existe correlação entre os tempos dos animais de uma mesma ninhada ou em humanos, entre os tempos dos membros de uma mesma família. De acordo com Giolo (2003), as correlações surgem devido a influências genéticas ou ambientais compartilhadas.

Para considerar a existência de correlação, uma classe de modelos denominada modelos de fragilidade tem sido proposta. Nessa classe de modelos, um efeito aleatório, denominado fragilidade, é introduzido na função de taxa de falha para descrever essa possível associação. Assim, de acordo com Giolo (2003), o nome modelo de fragilidade surge devido ao fato de que quanto maior for o valor da fragilidade, maior será o risco de uma falha ocorrer, ou seja, mais frágeis os indivíduos de um grupo estarão para falhar.

De acordo com Giolo & Demétrio (2011), a primeira abordagem baseada no conceito de fragilidade compartilhada, considerou um efeito aleatório comum sobre a taxa de falha de todos os indivíduos em um determinado grupo. No entanto, esses modelos apresentam algumas limitações. Uma delas é que todos os fatores de risco são assumidos serem os mesmos dentro de um determinado grupo, ou seja, os indivíduos são homogêneos, sendo que pode existir heterogeneidade também a nível individual. Além disso, estes modelos não incorporam relações genéticas entre indivíduos do mesmo grupo. Com isso, uma abordagem de modelos de fragilidade correlacionada foram propostos por Ripatti & Palmgreen (2000) e Therneau et al. (2003) para o caso de dados de sobrevivência de n indivíduos.

Um dos modelos que tem se destacado na análise de dados de sobrevivência com heterogeneidade é o modelo de Cox com fragilidade ou modelo misto de Cox,

em que um efeito aleatório é incluído na função de taxa de falha $h(t)$. Semelhante à forma já apresentada dos dados e considerando que os tempos de falha estão agrupados em n_i grupos e representados por $T_j = (T_{1j}, T_{2j}, \ldots, T_{n_j j})'$, a função de taxa de falha para o i – ésimo indivíduo no j – ésimo grupo é dada por

$$h_{ij}(t) = h_0(t) \exp\left\{ \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{g} \right\},\ i = 1, \ldots, n_j \text{ e } j = 1, \ldots q, \tag{2.15}$$

em que $h_{ij}(t)$ é a função de taxa de falha para T_{ij} condicionalmente ao efeito aleatório $\boldsymbol{g}$, $h_0(t)$ é a taxa base de falha não especificada, $\boldsymbol{\beta}$ é o vetor de efeitos fixos associados ao vetor de covariáveis $\boldsymbol{x}_{ij}$ de dimensão p e $\boldsymbol{g}$ é um vetor de efeitos aleatórios associados ao vetor de covariáveis $\boldsymbol{z}_{ij}$. Os efeitos aleatórios $\boldsymbol{g}$ são assumidos independentes e identicamente distribuídos com distribuição normal com média zero e matriz de covariâncias $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ que depende de um vetor d – dimensional de parâmetros $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$. A função densidade de probabilidade de $\boldsymbol{g}$ é denotada por $f(\boldsymbol{g}; \boldsymbol{\theta})$.

Devido as fragilidades serem consideradas variáveis não observadas (ou latentes), uma abordagem que pode ser utilizada para estimar modelos de sobrevivência com efeito aleatório é por meio do algoritmo EM (Esperança-Maximização) em que as fragilidades são estimadas no passo E do algoritmo e os valores dos $\beta's$ que maximizam a função de verossimilhança parcial são obtidos no passo M (Carvalho et al., 2011). Entretanto, de acordo com Therneau et al. (2003), o algoritmo EM é lento e as estimativas de variância exigem maior esforço computacional, além disso, o algoritmo não está implementado na maioria dos softwares estatísticos. Uma abordagem alternativa foi proposta por Therneau e Grambsch (2000), em que modelos de fragilidade podem ser abordados por meio de verossimilhança penalizada. Nesse caso, as fragilidades são tratadas como coeficientes de regressão adicionais, que são penalizados por uma função adicionada ao log-verossimilhança.

Sabe-se que o interesse maior quando se utiliza o modelo dado em (2.15) está na estimação dos parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\theta}$. Assumindo a independência condicional das

observações dentro de um determinado grupo de efeitos aleatórios g_j, a função de verossimilhança condicional $L(h_0,\boldsymbol{\beta}|\boldsymbol{g})$ é dada por

$$L(h_0,\boldsymbol{\beta}|\boldsymbol{g})=\prod_{j=1}^{q}\prod_{i=1}^{n_j}\left[h_{ij}(t)\right]^{\delta_{ij}}\exp\left\{-\int_0^t h_{ij}(v)\,dv\right\} \tag{2.16}$$

em que $h_{ij}(\cdot)$ é obtida conforme (2.15). A função de verossimilhança completa é dada por

$$L(h_0,\boldsymbol{\beta},\boldsymbol{g})=L(h_0,\boldsymbol{\beta}|\boldsymbol{g})\prod_{j=1}^{q} f(\boldsymbol{g};\boldsymbol{\theta}), \tag{2.17}$$

sendo que o logaritmo da função de verossimilhança condicional dada em (2.16) é

$$l\left(h_0,\boldsymbol{\beta}|\boldsymbol{g}\right)=\sum_{j=1}^{q}\sum_{i=1}^{n_j}\left[\delta_{ij}\left\{\log h_0(t)+\boldsymbol{x}_{ij}'\boldsymbol{\beta}+\boldsymbol{z}_{ij}'\boldsymbol{g}\right\}-H_0(t)\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}+\boldsymbol{z}_{ij}'\boldsymbol{g}\right)\right], \tag{2.18}$$

em que $H_0(t)=\int_0^t h_0(v)\,dv$ é a função distribuição acumulada da taxa base de falha. Integrando-se (2.17) em relação a $\boldsymbol{g}$, obtém-se a função de verossimilhança marginal, expressa por

$$\begin{aligned}L_{m\arg}(h_0,\boldsymbol{\theta},\boldsymbol{\beta}) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{j=1}^{q}\prod_{i=1}^{n_j}\left[\left(h_0(t)\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}+\boldsymbol{z}_{ij}'\boldsymbol{g}\right)\right)^{\delta_{ij}}\right.\\ &\qquad \left.\exp\left(-H_0(t)\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}+\boldsymbol{z}_{ij}'\boldsymbol{g}\right)\right)\right] f(\boldsymbol{g})\,d\boldsymbol{g}\\ &= c\left|\boldsymbol{\Sigma}(\boldsymbol{\theta})\right|^{-1/2}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left(-K_q(\boldsymbol{g})\right)d\boldsymbol{g}\end{aligned} \tag{2.19}$$

com $K_q(\boldsymbol{g})=l\left(h_0,\boldsymbol{\beta}|\boldsymbol{g}\right)-\frac{1}{2}\boldsymbol{g}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{g}$ e $l\left(h_0,\boldsymbol{\beta}|\boldsymbol{g}\right)$ dado em (2.18). Usando teorema de Laplace, Ripatti & Palmgren (2000) mostraram que o logaritmo da função de verossimilhança dada em (2.19) pode ser aproximado por

$$l_{m\arg}(h_0,\boldsymbol{\theta},\boldsymbol{\beta})\approx-\frac{1}{2}\log\left|\boldsymbol{\Sigma}(\boldsymbol{\theta})\right|-\frac{1}{2}\log\left|K_q''(\tilde{\boldsymbol{g}})\right|-K_q(\tilde{\boldsymbol{g}}),$$

em que K_q' e K_q'' denotam, respectivamente, as derivadas parciais de 1ª e 2ª ordem de K_q com relação a $\boldsymbol{g}$, e $\tilde{\boldsymbol{g}}=\tilde{\boldsymbol{g}}(\boldsymbol{\beta},\boldsymbol{\theta})$ é a solução para $K_q'(\tilde{\boldsymbol{g}})=0$. Para $\boldsymbol{\theta}$ fixado,

Ripatti & Palmgren (2000) relatam que os valores $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ e $\hat{\boldsymbol{g}}(\boldsymbol{\theta})$ que maximizam o logaritmo da função de verossimilhança penalizada dado por $K_q(\boldsymbol{g}) = l(h_0, \boldsymbol{\beta}|\boldsymbol{g}) - \frac{1}{2}\boldsymbol{g}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{g}$ também maximizam o logaritmo da função de verossimilhança parcial penalizada dado por

$$l_{ppl}(\boldsymbol{\beta}(\boldsymbol{\theta}), \boldsymbol{g}(\boldsymbol{\theta}), \boldsymbol{\theta}) = \sum_{j=1}^{q}\sum_{i=1}^{n_j}\left[\delta_{ij}\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{g}\right) - \log\sum_{sw\in R(t_{ij})}\exp\left(\boldsymbol{x}'_{sw}\boldsymbol{\beta} + \boldsymbol{z}'_{sw}\boldsymbol{g}\right)\right]$$

$$-\frac{1}{2}\boldsymbol{g}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{g} \qquad (2.20)$$

Com base em (2.20), as estimativas dos parâmetros são obtidas em duas etapas. Primeiramente, gera-se um valor inicial para $\boldsymbol{\theta}$ e o método iterativo de Newton-Raphson é utilizado para obter as estimativas $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ e $\hat{\boldsymbol{g}}(\boldsymbol{\theta})$. Em seguida, com base nos valores $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ e $\hat{\boldsymbol{g}}(\boldsymbol{\theta})$ obtidos, $\boldsymbol{\theta}$ é atualizado por maximizar a seguinte verossimilhança perfilada aproximada:

$$l_{ppl}(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \hat{\boldsymbol{g}}(\boldsymbol{\theta}), \boldsymbol{\theta}) \approx -\frac{1}{2}\left|\log\boldsymbol{\Sigma}(\boldsymbol{\theta})\right| - \frac{1}{2}\log\left|K''(\hat{\boldsymbol{g}})\right| - \frac{1}{2}\hat{\boldsymbol{g}}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\hat{\boldsymbol{g}}. \qquad (2.21)$$

Ripatti & Palmgren (2000) propuseram usar em (2.21) $K''_{ppl}(\hat{\boldsymbol{g}}) = \frac{\partial^2 l_{ppl}}{\partial g \partial g'}$. Para obter os erros-padrão dos efeitos fixos estimados, pode-se utilizar o modelo de Cox usual com os efeitos aleatórios estimados como um *offset*, isto é, segundo Pinto Júnior (2009), o termo *log(g)* é adicionado ao preditor linear durante o ajuste com coeficiente conhecido. Para calcular os erros-padrão das estimativas de variância e covariância, Ripatti & Palmgren (2000) sugeriram a obtenção do valor esperado em relação ao efeito aleatório, da segunda derivada de (2.21) em relação a $\boldsymbol{\theta}$. Maiores detalhes podem ser encontrados em Ripatti & Palmgren (2000); Therneau et al. (2003); Abrahantes & Burzykowski (2005) e Duchateau & Janssen (2008).

Pankratz et al. (2005) apresentam um paralelo entre modelos lineares mistos e modelos de riscos proporcionais com efeitos aleatórios e citam pelo menos três principais diferenças entre esses modelos. A primeira delas é que os efeitos que são modelados na regressão de riscos proporcionais são multiplicativos, apresentando

forma aditiva na escala logarítmica, o que leva a implicações na interpretação dos resultados, pois, as estimativas nos modelos de riscos proporcionais com fragilidade refletem um risco relativo de se atingir um estado do evento de interesse, em vez de riscos aditivos.

A segunda diferença é que no modelo misto de Cox, a herdabilidade não pode ser obtida diretamente, como ocorre no modelo linear misto, pois não existe o componente de variância do erro aleatório. No entanto, os componentes de variância obtidos a partir do modelo de Cox com fragilidade podem ser interpretados diretamente, pois são modelados na escala logarítmica de taxa de falha. E por último, a terceira diferença se refere aos métodos computacionais utilizados. Em modelos lineares mistos, a estimação dos parâmetros é feita diretamente com base na maximização da log-verossimilhança, enquanto que, para modelos de riscos proporcionais com efeitos aleatórios, como o modelo misto de Cox, utiliza-se a aproximação de Laplace. Mesmo após empregar esta aproximação, as demandas computacionais ainda são extensas.

A respeito do cálculo da herdabilidade, Ducrocq (1987) demonstrou que a mesma, na escala latente com função de ligação complemento log-log, para o modelo de riscos proporcionais de Weibull, pode ser aproximada assumindo que o efeito do erro aleatório segue uma distribuição valor extremo com variância igual a $\pi^2/6$, o que posteriormente foi estendido para o modelo de riscos proporcionais de Cox por Korsgaard et al. (1999). Já Yazdi et al. (2002) demonstraram que para o modelo de Weibull, a herdabilidade pode ser aproximada em escala linearizada, de acordo com Resende et al. (2002), com base somente na variância genética e na proporção de dados censurados, sendo essa aproximação utilizada para o modelo de Cox nos trabalhos de Schneider et al. (2005) e Anderson et al. (2007). Assim, a expressão da herdabilidade é dada por:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{1}{1-c}}, \qquad (2.22)$$

em que *c* é a proporção de dados censurados.

3. REFERÊNCIAS BIBLIOGRÁFICAS

ABRAHANTES, J. C.; BURZYKOWSKI, T. A version of the EM algorithm for proportional hazard model with random effects. **Biometrical Journal**, v. 47, n. 6, p. 847 – 862, 2005.

ANDERSON, C. A.; DUFFY, D. L.; MARTIN, N. G.; VISSCHER, P. M. Estimation of variance components for age menarche in twin families. **Behavior genetics**, v. 37, n. 5, p. 668 – 677, 2007.

ANUÁRIO BRASILEIRO DE AVES E SUÍNOS 2011. Santa Cruz do Sul: Editora Gazeta Santa Cruz, 2011. 116 p. Disponível em: <http://www.gaz.com.br/tratadas/eo_edicao/21/2011/05/20110511_7263f739d/pdf/2804_aves2011_flip.pdf>. Acesso em: 27 nov. 2012.

AZEVEDO, C. F. **Métodos de redução de dimensionalidade aplicados na seleção genômica para características de carcaça em suínos**. 2012, 59 f. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2012.

BRESLOW, N. E. Discussion of Professor Cox' paper. **Journal of the Royal Statistical Society B**, v. 34, n. 2, p. 216 – 217, 1972.

CARVALHO, M. S.; ANDREOZZI, V. L.; CODEÇO, C. T.; CAMPOS, D. P.; BARBOSA, M. T. S.; SHIMAKURA, S. E. **Análise de sobrevivência: teoria e aplicações em saúde**. Rio de Janeiro: Editora Fiocruz, 2011. 432 p.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência**. São Paulo: Edgard Blücher, 2006. 369 p.

COX, D. R. Regression models and life tables (with discussion). **Journal Royal Statistical Society**, B, v. 34, n. 2, p. 187 – 220, 1972.

COX, D. R. Partial likelihood. **Biometrika**, v. 62, n. 2, p. 269 – 276, 1975.

DE LOS CAMPOS, G.; GIANOLA, D.; ALLISON, D. B. Predicting genetic predisposition in humans: the promise of whole-genome markers. **Nature Reviews Genetics**, London, v. 11, p. 880 – 886, 2010.

DE LOS CAMPOS, G.; HICKEY, J. M.; PONG-WONG, R.; DAETWYLER, H. D.; CALUS, M. P. L. Whole Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. **Genetics**, doi:10.1534/genetics.112.143313, 2012.

DUCROCQ, V. **An analysis of productive life in dairy cattle**. Ph.D. Dissertation, Cornell University, Ithaca, New York, 1987.

DUCHATEAU, L.; JANSSEN, P. **The Frailty Model**. New York: Springer, 2008. 316 p.

DUCROCQ, V. **Survival Analysis applied to Animal Breeding and Epidemiology**. Station de Génétique Quantitative et Appliquée , INRA , France, 1997. 74 p.

GIOLO, S. R.; HENDERSON, R.; DEMÉTRIO, C. G. B. Um critério para a seleção de touros nelore usando modelos de sobrevivência. **Revista Brasileira de Biometria**, v. 21, n. 3, p. 115 – 223, 2003.

GIOLO, S. R. **Variáveis latentes em análise de sobrevivência e curvas de crescimento**. 2003, 100 f. Tese (Doutorado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2003.

GIOLO, S. R.; DEMÉTRIO, C. G. B. A frailty modeling approach for parental effects in animal breeding. **Journal of Applied Statistics**, v. 38, n. 3, p. 619 – 629, 2011.

GODDARD, M. E.; HAYES, B. J. Genomic selection. **Journal of Animal Breeding and Genetics**, v. 124, n. 6, p. 323 – 330, 2007.

GOUVÊA, G. D. R. **Métodos Bayesianos para análise de dados de eventos recorrentes considerando uma classe geral de modelos com fragilidade multiplicativa**. 2010, 151 f. Tese (Doutorado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, 2010.

KORSGAARD, I. R.; ANDERSEN, A .H.; JENSEN, J. Discussion of heritability of survival traits. In Proc. Int. **Workshop on Genetic Improvement of Functional Traits in cattle**, Longevity, Jouy-en-Josas, France. INTERBULL Bull. No. 21, p. 31. Int. Bull Eval. Serv., Uppsala, Sweden, 1999.

LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics**, v. 124, n. 3, p. 743-756, 1990.

LEE, E. T.; WANG, W. J. **Statistical methods for survival data analysis**. Wiley series in probability and statistics, 2003. 513 p.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819 – 1829, 2001.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to theory of statistics**. 3. Ed. New York: J. Wiley & Sons, 1974. 564 p.

PANKRATZ, V. S.; ANDRADE, M.; THERNEAU, T. M. Random-effects Cox proportional hazards model: general variance components methods for time-to-event data. **Genetic Epidemiology**, v. 28, n. 2, p. 97 – 109, 2005.

PASCOA, M. A. R. **Intervalos de credibilidade para a razão de riscos do modelo de Cox, considerando estimativas pontuais bootstrap**. 2008, 84 f. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, 2008.

PINTO JÚNIOR, J. A. **Seleção de covariáveis para modelos de sobrevivência via verossimilhança penalizada**. 2009, 97 f. Dissertação (Mestrado em Estatística), Universidade de São Paulo, São Paulo, 2009.

RESENDE, M. D. V.; LOPES, P. S.; SILVA, R. L.; PIRES, I. E. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, n. 56, p. 63 – 78, 2008.

RESENDE, M. D. V. **Genômica Quantitativa e Seleção no Melhoramento de Plantas Perenes e Animais.** Colombo: EMBRAPA Florestas, 2008. 330p.

RESENDE M. D. V.; SILVA, F. F.; VIANA, J. M. S.; PETERNELLI, L. A.; RESENDE JUNIOR, M. F. R.; VALLE, P.R.M. **Métodos estatísticos na seleção genômica ampla**. Colombo: Embrapa Florestas. 2011. 106 p.

RESENDE, M. D. V.; SILVA, F. F.; LOPES, P. S.; AZEVEDO, C. F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial**. Viçosa: Universidade Federal de Viçosa/Departamento de Estatística. 2012. 291 p. <http://www.det.ufv.br/ppestbio/corpo_docente.php>. Acesso em: 07 jan. 2013.

RESENDE JR., M. F. R. **Seleção genômica ampla no melhoramento vegetal**. UFV, 2010. 67 p. Dissertação (Mestrado em Genética e Melhoramento) – Universidade Federal de Viçosa, Viçosa, 2010.

RIPATTI S, PALMGREN J. Estimation of multivariate frailty models using penalized partial likelihood. **Biometrics**, v. 56, n. 4, p. 1016–1022, 2000.

ROCHA, G. S. **Métodos estatísticos na seleção genômica ampla para curvas de crescimento em animais**. 2011, 56 f. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2011.

SCHNEIDER, M. D. P.; STRANDBERG, E.; DUCROCQ, V.; ROTH, A. Survival analysis applied to genetic evaluation for female fertility in dairy cattle. **Journal Dairy Science**, v. 88, n. 6, p. 2253 – 2259, 2005.

SIMKO, I.; EUJAYL, I.; VAN HINTUM, T. Empirical evaluation of DArT, SNP, and SSR marker-systems for genotyping, clustering, and assigning sugar beet hybrid varieties into populations. **Plant Science,** v. 184, p. 54 – 62, 2012.

THERNEAU, T. M.; GRAMBSCH, P. M. **Modeling survival data: extending the Cox model**. New York: Springer-Verlag, 2000. 350 p.

THERNEAU, T. M.; GRAMBSCH, P. M.; PANKRATZ, V. S. Penalized survival models and frailty. **Journal of Computational and Graphical Statistics**, v. 12, n. 1, p. 156 – 175, 2003.

YAZDI, M. H.; VISSHER, P. M.; DUCROCQ, V. THOMPSON, R. Heritability, reability of genetic evaluations and response to selection in proportional hazard models. **Journal Dairy Science**. v. 85, p. 1563 – 1577, 2002.

SELEÇÃO GENÔMICA AMPLA PARA IDADE AO ABATE EM SUÍNOS POR MEIO DO MODELO MISTO DE COX

1. INTRODUÇÃO

A idade ao abate influencia significativamente a qualidade da carne, ocasionando variações na composição e nas características metabólicas dos músculos. Assim, é economicamente desejável identificar e selecionar animais que apresentem maior ganho de peso em tempos mais curtos. Uma forma de verificar isso é por meio da variável idade ao abate, em que se espera que os melhores animais apresentem tempos não muito longos para um específico ganho de peso ao abate. No entanto, o tempo exato em que o animal obteve o peso desejado é desconhecido, uma vez que a pesagem diária é inviável.

Uma forma de classificar o tempo dos animais que obtiveram o ganho de peso desejável para o abate é por meio de uma variável independente da variável resposta tempo, denominada *status*, em que os animais com peso maior que o pré-estabelecido recebem valor 1 e o tempo entre o nascimento e o abate é classificado como tempo de falha e os animais que não alcançaram o peso pré-estabelecido recebem valor 0, sendo o tempo classificado como censurado. Neste trabalho, o peso desejável para o abate dos animais foi estabelecido em torno de 65 kg.

Dados caracterizados pela presença de censura e por não apresentarem distribuição normal são denominados dados de sobrevivência (Colosimo & Giolo, 2006). Em estudos genéticos, indivíduos de uma mesma família ou animais de uma mesma ninhada apresentam tempos de sobrevivência correlacionados entre si, por exemplo, devido ao parentesco; violando a pressuposição de independência entre os tempos dos indivíduos ou animais nos modelos de riscos proporcionais, tais como o modelo original de Cox. Uma forma de analisar dados dessa natureza é por meio de modelos de fragilidade correlacionada ou modelo de sobrevivência com efeitos aleatórios.

Vários trabalhos tem utilizado modelos de sobrevivência no melhoramento animal. Em gado de leite, por exemplo, podem ser citados os trabalhos de Ojango et al. (2005); Roxstrom et al. (2001) e Buenger et al. (2001), os quais avaliaram o

intervalo de dias de permanência de vacas no rebanho; e em gado de corte, os trabalhos de Giolo & Demétrio (2011) e Giolo et al. (2003), os quais avaliaram o intervalo de dias necessário para obter um ganho de peso padrão, em animais da raça Nelore. No melhoramento de suínos, podem-se citar os trabalhos de Mészáros et al. (2010) e Tarres et al. (2006) que avaliaram o intervalo de dias entre o primeiro parto e o abate de matrizes das raças Large White e Landrace por meio do modelo de Weibull, e o trabalho de Brandt et al. (1999) que analisaram a mesma variável por meio do modelo de Cox.

A seleção genômica ampla (GWS) surgiu em 2001 e consiste no uso de um grande número de marcadores genéticos (SNPs) os quais cobrem o genoma de uma maneira densa, a fim de predizer os valores genéticos de indivíduos, os quais podem ser pessoas em que se deseja predizer o risco genético de desenvolver determinada doença, animais ou plantas em que as estimativas dos valores genéticos serão usadas para selecionar indivíduos para a próxima geração (Goddard et al., 2011).

No melhoramento animal, a seleção dos melhores indivíduos é baseada nos valores genéticos estimados (EBV's) por meio do uso de dados fenotípicos e da matriz de parentesco usando o método conhecido como BLUP (*Best Linear Unbiased Prediction – melhor preditor linear não-viesado*). De acordo com Resende et al. (2012) e Azevedo (2012), com o advento de marcadores, entre eles, os SNPs (*Single Nucleotide Polymorphisms*), tem-se um aumento na acurácia da seleção de indivíduos geneticamente superiores, visto que, utilizam-se além de informações fenotípicas, informações oriundas do desequilíbrio de ligação entre marcadores e locos de características quantitativas (*Quantitative Trait Loci* – QTL).

Na GWS muitos métodos estatísticos têm sido propostos para estimar os efeitos de marcadores na população de estimação. Com base nos pressupostos sobre a distribuição dos efeitos de marcador, Zhang et al. (2010) afirmam que estes métodos podem ser classificados em dois grupos. O primeiro grupo assume que todos os marcadores têm algum efeito sobre a característica de interesse e que a variância de cada efeito de marcador é igual, sendo o método RR-BLUP o mais utilizado nesse contexto. Já o segundo grupo contempla os métodos que permitem assumir diferentes distribuições para os efeitos de marcadores, os quais se destacam os métodos bayesianos (Bayes A, Bayes B, Bayes Cπ, Bayes Dπ).

Ainda segundo os autores, uma alternativa para estimar os valores genéticos genômicos (GEBV's), é por meio da estimação direta no contexto das equações de modelo misto. No método clássico de seleção, uma matriz de parentesco baseada no *pedigree* é usada para descrever a dependência entre todos os pares de indivíduos na população. Os elementos dessa matriz são dados por duas vezes o valor esperado que dois alelos distribuídos aleatoriamente de um mesmo locus em dois indivíduos sejam idênticos por descendência (IBD).

Nos últimos anos, com a disponibilidade de um grande número de marcadores cobrindo o genoma, Nejati-Javaremi et al. (1997) propuseram substituir a matriz de parentesco baseada no *pedigree* por uma matriz de parentesco realizada ou matriz de parentesco baseada em marcadores, em que, no contexto da seleção genômica ampla, esse método é denominado GBLUP ou BLUP genômico (Van Raden, 2008), o qual foi mostrado ser equivalente ao RR-BLUP (Hayes et al., 2009; Goddard, 2009; Habier et al., 2007). No melhoramento genético de suínos, esse método pode ser visto no trabalho de Su et al. (2012).

No melhoramento animal, existem na literatura alguns trabalhos em que são comparadas as metodologias utilizando modelos lineares e modelos de sobrevivência contemplando informações de parentesco entre os animais. Entre esses trabalhos, pode ser citado o de Hou et al. (2009), que compararam cinco modelos (modelo linear convencional, modelo linear com limiar, modelo de riscos proporcionais de Weibull e o modelo de riscos proporcionais de Cox com uma função base de risco "*piecewise*") para avaliação genética do intervalo entre o parto e a primeira inseminação e o período entre o parto e a concepção em vacas *Danish Holstein.* Os autores concluíram que o modelo de Cox foi o que apresentou melhor desempenho para predizer os valores genéticos das duas características avaliadas.

De forma geral, vale ressaltar que o modelo linear misto não contempla a possibilidade de considerar observações censuradas e pressupõem-se dados normais para a variável resposta. Vale notar também que, no modelo de Cox, o que se modela é o risco, logo, quanto maior o risco, menor o intervalo até a ocorrência do evento de interesse. Já no modelo linear misto, o que está sendo modelado é a variável tempo diretamente. Devido a essas diferenças, espera-se obter correlações negativas entre os efeitos estimados em ambos os modelos.

Diante do exposto, o objetivo principal deste trabalho foi comparar a metodologia GBLUP usando o modelo linear misto e o modelo de sobrevivência de Cox, aplicados aos dados reais de idade ao abate de uma população F_2 de suínos (Comercial x Piau), além de avaliar diferentes subconjuntos de marcadores e selecionar aquele com maior acurácia.

2. MATERIAL E MÉTODOS

2.1 Descrição dos Dados

Os dados utilizados neste estudo foram cedidos pelo Departamento de Zootecnia da Universidade Federal de Viçosa (UFV) e constam de 345 suínos oriundos de uma população F_2 obtida do cruzamento de dois varrões da raça brasileira Piau com 18 fêmeas de linhagem desenvolvida na UFV, pelo acasalamento de animais de raça comercial (Landrace × Large White × Pietrain). A formação da população e a coleta dos dados fenotípicos foram realizadas na Granja de Melhoramento de Suínos do Departamento de Zootecnia da Universidade Federal de Viçosa (UFV), em Viçosa, Minas Gerais, Brasil, no período de novembro de 1998 a julho de 2001.

A extração do DNA dos animais foi realizada no Laboratório de Biotecnologia Animal do Departamento de Zootecnia da Universidade Federal de Viçosa e a genotipagem para os 384 SNPs selecionados de acordo com seu espaçamento entre cromossomos que continham QTLs previamente detectados foi realizada via tecnologia Golden Gate/VeraCode®, para o leitor BeadXpress de Illumina, no Laboratório de Genética Animal (LGA), Embrapa Recursos Genéticos e Biotecnologia (CENARGEN), Brasília, DF. Dos 384 SNPs, 66 foram descartados devido a ausência de amplificação, e 81 foram descartados por apresentar menor frequência alélica (MAF < 0,05), restando 237 marcadores distribuídos da seguinte forma nos cromossomos de *Sus scrofa*: SSC1 (56), SSC4 (54), SSC7 (59), SSC8 (31), SSC17 (25) e SSCX (12).

O gene halotano também foi considerado na análise como marcador adicional. Conhecido como gene do estresse, esse gene surgiu de uma mutação no

cromossomo 6 de suínos e está associado com carne PSE (pálida, flácida e exudativa). Sua presença contribui para o aumento do percentual de carne na carcaça (Band et al., 2005) porém, provoca o aumento de mortes súbitas, especialmente na movimentação e transporte dos animais quando não manejados adequadamente. Maiores detalhes sobre a constituição da população F_2 de suínos bem como detalhes sobre a extração do DNA e genotipagem dos indivíduos podem ser encontrados em Peixoto et al. (2006).

Os indivíduos da população F_2 possuem informações fenotípicas da idade ao abate e do peso dos animais em 7 momentos distintos (0, 21, 42, 63, 77, 105, 150 dias), em que o peso aos 150 dias indica o peso ao abate. Para alguns animais existem informações perdidas em determinadas idades.

2.2 Censura

Sabe-se que tempos não muito longos para um específico ganho de peso do nascimento até o abate de suínos é economicamente desejável, ou seja, é interessante identificar animais com ganho de peso mais rápido, dadas as mesmas condições ambientais. Segundo Band et al. (2005), o peso desejável para o abate em animais dessa população é em torno de 65 kg. Semelhante aos dados de gado Nelore analisados por Giolo et al. (2003) e Giolo & Demétrio (2011), não se conhece o tempo exato que um animal levou para ganhar o peso desejado de 65 kg, uma vez que a pesagem diária dos animais se torna inviável na prática. Sabe-se apenas a idade em que o animal foi abatido e sua pesagem.

Nesse contexto, para a análise do modelo de Cox, a censura foi criada, de forma independente à variável resposta, com base na variável peso ao abate, ou seja, animais que não alcançaram o peso de 65 kg foram considerados censurados e receberam valor 0 na variável indicadora de censura, e os animais que alcançaram esse peso ou superior foram considerados falha e receberam valor 1.

Os dados utilizados na análise foram previamente corrigidos para seus efeitos fixos de sexo e lote de manejo, sendo que o gene halotano foi considerado como marcador adicional. Essa correção se fez necessária, pois, segundo Resende et al. (2012), visa eliminar efeitos ambientais, reduzindo a amplitude de variação da população de mapeamento e tornando os resultados mais realísticos.

2.3 Método GBLUP

O método GBLUP consiste em predizer os valores genéticos genômicos substituindo a matriz de parentesco baseada no *pedigree* por uma matriz de parentesco baseada nos marcadores SNP's, denominada matriz de parentesco genômica (Van Raden, 2008). Assume-se nesse método que cada marcador explica igual proporção da variância genética, ou seja, estima-se os valores genéticos genômicos considerando um efeito poligênico em que todos os marcadores possuem variância constante. A seguir serão apresentados o modelo linear misto e o modelo misto de Cox utilizados no método GBLUP.

2.3.1 Modelo Linear Misto

O modelo linear misto utilizado foi definido da seguinte forma:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zg} + \boldsymbol{e}, \quad (1)$$

em que $\boldsymbol{y}$ é o vetor de tamanho *n* de fenotípicos corrigidos para seus efeitos fixos de sexo e lote de manejo; $\boldsymbol{\beta}$ é o vetor de efeitos fixos, $\boldsymbol{g}$ é o vetor dos efeitos genéticos aditivos individuais (efeitos aleatórios), com média **0** e matriz de variância e covariância $\boldsymbol{\Sigma} = \boldsymbol{G}_m\sigma_g^2$, em que $\boldsymbol{G}_m$ é a matriz de parentesco genômico nos locos marcadores; e $\boldsymbol{e}$ é o vetor de erros aleatórios normalmente distribuídos com media **0** e matriz de variância $\boldsymbol{I}\sigma_e^2$, sendo $\boldsymbol{X}$ e $\boldsymbol{Z}$ as matrizes de incidência para $\boldsymbol{\beta}$ e $\boldsymbol{g}$, respectivamente**.** Usando informações fenotípicas e dos marcadores, tem-se o modelo equivalente a (1), dado pela seguinte expressão:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{ZWm} + \boldsymbol{e}, \quad (2)$$

em que $\boldsymbol{g} = \boldsymbol{Wm}$, com variância de $\boldsymbol{Wm} = \boldsymbol{WW'}\sigma_{\boldsymbol{m}}^{\boldsymbol{2}}$ onde $\boldsymbol{W}$ é a matriz de incidência dos marcadores e $\boldsymbol{m}$ é o vetor dos efeitos aleatórios de marcas. A matriz de incidência $\boldsymbol{W}$ contém os valores w_{ij} iguais a 0, 1 e 2 ou -1, 0 e 1 para os marcadores mm, Mm e MM, respectivamente (Resende 2007; 2008; Resende et al., 2010; Resende et al., 2012). As equações do modelo linear misto para a predição de $\boldsymbol{g}$ no método GBLUP equivalem a:

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \boldsymbol{G_m^{-1}} \dfrac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{bmatrix},$$

em que $\boldsymbol{G}_m$ *é* a matriz de parentesco genômica. Resende et al. (2012) mostraram que a matriz de parentesco dos locos marcadores $\boldsymbol{G_m} = \boldsymbol{WW'}\sigma_{\boldsymbol{m}}^{\boldsymbol{2}} / \sigma_g^2$, uma vez que, igualando a variância genética assumida em (1) com a variância genética em (2), tem-se: $Var(\boldsymbol{g}) = Var(\boldsymbol{Wm})$, portanto $\boldsymbol{G_m}\sigma_g^2 = \boldsymbol{WW'}\sigma_{\boldsymbol{m}}^{\boldsymbol{2}}$. Desenvolvendo, tem-se: $\boldsymbol{G_m} = \boldsymbol{WW'} \Big/ 2\sum_i^k p_k (1 - p_k)$ pois $\sigma_g^2 = \left[2\sum_i^k p_k (1 - p_k) \right] \sigma_m^2$.

Segundo Resende et al. (2012), o valor genético genômico global de um individuo *j* no método GBLUP é dado por $\hat{g}_j = \sum_i w_{ij} \hat{m}_i$. A partir da estimação dos efeitos genéticos $(\hat{g})$ pelo GBLUP, os efeitos estimados de marcas $(\hat{m})$ podem ser obtidos por: $\hat{g} = W\hat{m} \Rightarrow W'\hat{g} = W'W\hat{m} \Rightarrow \hat{m} = (W'W)^{-1} W'\hat{g}$. Uma vez obtidas as variâncias genéticas aditivas e residual, o cálculo da herdabilidade é dado por:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}, \tag{3}$$

em que σ_g^2 é a variância explicada pelos efeitos genéticos aditivos e σ_e^2 é a variância residual.

2.3.2 Modelo Misto de Cox

Adotando forma semelhante ao modelo linear misto, dados em (1) e (2), e evitando algumas complexidades de notação da literatura sobre modelos de fragilidade, Therneau (2007) definiu o seguinte modelo misto de Cox, o qual, no contexto da GWS, é dado por:

$$h(t) = h_0(t) \exp\{\boldsymbol{X\beta} + \boldsymbol{Zg}\} \tag{4}$$

em que $h_0(t)$ é a taxa base de falha não especificada de forma semelhante ao modelo de Cox, $\boldsymbol{X}$ e $\boldsymbol{Z}$ são as matrizes de incidência para os efeitos fixos e aleatórios,

respectivamente, $\boldsymbol{\beta}$ é o vetor de efeitos fixos e $\boldsymbol{g}$ é o vetor dos efeitos genéticos aditivos individuais no modelo de Cox, assumidos normais com média **0** e matriz de covariâncias $\boldsymbol{\Sigma} = \boldsymbol{G}_m$. Semelhante ao modelo linear misto apresentado acima, a matriz de variâncias e covariâncias é dada com base na matriz de parentesco genômica que tem a vantagem sobre a matriz de parentesco baseada no *pedigree*, pois captura a proporção realizada de parentesco e não uma proporção média esperada como a calculada pela matriz de parentesco baseada no *pedigree*.

Assumindo que os efeitos aleatórios ou fragilidades $\boldsymbol{g}$ seguem distribuição normal com média **0** e matriz de variâncias e covariâncias $\boldsymbol{\Sigma} = \boldsymbol{G}_m\sigma_g^2$ e que a censura é independente e não informativa de $\boldsymbol{g}$, a função de verossimilhança parcial para o modelo misto de Cox é dada por

$$L = \int PL(\boldsymbol{\beta}, \boldsymbol{g}) \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}\boldsymbol{g}'\boldsymbol{\Sigma}^{-1}\boldsymbol{g}\right] d\boldsymbol{g}, \tag{5}$$

em que PL é a função de verossimilhança parcial para o modelo de Cox usual. Devido a integral acima não apresentar forma fechada, pois conforme Pankratz et al. (2005), PL é um produto de razões e o vetor de efeitos aleatórios $\boldsymbol{g}$ tem dimensão *n*, Ripatti & Palmgren (2000), com base nos métodos de Breslow & Clayton (1993), utilizaram aproximação de Laplace para obter o logaritmo da função de verossimilhança (5).

Conforme pode ser observado no modelo misto de Cox, o termo referente ao erro aleatório não é incluído no modelo. O fato é que, segundo Pankratz et al. (2005), esse termo é incorporado na função base de risco. Logo, a herdabilidade não pode ser diretamente calculada. Uma aproximação foi inicialmente proposta por Yazdi et al. (2002), e utilizada por Schneider et al. (2005) e Anderson et al. (2007) para o modelo misto de Cox, a qual, para o cálculo semelhante ao utilizado no modelo linear misto, é dada por:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{1}{1-c}}, \tag{6}$$

em que σ_g^2 é a variância explicada pelos efeitos genéticos aditivos e *c* é a proporção de dados censurados. Na verdade, os autores comentam que a variância do erro pode ser substituída por $\frac{1}{1-c}$. No caso de dados não censurados, a variância do erro passa a ser 1 com o erro aleatório na escala latente normal padrão com função de ligação probito, conforme apresentado em Resende (2002). Mézsaros et al. (2010) relatam que em alguns trabalhos envolvendo modelos de sobrevivência, a herdabilidade é computada assumindo erro aleatório na escala latente valor extremo, conforme Resende (2002), com função de ligação complemento log-log e variância do erro dada por $\frac{\pi^2}{6}$.

2.3.3 Matriz de Parentesco Genômica

A matriz de parentesco genômica foi formada conforme mostrado em Habier et al. (2007), Hayes et al. (2009) e Van Raden et al. (2009), em que a matriz $\boldsymbol{G}_m$ é dada por

$$\frac{\boldsymbol{W}^*\boldsymbol{W}^{*\prime}}{2\sum_i^k p_i\left(1-p_i\right)},$$

onde $\boldsymbol{W}^*$ é a matriz **W** corrigida para suas médias em cada loco ($2p_i$) e p_i é a frequência de um dos alelos no marcador *i*. A matriz **W** é codificada como -1, 0 e 1 para os marcadores mm, Mm e MM, respectivamente. Essa matriz foi computada por meio da função *A.mat* ($\boldsymbol{G}_m$) do pacote *RR-BLUP* do software livre R (Endelman, 2011). Para tornar $\boldsymbol{G}_m$ uma matriz positiva definida, foi obtida a matriz $\boldsymbol{G}_m^* = \boldsymbol{G}_m + 10^{-6}\boldsymbol{I}$, em que $\boldsymbol{I}$ é uma matriz identidade. Obtida a $\boldsymbol{G}_m^*$, a mesma foi utilizada para o cálculo das estimativas dos componentes de variância bem como dos efeitos aleatórios, para o modelo linear misto e de fragilidade de Cox, por meio das funções *lmekin* e *coxme* respectivamente, implementadas no pacote *coxme* do software livre R (Therneau, 2012).

2.4 Comparação entre os modelos

A fim de verificar se ambos os modelos linear e de Cox com efeitos aleatórios, nas situações de equivalência, são concordantes entre si em predizer os valores genéticos genômicos (GBV's) e em estimar os efeitos de marcas, foram comparados os seguintes modelos: modelo linear misto considerando todas as observações como não censuradas (L1), modelo linear misto com dados imputados (L2), modelo misto de Cox com censura (S1) e modelo misto de Cox sem censura (S2). Na presença de observações incompletas, uma forma de análise é por meio da imputação de dados, sendo a mais conhecida, a imputação pela média das observações. Ou seja, a fim de comparar os modelos L2 e S1, as observações da variável resposta que foram consideradas censuradas no modelo S1 foram substituídas pelo valor médio no modelo L2.

No modelo de Cox, a estimação via o método da máxima verossimilhança restrita (REML) é de difícil obtenção e não está implementada no pacote *coxme* do software R. Logo, para efeito de comparação, ambos os modelos, de Cox e linear, foram estimados por meio do método da máxima verossimilhança (ML).

2.4.1 Validação cruzada

A comparação dos modelos foi realizada por meio da técnica de validação cruzada jacknife, a qual foi inicialmente proposta por Quenouille (1956) e consiste, de modo geral, na divisão da amostra de tamanho *N* em *g* grupos de tamanho igual a *k*, de forma que *N*=*g* quando *k*=1 (Resende, 2002).

O procedimento foi realizado conforme citado em Resende et al. (2010) e Resende et al. (2012), no qual consistiu em excluir da análise um indivíduo em cada repetição a fim de compor a população de validação, sendo que os outros 334 indivíduos foram utilizados na estimação dos efeitos de marcas por meio da expressão $\hat{m} = (W'W)^{-1} W'\hat{g}$, ou seja, esses indivíduos constituíram a população de estimação. Depois de obtidos os efeitos de marcas, esses foram aplicados na população de validação para predição dos valores genéticos genômicos GEBV's por meio da expressão $\hat{g} = W\hat{m}$.

Após a obtenção dos GBV's preditos na validação cruzada, esses foram submetidos à correlação com os valores fenotípicos corrigidos a fim de avaliar a acurácia da predição de valores genéticos genômicos (GBV's) dos modelos utilizados em uma amostra independente. De acordo com Resende et al. (2010), Resende et al. (2012), essa correlação indica a capacidade do método em predizer de forma acurada os fenótipos.

No entanto, surge um problema em obter a correlação entre os GBV's preditos no modelo de fragilidade de Cox (S1 e S2) e o fenótipo, uma vez que os valores genéticos genômicos estimados nesses modelos estão representados na escala de taxa de falha e os fenótipos corrigidos na escala de dias. Uma forma para solucionar isso foi obter a correlação entre os postos dos GBV's ordenados de forma crescente para o modelo linear (L1 e L2) e de forma decrescente para o modelo misto de Cox com (S1) e sem censura (S2). Ou seja, os indivíduos candidatos à seleção foram aqueles que apresentaram menor tempo para o ganho de peso ao abate e maior fragilidade, e consequentemente os descartados foram aqueles que apresentarem maior tempo para o ganho de peso e menor fragilidade. O coeficiente de correlação entre os postos foi obtido pela seguinte expressão:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = r_s.$$

Ou seja, foi calculado o coeficiente de correlação de Pearson aplicado à dados ordenados, o qual equivale ao coeficiente de correlação de Spearman, dado pela seguinte expressão:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n},$$

onde d_i é a diferença de postos dos escores X e Y. Foram obtidas também a acurácia de cada modelo, a qual é dada pela razão da correlação entre os postos e a raiz quadrada da herdabilidade estimada via máxima verossimilhança (ML), conforme expressão abaixo:

$$r_{g\hat{g}} = \frac{r_{y\hat{y}}}{\sqrt{h^2}}.$$

Além da capacidade preditiva e da acurácia, foram obtidas também as correlações de Pearson e de Spearman entre os GBV's preditos na população de estimação e de validação nos quatro modelos utilizados.

A correlação entre os GBV's estimados nessas populações para os modelos L1 e S2 permite verificar se ambos os modelos na situação em que a variável é assumida normalmente distribuída e 0% de censura são concordantes entre si em predizer o fenótipo. E no caso dos modelos L2 e S1, uma forma mais simples de comparar os dois modelos é considerando a imputação dos valores perdidos para a variável resposta no modelo linear. Em ambas as comparações, esperam-se correlações negativas, uma vez que no modelo linear misto se modela a variável tempo diretamente, que nesse estudo, é o tempo para alcançar o peso ideal de abate, enquanto que no modelo de Cox, o que se modela é o risco.

2.5 Método GBLUP supervisionado

Além do método GBLUP, foi aplicado também o método denominado GBLUP supervisionado, o qual é semelhante o método RRBLUP-B aplicado por Resende Jr et al. (2012). Esse método consiste na análise do método GBLUP em duas etapas, no entanto, utiliza um número menor de efeitos de marcas estimadas.

Os seguintes passos foram efetuados na análise do método GBLUP supervisionado, conforme proposto por Resende et al. (2010): primeiramente foi computada a predição dos GBV's usando todos os marcadores e a correlação entre os GBV's preditos e os valores fenotípicos corrigidos é obtida. Em seguida, os marcadores são classificados em ordem decrescente com base nos seus efeitos estimados na população de estimação e agrupados em subconjuntos de marcas (2,5%, 5%, 10%, 20%, 30%, 40%, 50% e 75% do total de marcadores com maiores efeitos em módulo, correspondendo respectivamente a 6, 12, 24, 48, 72, 96, 120 e 180 marcas). O subconjunto de marcas que maximiza a capacidade preditiva obtida na população de estimação é selecionado. Em seguida, é feita a validação nesse subconjunto selecionado e em todos os menores e um maior para verificar tendências.

Estudos em pinheiros usando metodologia semelhante à citada aqui (Resende Jr et al., 2012) e com dados simulados (Zhang et al. 2010; Zhang et al. 2011), tem

mostrado um aumento significativo nos valores de capacidade preditiva, com resultados equivalentes aos obtidos por métodos bayesianos. A vantagem desse método, segundo Resende Jr et al. (2012), é sua simplicidade e demanda computacionalmente menor que os métodos bayesianos.

Além da seleção inicial de marcadores no método GBLUP supervisionado, foi construído também o gráfico denominado *Manhattan plot* com base nos efeitos estimados, na população de estimação, de todas as marcas e das selecionadas no GBLUP supervisionado, em que, cada ponto representa um marcador SNP, onde o eixo X mostra sua posição no cromossomo e o eixo Y o módulo de seu efeito (Azevedo, 2012).

2.6 Seleção de indivíduos

A fim de identificar possíveis indivíduos candidatos à seleção e ao descarte, foram obtidas as curvas de sobrevivência com base na média dos 10% maiores e 10% menores valores genéticos genômicos preditos no modelo de Cox com censura - S1 (modelo verdadeiro), além das curvas de sobrevivência individuais para o animal com maior e menor GBV predito. A função de sobrevivência para o modelo (4) é dada por (Giolo & Demétrio, 2011):

$$S\left(t|\boldsymbol{g}\right)=\left[S_0\left(t\right)\right]^{\exp\{\boldsymbol{X\beta}+\boldsymbol{Zg}\}}, \qquad (7)$$

em que $S_0\left(t\right)=\exp\left\{-H_0\left(t\right)\right\}$ é a função base de sobrevivência e $H_0\left(t\right)=\int_0^t h_0\left(v\right)dv$ é a função distribuição acumulada da taxa base de falha.

A fim de verificar a equivalência entre os modelos em selecionar os melhores indivíduos e os maiores efeitos de marcas, foram calculadas as taxas de concordância para os 10% maiores efeitos de GBV's preditos utilizando todas as 238 marcas e somente as selecionadas no método GBLUP supervisionado entre os modelos linear misto (L1) e do modelo de Cox sem censura (S2) e entre os modelos linear misto com imputação de dados (L2) e de sobrevivência de Cox com censura (S1). Também foram calculadas as taxas de concordância entre as médias dos maiores efeitos em valor absoluto de todas as marcas e as selecionadas no GBLUP supervisionado, obtidos nos quatro modelos pela validação cruzada.

2.7 Índice Kappa

Além do cálculo da taxa de concordância entre os modelos, foi obtido também o índice *Kappa* (*k*), proposto por Cohen (1960), no qual consiste em medir o grau de concordância entre dois julgamentos independentes. No presente trabalho, o índice *Kappa* foi utilizado para avaliar a concordância entre os dois modelos utilizados nas situações equivalentes: modelo linear misto com observações completas (L1) *versus* o modelo de sobrevivência de Cox sem censura (S2), e o modelo linear misto com imputação de dados (L2) *versus* o modelo de sobrevivência de Cox com censura (S1).

Fleiss et al. (1981) relatam que para uma única categoria *j*, pode-se obter um índice de concordância simples dado por $p_0 = a + d$ e ainda, as proporções esperadas ao acaso podem ser obtidas conforme apresentado na Tabela 1 abaixo.

Tabela 1: Proporção de concordância observada e esperada ao acaso em uma única categoria.

Proporção de Concordância				Proporção Esperada ao Acaso			
Método 1	Método 2			Método 1	Método 2		
	1	*j*-1	Total		1	*j*-1	Total
1	*a*	*b*	p_1	*1*	p_1p_2	p_1q_2	p_1
j-1	*c*	*d*	q_1	*j-1*	q_1p_2	q_1q_2	q_1
Total	p_2	q_2	*1*	*Total*	p_2	q_2	*1*

Assim, o índice *Kappa* para uma única categoria pode ser obtido por meio da seguinte expressão (Fleiss et al., 1981):

$$\hat{k} = \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1}, \tag{8}$$

em que *a*, *b*, *c*, e *d* referem-se às proporções das observações.

Segundo Fonseca et al. (2007), apesar de não existir um valor específico a partir do qual se deva considerar o valor do *kappa* como adequado, encontram-se na literatura algumas sugestões que orientam essa decisão, como a proposta por Landis & Koch (1977), apresentada na Tabela 2.

Tabela 2: Índice Kappa e sua correspondente classificação de desempenho.

Índice *Kappa*	Desempenho
$\hat{k} \leq 0,2$	Ruim
$0,2 < \hat{k} \leq 0,4$	Razoável
$0,4 < \hat{k} \leq 0,6$	Bom
$0,6 < \hat{k} \leq 0,8$	Muito bom
$0,8 < \hat{k} \leq 1,0$	Excelente

Fleiss et al. (1981) também relatam sobre a obtenção de um índice *Kappa global* baseado em todas as categorias *j*, o qual, segundo os autores, pode ser definido como uma média ponderada de todos os valores *kappas* individuais obtidos para cada categoria, onde os pesos são os denominadores desses índices individuais, ou seja, as quantidades $p_1q_2 + p_2q_1$ apresentadas em (8).

Tabela 3: Proporção conjunta de classificação por dois métodos em uma escala com *j* categorias.

	Método 2				
Método 1	1	2	⋯	*j*	Total
1	p_{11}	p_{12}	⋯	p_{1j}	p_{1+}
2	p_{21}	p_{22}	⋯	p_{2j}	p_{2+}
⋮	⋮	⋮	⋱	⋮	⋮
j	p_{j1}	p_{j2}	⋯	p_{jj}	p_{j+}
Total	p_{+1}	p_{+2}	⋯	p_{+j}	*1*

Assim, com base na Tabela 3, a proporção de concordância observada é dada por

$$p_0 = \sum_{i=1}^{j} p_{ii}$$

e a proporção de observações nas quais a concordância foi devida ao acaso é dada por

$$p_e = \sum_{i=1}^{j} p_{i+} p_{+i} \,.$$

O valor *kappa global* é dado então por

$$\hat{\bar{k}} = \frac{p_0 - p_e}{1 - p_e},$$

em que $p_0 - p_e$ representa a proporção de observações em que a concordância ocorreu além do que se esperava aleatoriamente e $1 - p_e$ a proporção de observações em que não ocorreu concordância.

Para avaliar a significância estatística do valor *Kappa global*, vários autores [Cohen (1960); Congalton & Green (1998); Fleiss et al. (1981); Galparsoro & Fernández (2001)] sugerem usar o teste Z, em que as hipóteses testadas são: H_0: $k = 0$ (a concordância entre os dois métodos é nula) *vs* H_1: $k > 0$ (concordância maior do que o acaso). Fleiss et al. (1981) afirmam que o teste unilateral é mais apropriado que o bilateral, uma vez que, segundo Landis & Koch (1977), valores negativos do índice *kappa* não tem interpretação plausível.

A estatística do teste Z é dada por (Fleiss et al., 1981):

$$Z = \frac{\hat{\bar{k}}}{\widehat{ep}(\hat{\bar{k}})} \sim N(0,1)$$

em que $\widehat{ep}\left(\hat{\bar{k}}\right) = \frac{1}{(1-p_e)\sqrt{n}}\sqrt{p_e + p_e^2 - \sum_{i=1}^{j} p_{i+}p_{+i}(p_{i+}+p_{+i})}$.

Se o valor do teste Z calculado for maior que o valor crítico Z_α, há indícios de que a classificação entre os dois métodos não ocorreu pelo acaso.

3. RESULTADOS E DISCUSSÃO

A análise foi feita com base nos modelos linear e de Cox com efeitos aleatórios, estimados via máxima verossimilhança nas situações de dados completos e censurados. Para estimar o modelo linear misto na presença de dados censurados (L2), os mesmos foram considerados como observações perdidas, sendo substituídos pela média amostral.

A censura no modelo de sobrevivência de Cox (S1) foi criada com base no peso ao abate dos animais, em torno de 65 kg. Ou seja, os tempos dos animais que apresentaram pesos ao abate menor que 65 kg, foram considerados tempos censurados e receberam o valor 0, e os tempos dos animais que ganharam peso maior que 65 kg, foram considerados tempos de falha e receberam o valor 1. A proporção de censura foi em torno de 0,561, ou seja, foi observado que cerca de 44% dos animais tiveram ganho de peso superior a 65 kg.

Nos modelos linear misto (L1) e de fragilidade de Cox (S2), a natureza da censura não foi levada em conta, ou seja, todos os indivíduos foram considerados não censurados. Isso foi feito com o objetivo de verificar a concordância em se estimar os valores genéticos genômicos por meio do modelo linear misto e do modelo misto de Cox. Com esse mesmo objetivo, também foram comparados os modelos linear misto com imputação de dados (L2) e o modelo de fragilidade de Cox com censura (S1). No modelo linear misto, os animais selecionados foram aqueles cujos correspondentes valores do efeito genético aleatório minimizaram o tempo de abate. Já no modelo misto de Cox, os animais selecionados foram aqueles cujos correspondentes valores do efeito aleatório maximizaram o risco.

As estimativas dos componentes de variância obtidas pelo método GBLUP para a variável idade ao abate em suínos, corrigida para os efeitos fixos de sexo e lote de manejo são apresentadas na Tabela 4. As estimativas de herdabilidade para o modelo linear e de Cox foram obtidas conforme as expressões (3) e (6), respectivamente, e foram em torno de 0,08 para o modelo linear com observações completas e de 0,02 para o modelo misto de Cox nas situações com e sem censura. Para o modelo L2, a herdabilidade foi menor do que a do modelo L1. Isso ocorreu provavelmente devido no modelo L2 os dados considerados perdidos terem sido substituídos pelo valor médio amostral.

Nota-se, de maneira geral, baixos valores de herdabilidade, conforme já esperado para a característica. Será mostrado mais adiante (Tabela 9), ao considerar na análise somente os marcadores de maior efeito, as estimativas de herdabilidade aumentaram para os quatro modelos em relação ao uso de todos os marcadores.

Tabela 4: Componentes de variância estimados pelo método GBLUP para a característica idade ao abate em suínos, considerando o modelo linear misto e de Cox, e todas as marcas.

Modelos	Componentes de variância		
	σ_g^2	σ_e^2	h^2
L1	9,103	108,985	0,077
L2	0,595	41,206	0,014
S1	0,057	-	0,024*
S2	0,018	-	0,018*

L1: Modelo linear misto com dados completos; L2: Modelo linear misto com dados imputados; S1: Modelo misto de Cox com censura; S2: Modelo misto de Cox sem censura. *Herdabilidade calculada pela expressão dada em (6).

A Tabela 5 apresenta as estimativas dos coeficientes de correlação envolvendo valores fenotípicos (y) e valores genéticos genômicos preditos ($\widehat{GBV}$) no método GBLUP, obtidos com base nos modelos L1, L2, S1 e S2 e com todos os 238 marcadores SNP's na população de estimação. Observa-se correlações negativas dos GBV's estimados nos modelos S1 (modelo misto de Cox com censura) e S2 (modelo misto de Cox sem censura) com os GBV's estimados nos modelos L1 (modelo linear misto com dados completos) e L2 (modelo linear misto com dados imputados). Isso já era esperado, uma vez que no modelo de Cox o efeito aleatório reflete o risco, enquanto que no modelo linear misto, reflete os dias que o animal leva para ganhar o peso de abate de 65 kg. Ou seja, os efeitos aleatórios estão em escalas diferentes, por isso ocorre a correlação negativa.

Observa-se também coeficientes de correlação de Pearson e de Spearman de menor magnitude entre os GBV's estimados no modelo linear misto - L1 com o modelo de fragilidade de Cox com censura - S1 (-0,67 e -0,62) em relação ao L1 com o modelo de fragilidade de Cox sem censura - S2 (-0,94 e -0,93), respectivamente, o que mostra a influência da censura, uma vez que no modelo linear misto (L1) considerou-se todos os dados não censurados. Já o modelo linear misto com dados imputados (L2) apresentou maiores coeficientes de correlação de Pearson e de Spearman com o modelo S1 (-0,72 e -0,69) do que com o modelo S2 (-0,63 e -0,59), respectivamente.

Tabela 5: Estimativas dos coeficientes de correlação envolvendo valores fenotípicos (y) e valores genéticos genômicos estimados ($\widehat{GBV}$) no método GBLUP, considerando o modelo linear misto e o modelo misto de Cox, com base em todos marcadores na população total.

Modelos	L1	L2	S1	S2
L1	0,42	0,62	-0,62	-0,93
L2	0,65	0,26	-0,69	-0,59
S1	-0,67	-0,72	0,22	0,53
S2	-0,94	-0,63	0,59	0,38

L1: Modelo linear misto com dados completos; L2: Modelo linear misto com dados imputados; S1: Modelo misto de Cox com censura; S2: Modelo misto de Cox sem censura. Na diagonal estão apresentadas as correlações entre os postos do $\widehat{GBV}$ e do y; acima da diagonal estão apresentadas as correlações de Spearman e abaixo da diagonal as correlações de Pearson entre os $\widehat{GBV}$'s estimados nos modelos utilizados.

Nota-se também que, na situação de equivalência, os valores dos coeficientes de correlação de Pearson (-0,94) e de Spearman (-0,93) entre os modelos L1 e S2 foram maiores do que entre os modelos L2 e S1 [-0,72 (Pearson) e -0,69 (Spearman)]. Um dos possíveis fatores dessa ocorrência é devido a amostra, uma vez que nos modelos L1 e S2 não se tem a presença de censura, enquanto que nos modelos L2 e S1 a censura foi levada em conta de alguma forma.

A capacidade preditiva entre os modelos linear e de fragilidade de Cox foi aproximadamente igual nas situações avaliadas. Na situação de dados não censurados, os modelos L1 e S2 apresentaram capacidade preditiva igual a 0,42 e 0,38, respectivamente; enquanto que os modelos L2 e S1 apresentaram capacidade preditiva igual a 0,26 e 0,22, respectivamente. Em ambos os casos, a capacidade preditiva foi maior nos modelos lineares do que nos modelos de sobrevivência.

As estimativas das correlações dos GBV's estimados com os fenótipos foram calculadas com base nos postos dos valores ordenados de forma decrescente para os GBV's estimados nos modelos de Cox e em ordem crescente para os GBV'S estimados nos modelos lineares e os fenótipos corrigidos, solucionando assim o problema de escala e tornando positivas as estimativas das correlações.

As estimativas de correlação entre os valores fenotípicos corrigidos e valores genéticos genômicos preditos ($\widehat{GBV}$) no modelo linear misto e modelo de Cox, pela validação cruzada e com base em todos os marcadores, estão apresentadas na Tabela 6.

Tabela 6: Estimativas dos coeficientes de correlação envolvendo valores fenotípicos corrigidos (y) e valores genéticos genômicos preditos $\widehat{GBV}$, pela validação cruzada, no método GBLUP, considerando o modelo linear misto e de fragilidade de Cox, e todos os marcadores.

Modelos	L1	L2	S1	S2
L1	0,08	0,58	-0,62	-0,91
L2	0,57	-0,07	-0,65	-0,53
S1	-0,66	-0,62	0,006	0,52
S2	-0,89	-0,55	0,60	-0,02

L1: Modelo linear misto com dados completos; L2: Modelo linear misto com dados imputados; S1: Modelo misto de Cox com censura; S2: Modelo misto de Cox sem censura. Na diagonal estão apresentadas as correlações entre os postos do $\widehat{GBV}$ e do y; acima da diagonal estão apresentadas as correlações de Spearman e abaixo da diagonal as correlações de Pearson entre os $\widehat{GBV}$'s estimados nos modelos utilizados.

Observa-se que as estimativas de correlação foram menores do que as obtidas na Tabela 5. Isso ocorre pelo fato de que os valores genéticos genômicos dos indivíduos na população de validação são preditos com base nos efeitos dos marcadores estimados por meio da população de estimação. Resultados similares foram observados por Rocha et al. (2011) e Resende et al. (2010).

Nota-se também que a capacidade do método em predizer os valores genéticos genômicos de forma acurada foram baixas e até negativas conforme pode ser visto das correlações entre os postos dos GBV's preditos nos quatro modelos e dos fenótipos corrigidos (valores na diagonal). Esses valores arbitrários de correlação foram devidos à utilização de todos os marcadores na análise. Segundo Resende et al. (2010), uma forma de maximizar a capacidade do método em predizer de forma acurada é por meio do cômputo dos valores genéticos genômicos (GBV) em dois passos, método esse denominado GBLUP supervisionado, conforme apresentado no item 2.5. Os resultados dessa análise são apresentados nas Tabelas 7 e 8.

No primeiro passo, os GBV's foram obtidos usando todos os marcadores e a correlação entre GBV e o fenótipo foi calculada. Em seguida, foram criados vários arquivos com subconjuntos de marcadores ordenados de acordo com os maiores módulos dos seus efeitos estimados e a correlação entre GBV e fenótipo ($r_{GBV,y}$) foi calculada para todos esses arquivos de subconjuntos de marcadores, sendo o arquivo ótimo aquele que maximizou a $r_{GBV,y}$. Foi feita então a validação cruzada com base no número de marcadores selecionados nesse arquivo ótimo e nos arquivos menores e um maior para ver tendências (Resende et al., 2010; Resende et al., 2012).

A Tabela 7 apresenta a correlação entre os postos na população de estimação considerando o modelo linear misto com dados completos (L1) e imputados (L2) e o modelo de fragilidade de Cox nas situações com (S1) e sem censura (S2), e com diferentes números de marcadores em ordem de maior efeito em módulo.

Como já citado neste trabalho, a correlação dos GBV's estimados pelos modelos lineares mistos e de fragilidade de Cox com o fenótipo foi obtida pela ordem crescente dos GBV's para o modelo linear e fenótipos corrigidos, e em ordem decrescente para o modelo de fragilidade de Cox com e sem censura. Isso foi feito a fim de contornar o problema de escalas diferentes em que os valores genéticos genômicos foram estimados nos quatro modelos.

Observa-se que para o modelo linear com dados completos (L1), a correlação entre os postos foi maior quando se utilizou os 96 marcadores de maiores efeitos em módulo, e praticamente igual quando se utilizou 120 marcadores. Já para o modelo linear misto com dados imputados (L2), o maior valor de correlação entre os postos foi para 120 marcas e praticamente igual quando se utilizou 96 e 180 marcas. Para o modelo de fragilidade de Cox nas situações com e sem censura, a maior correlação foi obtida também no uso de 120 marcadores. Vale notar também que as correlações dos GBV's estimados nos modelos L1 e S2 com o fenótipo corrigido foram próximas entre si e maiores do que as correlações entre os GBV's estimados nos modelos L2 e S1 com o mesmo fenótipo para todos os subconjuntos de marcas.

Tabela 7: Capacidade preditiva da GWS (correlação entre os postos) na população de estimação com base no modelo linear misto e de fragilidade de Cox.

Número de marcas	Modelos			
	L1	*L2*	*S1*	*S2*
2,5% (6 marcas)	0,38	0,27	0,19	0,34
5% (12 marcas)	0,34	0,28	0,20	0,37
10% (24 marcas)	0,38	0,27	0,21	0,39
20% (48 marcas)	0,44	0,28	0,26	0,42
30% (72 marcas)	0,46	0,29	0,25	0,44
40% (96 marcas)	**0,48**	0,30	0,27	0,46
50% (120 marcas)	0,47	**0,31**	**0,28**	**0,48**
75% (180 marcas)	0,46	0,30	0,27	0,45
100% (238 marcas)	0,42	0,26	0,22	0,38

L1: Modelo linear misto com dados completos; L2: Modelo linear misto com dados imputados; S1: Modelo misto de Cox com censura; S2: Modelo misto de Cox sem censura.

Na Tabela 8 são apresentados os valores da correlação entre os postos dos GBV's preditos e da acurácia dos modelos, obtidos pela técnica de validação cruzada jacknife nos modelos linear e de Cox com diferentes números de marcadores de maior efeito. A validação cruzada foi feita em todos os arquivos com número de marcadores menores que o ideal e um arquivo com número de marcadores maior que o sugerido pela capacidade preditiva na população de estimação.

Nota-se uma maior diferença nas correlações, nos quatro modelos avaliados (L1, L2, S1 e S2), quando se utiliza os marcadores mais significativos ao invés de todos. Por exemplo, a correlação entre os postos para o modelo linear misto foi de 0,08 (Tabela 6 e 8) e valor da acurácia de 0,28 utilizando todos os marcadores, e com

base nos 120 mais informativos, a correlação passou a ser de 0,22 e valor da acurácia de 0,55. O mesmo ocorreu para os modelos linear (L2), de fragilidade de Cox com (S1) e sem (S2) censura, em que as correlações passaram de -0,07 para 0,16; de 0,006 para 0,12 e de -0,02 para 0,21; e valores de acurácia de -0,70 para 0,32; de 0,04 para 0,32 e de -0,14 para 0,49, respectivamente. Uma das possíveis razões para isso, é que no método GBLUP tradicional considera-se igual proporção de variância explicada por cada marcador, já no método GBLUP supervisionado, o número de parâmetros é reduzido devido à seleção dos marcadores mais significativos.

Semelhante ao que ocorreu na população de estimação, as correlações entre os postos e consequentemente as acurácias foram maiores nos modelos L1 e S2 do que nos modelos L2 e S1. Uma das possíveis causas para essa correlação menor nos modelos L2 e S1 é porque nos modelos L1 e S2 não foi considerada a natureza da censura, sendo que os tempos censurados foram considerados como tempos exatos. Já nos modelos L2 e S1, a correlação foi menor devido à maior distância esperada entre os GBV's preditos e o fenótipo corrigido para os tempos censurados.

Tabela 8: Capacidade preditiva ($r_{y\hat{y}}$) (correlação entre os postos) e acurácia ($r_{g\hat{g}}$) da GWS na população de validação com base no modelo linear misto e de fragilidade de Cox.

Número de marcas	Modelos							
	L1		*L2*		*S1*		*S2*	
	$r_{y\hat{y}}$	$r_{g\hat{g}}$	$r_{y\hat{y}}$	$r_{g\hat{g}}$	$r_{y\hat{y}}$	$r_{g\hat{g}}$	$r_{y\hat{y}}$	$r_{g\hat{g}}$
2,5% (6 marcas)	0,27	0,75	0,18	0,57	0,11	0,42	0,22	0,47
5% (12 marcas)	0,27	**0,78**	**0,23**	**0,58**	0,13	0,43	**0,30**	0,69
10% (24 marcas)	0,27	0,70	0,20	0,52	0,14	**0,49**	0,29	**0,75**
20% (48 marcas)	**0,29**	0,70	0,21	0,47	**0,16**	0,46	0,26	0,67
30% (72 marcas)	0,26	0,60	0,19	0,37	0,13	0,34	0,24	0,60
40% (96 marcas)	0,25	0,59	0,18	0,36	0,13	0,33	0,22	0,52
50% (120 marcas)	0,22	0,55	0,16	0,32	0,12	0,32	0,21	0,49
75% (180 marcas)	0,17	0,47	0,11	0,29	-0,11	-0,37	-0,12	-0,40
100% (238 marcas)	0,08	0,28	-0,07	-0,70	0,01	0,04	-0,02	-0,14

L1: Modelo linear misto com dados completos; L2: Modelo linear misto com dados imputados; S1: Modelo misto de Cox com censura; S2: Modelo misto de Cox sem censura.

Ainda na Tabela 8, observa-se que as estimativas das correlações entre os postos na população de validação foram menores do que na população de estimação, o que também ocorreu quando se utilizou todos os marcadores (Tabelas 5 e 6). Observa-se também que para os modelos L1 e S1, os maiores valores de correlação

foram de 0,29 e de 0,16, respectivamente, ambos com uso de 48 marcas. Já para esses mesmos modelos, os maiores valores de acurácia foram obtidos com o uso de 12 e 24 marcas.

Para os modelos L2 e S2, o uso de somente 12 marcas com maiores efeitos significativos já permitiu maior correlação (0,23 e 0,30 respectivamente), sendo que os maiores valores de acurácia foram obtidos também com o uso de 12 marcas para o modelo L2 e de 24 marcas para o modelo S2. Ou seja, os valores da capacidade preditiva na população de validação indicaram menores números de marcadores com maior efeito do que aqueles indicados na população de estimação. Pode-se destacar também que assim como nos resultados encontrados por Resende et al. (2010) e Fernando et al. (2007) para o método RR-BLUP, o aumento do número de marcadores não aumenta linearmente a acurácia da GWS. A análise com 120 marcadores para os quatro modelos utilizados já propicia aumento na capacidade preditiva. Essa mesma metodologia foi aplicada por Resende Jr. et al. (2012) em pinheiros (*Pinus taeda L.*) utilizando o método RR-BLUP, a qual os autores denominaram RR-BLUP-B. O método apresentou desempenho superior ao RR-BLUP e semelhante ao métodos bayesianos para as características de resistência a doença e densidade da madeira.

As estimativas de herdabilidade estimada pelo método GBLUP considerando o modelo linear misto e de fragilidade de Cox e diferentes números de marcas estão apresentadas na Tabela 9. Essas estimativas foram obtidas por meio dos maiores efeitos de marcadores estimados na população de estimação. Observa-se de maneira geral um aumento significativo nas estimativas de herdabilidade ao se utilizar um subconjunto de marcadores selecionados por seus maiores efeitos. Para o modelo de fragilidade de Cox com censura (S1) essa diferença já é notória quando se passa a utilizar 180 marcadores. Essas estimativas aumentam à medida que se diminui o número de marcas, até a quantidade de 96 marcadores, onde se obtém maior herdabilidade (0,16). No entanto, conforme Tabela 8, a herdabilidade que maximiza a capacidade preditiva na validação cruzada foi aquela estimada com base em 48 marcas ($r_{y\hat{y}}$ igual a 0,16).

Visando um balanço entre acréscimo de informação por meio do maior número de marcadores e diminuição no tamanho da amostra (N/n), conforme

Resende et al. (2010), considerou-se a análise com 120 marcadores de efeitos mais significativos, uma vez que, conforme a Tabela 8, esse número de marcadores já propicia aumento significativo na capacidade preditiva nos quatro modelos.

Tabela 9: Herdabilidade estimada pelo método GBLUP considerando o modelo linear misto e de fragilidade de Cox e diferentes números de marcas.

Número de marcas	Modelos			
	L1	*L2*	*S1**	*S2**
2,5% (6 marcas)	0,13	0,10	0,07	**0,22**
5% (12 marcas)	0,12	0,16	0,09	0,19
10% (24 marcas)	0,15	0,15	0,08	0,15
20% (48 marcas)	0,17	0,20	0,12	0,15
30% (72 marcas)	**0,19**	**0,26**	0,15	0,16
40% (96 marcas)	0,18	0,25	**0,16**	0,18
50% (120 marcas)	0,16	0,25	0,14	0,18
75% (180 marcas)	0,13	0,14	0,09	0,09
100% (238 marcas)	0,08	0,01	0,02	0,02

L1: Modelo linear misto com dados completos; L2: Modelo linear misto com dados imputados; S1: Modelo misto de Cox com censura; S2: Modelo misto de Cox sem censura. *Herdabilidade calculada pela expressão dada em (6).

A estimativa de herdabilidade obtida considerando o modelo linear misto com dados completos (L1) e com dados imputados (L2) além dos 120 marcadores de maiores efeitos foi de 0,16 e de 0,25 respectivamente. Observa-se também que as estimativas obtidas em todos os subconjuntos de marcas estão dentro dos limites de 0,00 a 0,36 apresentados por Catalan (1986) para a característica idade ao final do teste. Mendonça et al. (2012) relatam valor de herdabilidade em torno de 0,25 para essa mesma população. Torres Jr. et al. (1998), ao trabalharem com a característica idade, a certo peso, encontraram valores de herdabilidade variando entre 0,01 e 0,49 e Torres Filho et al. (2004) encontraram valores de herdabilidade para idade até atingir 100 kg em suínos da raça *Large White* de 0,13 a 0,20. Vale notar que as herdabilidades para o modelo L2 foram maiores do que para o modelo L1, sendo que para o modelo L2 foi utilizada a imputação pela média das observações, o que pode ter ocasionado esse aumento nas herdabilidades. Para o modelo de fragilidade de Cox com e sem censura, os valores de herdabilidades foram de 0,14 e 0,18, respectivamente, com base nos 120 maiores efeitos de marcadores.

No modelo de Cox com censura (S1), a maior herdabilidade encontrada foi no uso de 96 marcas enquanto que a validação cruzada indicou a quantidade de

marcadores significativos igual a 48 marcas (Tabela 8). Os valores de herdabilidade foram próximos quando se utilizou um intervalo de 48 a 120 marcas, com valores de 0,12 a 0,16, o que também ocorreu para o modelo L2, com valores de herdabilidade entre 0,20 e 0,26. Já no modelo de Cox sem censura, a maior herdabilidade foi indicada para as 6 marcas com maiores efeitos significativos, sendo que na validação cruzada, a maior capacidade preditiva foi obtida com o uso de 12 marcas.

Esses resultados diferem dos encontrados por Resende et al. (2010) em eucalipto usando o método RR-BLUP, em que constataram que a herdabilidade que maximiza a capacidade preditiva na validação cruzada foi a mesma estimada especificamente para cada subconjunto de marcadores.

No modelo misto de Cox com e sem censura, o cálculo da herdabilidade dado em (6) foi baseado na expressão apresentada por Anderson et al. (2007) e Schneider et al. (2005), e que foi inicialmente proposta por Yazdi et al. (2002). Nessa expressão, a variância do erro foi substituída por *1/(1-c)*, em que *c* é a proporção de dados censurados. Para o modelo misto de Cox com censura (S1), o valor de *c* foi igual a 0,56 e para o S2 a variância do erro foi igual a 1, em que o erro aleatório na escala latente tem distribuição normal padrão com função de ligação probito. Em modelos de riscos proporcionais, como o de Weibull e de Cox, o componente do erro aleatório é incorporado na função de risco basal $(h_0(t))$, não sendo possível obter a estimativa da herdabilidade diretamente.

No entanto, segundo Pankratz et al. (2005) e Giolo & Demétrio (2011), uma outra forma de obter informação sobre a herdabilidade é por meio de uma medida de risco, uma vez que os componentes de variância obtidos no modelo misto de Cox são modelados na escala logarítmica do risco. Ou seja, o valor obtido pela exponencial da raiz quadrada da variância genética nos fornece informação a respeito do risco relativo associado aos efeitos aleatórios.

As estimativas dos componentes de variância poligênicos foram iguais a 0,38 e 0,22 nos modelos de Cox com e sem censura, respectivamente. O risco do animal de obter o ganho de peso ao abate em um curto período de tempo é em torno de $\exp(\sqrt{0,38}) \approx 1,852$ para o modelo misto de Cox com censura e de $\exp(\sqrt{0,22}) \approx 1,598$ para o modelo misto de Cox sem censura. De acordo com Giolo

& Demétrio (2011) e Pankratz et al. (2005), podemos afirmar então, que existem animais com o risco de obter o peso ideal de abate 85% maior que o risco médio geral com base no modelo misto de Cox com censura, e de 60% com base no modelo misto de Cox sem censura.

Vale ressaltar que essas estimativas foram baseadas no parentesco genômico dos 120 marcadores de maiores efeitos em módulo, nos quatro modelos avaliados. De maneira geral, observou-se que, ao utilizar os marcadores mais informativos (120 marcas), a herdabilidade que era de 0,02 (Tabelas 4 e 9) passou a ser de 0,14 e de 0,18 para os modelos de Cox com (S1) e sem (S2) censura, respectivamente. Mészáros et al. (2010) avaliaram o período de dias entre idade ao primeiro parto e idade ao abate em suínos da raças *Large White* e *Landrace* por meio do modelo de sobrevivência de Weibull, e encontraram valores de herdabilidade de 0,08 e 0,05 para as respectivas raças com base no modelo animal. Para o cálculo da herdabilidade, os autores usaram a mesma expressão citada neste trabalho e proposta por Yazdi et al. (2002). Analisando a mesma variável nas mesmas raças, Serenius & Stalder (2004) encontraram valores de herdabilidades 0,16 e 0,17 para *Landrace* e entre 0,17 e 0,19 para *Large White.* Yazdi et al. (2000) relataram valores de herdabilidade para a variável longevidade em suínos *Landrace* variando de 0,11 a 0,27.

Segundo Mézsáros et al. (2010), nesses dois últimos trabalhos, as herdabilidades foram obtidas assumindo que o erro aleatório na escala latente tem distribuição valor extremo com função de ligação complemento log-log e variância $\pi^2/6$. Nesse contexto, a herdabilidade é definida na escala logarítmica e não na escala original conforme foi proposto por Yazdi et al. (2002); portanto, não sendo possível a comparação direta. Assim, Mézsáros et al. (2010) sugeriram substituir na expressão (6) os componentes de variância obtidos de acordo com o modelo utilizado a fim de tornar as estimativas de herdabilidade comparáveis. Ao fazer isso, os autores comentam que o novo valor de herdabilidade obtido com base em Yazdi et al. (2000) foi de 0,09 e no trabalho de Serenius & Stalder (2004), variou entre 0,18 e 0,21, sendo próximo aos já obtidos pelos autores (0,16 – 0,19). Não foram encontrados na literatura trabalhos avaliando mais precisamente a variável idade do nascimento até o abate do animal, sendo a variável citada acima a mais próxima para comparação,

além de que todos os valores de herdabilidade encontrados na literatura são baseados na matriz de parentesco associada ao *pedigree*. Vale ressaltar também que ao se comparar essas estimativas de herdabilidade, deve-se levar em conta o modelo de sobrevivência e as covariáveis utilizadas, além da expressão utilizada para o cálculo da herdabilidade.

A Tabela 10 apresenta as estimativas dos coeficientes de correlação envolvendo valores fenotípicos corrigidos e valores genéticos genômicos preditos ($\widehat{GBV}$), pela validação cruzada, com base nos 120 marcadores mais significativos, no método GBLUP, considerando o modelo linear misto e o modelo de Cox com fragilidade. Nota-se que, como já citado anteriormente, a capacidade preditiva do método em predizer de forma acurada foi maior ao se utilizar a seleção de marcadores mais significativos em módulo, conforme pode ser visto ao se comparar com os valores apresentados nas diagonais da Tabela 6.

A capacidade preditiva dos modelos L1 e S2 foram praticamente iguais, com valores de 0,22 e 0,21, respectivamente, indicando que ambos os modelos na situação de equivalência foram concordantes em predizer o verdadeiro valor fenotípico. Para os modelos L2 e S1, os valores da capacidade preditiva foram próximos entre si, com valores iguais a 0,16 e 0,12, respectivamente.

Semelhante ao uso de todas as marcas, as estimativas de correlações (Pearson e Spearman) entre os GBV's preditos pelo modelo linear (L1) e os GBV's preditos pelo modelo de fragilidade de Cox sem censura (S2) foram maiores que as obtidas pelo modelo de Cox com censura (S1).

Tabela 10: Estimativas dos coeficientes de correlação envolvendo valores fenotípicos corrigidos e valores genéticos genômicos preditos $\widehat{GBV}$, pela validação cruzada, com base nos 120 marcadores mais significativos, no método GBLUP, considerando o modelo linear misto e o modelo misto de Cox.

Modelos	L1	L2	S1	S2
L1	0,22	0,52	-0,57	-0,89
L2	0,55	0,16	-0,54	-0,57
S1	-0,60	-0,54	0,12	0,52
S2	-0,91	-0,59	0,54	0,21

L1: Modelo linear misto com dados completos; L2: Modelo linear misto com dados imputados; S1: Modelo misto de Cox com censura; S2: Modelo misto de Cox sem censura. Na diagonal estão apresentadas as correlações entre os postos do $\widehat{GBV}$ e do *y*; acima da diagonal estão apresentadas as correlações de Spearman e abaixo da diagonal as correlações de Pearson entre os $\widehat{GBV}$'s estimados nos modelos utilizados.

Já o modelo L2 apresentou maiores estimativas de correlações (Pearson e Spearman) entre os GBV's preditos com o modelo de fragilidade de Cox sem censura (S2) do que com o S1. Vale notar também que, de forma geral, as correlações de Spearman foram menores que as estimativas das correlações de Pearson. Resultados semelhantes foram encontrados por Hou et al. (2009) ao comparar cinco modelos (modelo linear convencional, modelo linear com limiar, modelo de riscos proporcionais de Weibull e o modelo de riscos proporcionais de Cox com uma função de risco basal "*piecewise*") para avaliação genética do intervalo entre o parto e a primeira inseminação e o período entre o parto e a concepção em vacas *Danish Holstein*, onde também foram obtidas correlações negativas entre os valores genéticos estimados pelos modelos lineares e de sobrevivência. E Zhang et al. (2010) em estudo da GWS com dados simulados, obtiveram correlações dos postos dos valores genéticos genômicos menores do que as estimativas de correlação de Pearson. Os autores também afirmam que do ponto de vista prático, a correlação dos postos é mais importante que a correlação de Pearson.

A Figura 1.A apresenta as curvas de sobrevivência estimadas com base na média dos 10% maiores e 10% menores valores genéticos genômicos preditos na validação. E a Figura 1.B apresenta a curva de sobrevivência individual para o animal com maior e menor valor genético genômico predito, ambas as figuras com base no modelo de fragilidade de Cox com censura (S1). Essas estimativas foram obtidas por meio da expressão (7) apresentada no item 2.6.

De acordo com Giolo et al. (2003) e Giolo & Demétrio (2011), o interesse está em selecionar indivíduos que apresentem curvas de sobrevivência que decrescem rapidamente, dado que a variável resposta é o tempo até atingir o peso de abate. Nesse sentido, observa-se que os animais com maiores valores genéticos genômicos preditos (média igual a 0,58) apresentam decréscimos mais acentuados ao longo do tempo do que os animais com menor $\widehat{GBV}$ (média igual a -0,62), o que também pode ser visto quando se avalia individualmente, conforme Figura 1.B, em que o melhor animal (indivíduo 588) apresentou $\widehat{GBV}$ igual a 0,82 e o pior animal (indivíduo 827) apresentou $\widehat{GBV}$ igual a -0,84.

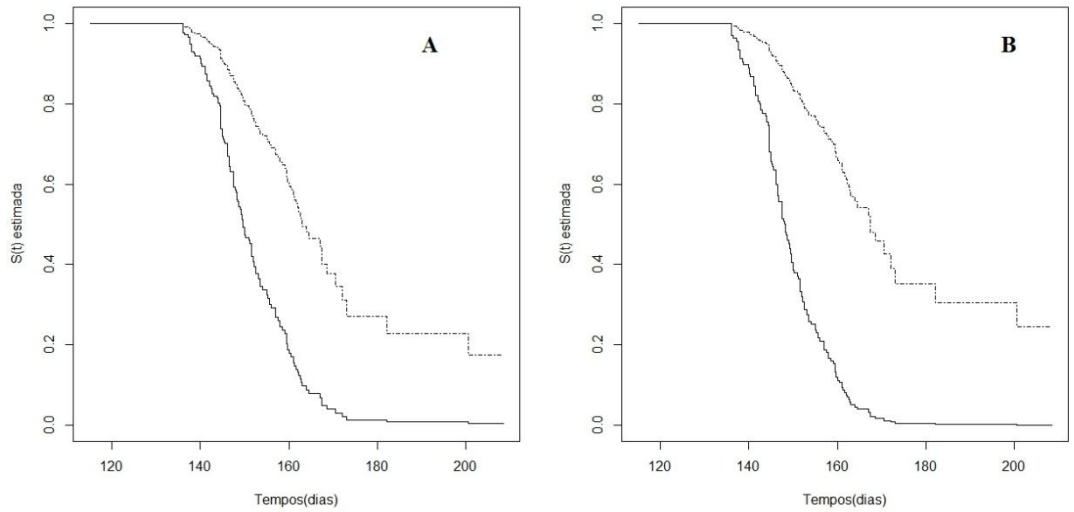

Figura 1: Curvas de sobrevivência estimadas com base na média dos 10% maiores (linha sólida) e os 10% menores (linha tracejada) valores genéticos genômicos preditos $\widehat{GBV}$ na validação (Figura 1.A). Curvas de sobrevivência estimada para o animal com o maior (linha sólida) e o menor (linha tracejada) valor genético genômico predito $\widehat{GBV}$ com base no modelo de fragilidade de Cox com censura (S1) e 120 marcas (Figura 1.B).

A Tabela 11 apresenta a proporção de concordância e o índice *Kappa* entre os 10% menores valores genéticos genômicos preditos ($\widehat{GBV}$) no modelo linear misto (L1) e os 10% maiores valores genéticos genômicos preditos ($\widehat{GBV}$) no modelo de fragilidade de Cox (S2), sem e com o método jacknife de validação cruzada, e a proporção de concordância dos efeitos de marcas em diferentes subconjuntos, sendo que, para todas as marcas e as 120 de maior efeito, foi obtida também a proporção de concordância entre os 10% maiores efeitos de marcas em valor absoluto.

O índice *Kappa* foi utilizado visando comparar os dois modelos (linear misto e de fragilidade de Cox) nas situações equivalentes, sendo sua significância verificada pelo teste Z por meio da expressão (8) apresentada no item 2.7. Por exemplo, considerando a quantidade de 120 marcadores de maiores efeitos, tem-se que, aproximadamente 74% dos 34 melhores indivíduos (25 animais) foram identificados em ambos os modelos linear (L1) e de fragilidade de Cox sem censura (S2), conforme pode ser visto na Tabela A do apêndice, e o índice *Kappa* foi igual a 0,71, sendo classificado conforme Landis & Koch (1977) como uma concordância muito boa.

Tabela 11: Proporção de concordância e índice *Kappa* entre os 10% maiores valores genéticos genômicos preditos ($\widehat{GBV}$) e entre os efeitos de marcas na população de estimação e os 10% maiores efeitos de marcadores na população de validação considerando os modelos L1 e S2.

Número de marcas	$\widehat{GBV}$		Ef. Marcas	
	Pop. Estimação	Pop. Validação	Pop. Estimação	Pop. Validação (10%)
2,5% (6 marcas)	0,74 (0,71**)	0,62 (0,57**)	0,67 (0,66**)	-
5% (12 marcas)	0,62 (0,57**)	0,56 (0,51**)	0,67 (0,65**)	-
10% (24 marcas)	0,76 (0,74**)	0,68 (0,64**)	0,71 (0,68**)	-
20% (48 marcas)	0,68 (0,64**)	0,68 (0,64**)	0,71 (0,63**)	-
30% (72 marcas)	0,68 (0,64**)	0,71 (0,67**)	0,78 (0,68**)	-
40% (96 marcas)	0,65 (0,61**)	0,74 (0,71**)	0,82 (0,70**)	-
50% (120 marcas)	**0,74** (**0,71****)	**0,71** (**0,67****)	**0,84** (**0,68****)	**0,67** (**0,63****)
100% (238 marcas)	0,76 (0,74**)	0,65 (0,61**)	-	0,71 (0,68**)

L1: Modelo linear misto com dados completos; S2: Modelo misto de Cox sem censura.
*: significativo a 5%; **: significativo a 1%; ^{ns}: não significativo.

Ao verificar a hipótese de que o índice *Kappa* é nulo, a mesma foi rejeitada ao nível de significância de 1% para todos os casos avaliados, indicando assim que há concordância direta e significativa entre os $\widehat{GBV}$'s preditos em ambos os modelos. No L1, os melhores animais são aqueles que apresentaram menores $\widehat{GBV}$'s, uma vez que alcançaram o peso desejável de abate em menor tempo. Já no modelo S2, os melhores animais são aqueles que apresentaram maiores $\widehat{GBV}$'s, dado que o efeito aleatório é a fragilidade; assim, a função de risco cresce rapidamente, indicando que o peso do animal também cresce rapidamente.

Observa-se, de maneira geral, que a proporção de concordância para os valores genéticos genômicos foi menor na população de validação do que na população de estimação. Fato este já evidenciado com base nas correlações dos postos dos GBV's preditos em ambos os modelos.

Para os efeitos de marcadores, observa-se que, das 120 marcas de maior efeito em módulo na população de estimação, 101 foram concordantes para ambos os modelos, com índice *kappa* igual a 0,68, sendo classificado como uma concordância muito boa e significativa ao nível de 1%. Ainda a respeito da seleção de marcas, observa-se que a proporção de 10% dos marcadores de maior efeito em ambos os modelos foi de 67% para as 120 marcas e de 71% para todos os 238 marcadores, com valores *kappa* iguais a 0,63 e 0,68 respectivamente. Ou seja, das 12 marcas de maior

efeito em módulo estimadas pelo L1 e S2 na validação cruzada, oito foram concordantes (Tabela G do apêndice). Vale observar novamente que em todos os casos a hipótese de concordância nula foi rejeitada ao nível de 1% de significância, evidenciando assim que as concordâncias não foram obtidas ao acaso, com valores *kappa* de classificação boa e muito boa.

Isso nos permite afirmar que o método proposto utilizando a matriz de parentesco genômica com base no modelo de Cox com efeito aleatório normal apresentou desempenho satisfatório, uma vez que selecionou cerca de 71% dos mesmos indivíduos e cerca de 84% (população de estimação) e de 67% (população de validação) dos mesmos marcadores quando se utilizou o modelo linear misto nas situações de dados normais, 0% de censura e seleção de 120 marcadores mais significativos.

A Tabela 12 apresenta a proporção de concordância e o índice *Kappa* entre os 10% menores valores genéticos genômicos preditos ($\widehat{GBV}$) no modelo linear misto (L2) e os 10% maiores valores genéticos genômicos preditos ($\widehat{GBV}$) no modelo de fragilidade de Cox (S1), sem e com o método jacknife de validação cruzada, e a proporção de concordância dos efeitos de marcador em diferentes subconjuntos, sendo que, para todas as marcas e as 120, foi obtida também a proporção de concordância entre os 10% maiores efeitos de marcas em valor absoluto.

Para a quantidade de 120 marcadores de maiores efeitos, tem-se que, somente 32% dos 34 melhores indivíduos (11 animais) foram identificados em ambos os modelos linear (L2) e de fragilidade de Cox com censura (S1), conforme pode ser visto na Tabela B do apêndice, e o índice *Kappa* foi igual a 0,28, sendo considerado razoável conforme classificação proposta por Landis & Koch (1977) e apresentada na Tabela 2 do item 2.7.

Ao verificar a hipótese de que o índice *Kappa* é nulo, a mesma foi rejeitada ao nível de significância de 1%, indicando assim que há concordância direta e significativa entre os $\widehat{GBV}$'s preditos em ambos os modelos. Da mesma forma que no modelo L1, no modelo L2, os melhores animais são aqueles que apresentaram menores $\widehat{GBV}$'s, uma vez que alcançaram o peso desejável de abate em menor tempo. Já no modelo S1, os melhores animais são aqueles que apresentaram maiores $\widehat{GBV}$'s,

dado que o efeito aleatório é a fragilidade; assim, a função de risco cresce rapidamente, indicando que o peso do animal também cresce rapidamente.

Tabela 12: Proporção de concordância e índice *Kappa* entre os 10% maiores valores genéticos genômicos preditos ($\widehat{GBV}$) e entre os efeitos de marcas na população de estimação e os 10% maiores efeitos de marcadores na população de validação considerando os modelos L2 e S1.

Número de marcas	$\widehat{GBV}$		Ef. Marcas	
	Pop. Estimação	Pop. Validação	Pop. Estimação	Pop. Validação (10%)
2,5% (6 marcas)	0,35 (0,28**)	0,44 (0,38**)	0,33 (0,32**)	-
5% (12 marcas)	0,24 (0,15*)	0,18 (0,08^{ns})	0,17 (0,12 ^{ns})	-
10% (24 marcas)	0,29 (0,21**)	0,26 (0,18*)	0,33 (0,26**)	-
20% (48 marcas)	0,41 (0,35**)	0,32 (0,25**)	0,35 (0,19**)	-
30% (72 marcas)	0,29 (0,21**)	0,15 (0,05 ^{ns})	0,44 (0,20**)	-
40% (96 marcas)	0,41 (0,35**)	0,32 (0,25**)	0,49 (0,15*)	-
50% (120 marcas)	**0,35** (**0,28****)	**0,32** (**0,25****)	**0,60** (**0,21****)	**0,25** (**0,17***)
100% (238 marcas)	0,35 (0,28**)	0,41 (0,35**)	-	0,33 (0,26**)

L2: Modelo linear misto com dados imputados; S1: Modelo misto de Cox com censura.
*: significativo a 5%; **: significativo a 1%; ^{ns}: não significativo.

Para os efeitos de marcadores, observa-se que, das 120 marcas de maior efeito em módulo na população de estimação, 72 foram concordantes para ambos os modelos, com índice *kappa* igual a 0,21, sendo classificado como uma concordância razoável, porém significativa ao nível de 1%.

A respeito da seleção de marcas na população de validação, observa-se que a proporção de 10% dos marcadores de maior efeito em ambos os modelos foi de apenas 25% para as 120 marcas e de 33% para todos os 238 marcadores, com valores *kappa* iguais a 0,17 e 0,26 respectivamente. Ou seja, das 12 marcas de maior efeito em módulo estimadas pelo L2 e S1 na validação cruzada, somente três foram concordantes e das 24, apenas oito.

Vale observar também que, apesar de na grande maioria dos casos a hipótese de concordância nula ter sido rejeitada aos níveis de 5% e 1% de significância, indicando que as concordâncias entre os modelos não foram obtidas ao acaso, os valores *kappa* foram classificados com desempenho ruim ou razoável.

Isso indica que, assumindo como verdadeiro o modelo S1, tem-se que o modelo linear com imputação pela média para os valores censurados no modelo S1

não foi eficiente em selecionar os melhores indivíduos e as marcas de maior efeito, uma vez que selecionou apenas 32% dos mesmos indivíduos (*kappa* = 0,25) e cerca de 60% (população de estimação, *kappa* = 0,21) e de 25% (população de validação, *kappa* = 0,17) dos mesmos marcadores quando comparado ao modelo de fragilidade de Cox na presença de censura e seleção de 120 marcadores mais significativos.

Os gráficos *Manhattan plot* são apresentados nas Figuras 2, 3 e 4. Optou-se por não apresentar a análise gráfica dos efeitos de marcas estimadas com base no modelo linear com dados imputados (L2), uma vez que a concordância com o modelo de fragilidade de Cox com censura (S1) praticamente não ocorreu.

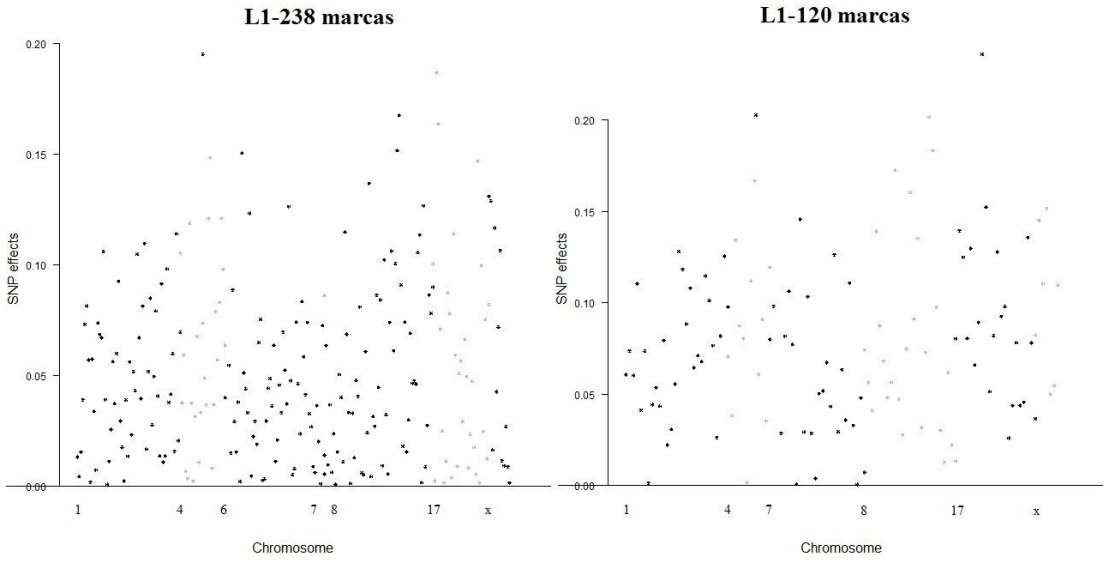

Figura 2: *Manhattan plot* dos efeitos de marcadores padronizados considerando o modelo linear misto com todas as marcas e com as 120 marcas com maiores efeitos.

Além dos gráficos, são apresentadas as Tabelas (apêndice) com base nos 120 marcadores de maior efeito e suas respectivas posições no cromossomo para os quatro modelos avaliados, permitindo assim verificar a existência de QTLs que afetam o caráter quantitativo (Azevedo, 2012). Os gráficos foram construídos com base nos efeitos estimados e padronizados na população de estimação.

Observa-se na Figura 2 que para ambas as análises com base no modelo L1, com todas as marcas e somente com as 120 de maiores efeitos em valor absoluto, foram encontrados SNPs de maior efeito. Nota-se efeito de marcador expressivo no cromossomo 4, mais precisamente na posição 80,1968 cM (Tabela C do apêndice),

para ambas as análises (todas as marcas e as 120 de maior efeito). Esse resultado está de acordo com o obtido por Marklund et al. (1999), onde foram encontrados um ou mais QTLs significativos nesse mesmo cromossomo em uma população F_2 obtida do cruzamento das raças *Large White* e *Landrace*. Ao considerar somente as 120 marcas de maiores efeitos, foi detectado maior efeito de marcador também no cromossomo 8, na posição 60,0367 cM, conforme Tabela C do apêndice.

Houston et al. (2006) avaliando uma população F2 (*Meishan x Large White*) encontraram QTL nessa mesma região cromossômica para a variável ganho de peso médio diário e Koning et al. (2003) também detectaram efeito de QTL na posição cromossômica entre 27 e 60,4 cM para a característica ganho de peso médio diário entre 25 e 90 kg. O efeito expressivo do marcador ALGA0048133 na posição 35,0374 cM do cromossomo 8 está de acordo ao resultado obtido por Quintanilla et al. (2002), os quais detectaram QTL no alcance -1.3 a 38.3 cM para a variável ganho de peso diário no intervalo de 70 a 154 dias, com base no cruzamento das raças *Large White* e *Meishan*.

A Figura 3 mostra o *Manhattan plot* dos efeitos de marcadores padronizados considerando o modelo de fragilidade de Cox com censura (S1) e com todas as marcas e as 120 marcas com maiores efeitos. Observa-se que, assim como no modelo linear, o marcador ALGA0026242 foi também de efeito expressivo no modelo de fragilidade de Cox com censura (Tabela E do apêndice). Além desse, foram encontrados marcadores de maiores efeitos no cromossomo 17, tais como ALGA0096093, ALGA0095323 e ALGA0094911, nas posições 50,2902 cM, 40,1282 cM e 35,0202 cM, respectivamente.

Esses resultados estão de acordo aos obtidos por Ramos et al. (2009) os quais encontraram QTLs significativos para a variável idade ao abate nesse mesmo cromossomo, porém em posições diferentes (95,2 cM, 95,4 cM e 11,4 cM). No entanto, Houston et al. (2005) ao avaliar animais de uma população F2 obtida do cruzamento das raças *Meshan* x *Large White* encontraram QTL significativo nesse cromossomo na posição 39,4 cM para a característica ganho de peso entre 35 e 80 kg.

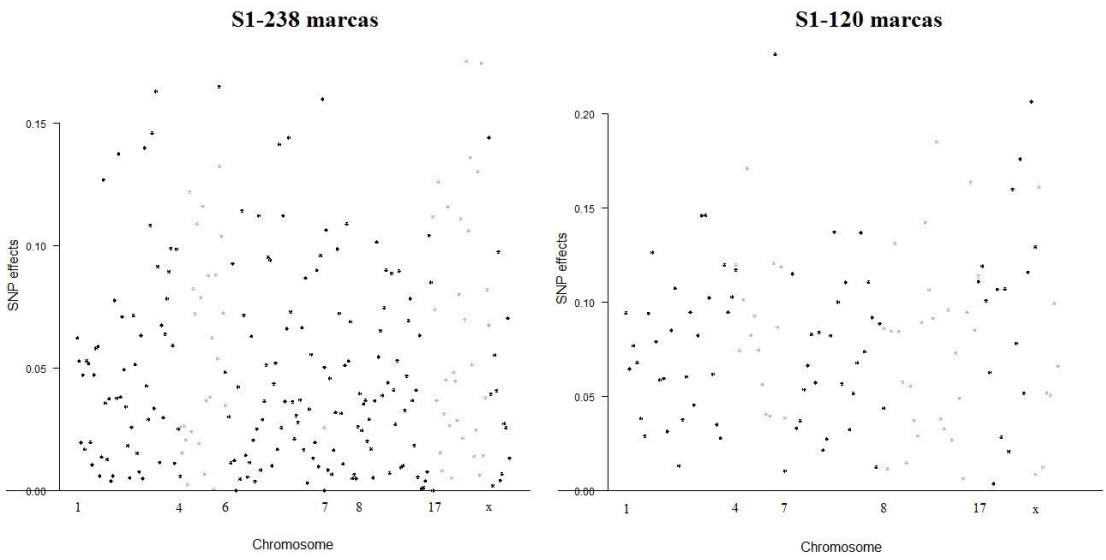

Figura 3: *Manhattan plot* dos efeitos de marcadores padronizados considerando o modelo de fragilidade de Cox com censura e com todas as marcas e as 120 marcas com maiores efeitos.

A Figura 4 mostra o *Manhattan plot* dos efeitos de marcadores padronizados considerando o modelo de fragilidade de Cox sem censura (S2), com todas as marcas e as 120 marcas com maiores efeitos. Observa-se novamente que os modelos de Cox com e sem censura e o modelo linear foram concordantes em estimar o maior efeito de marcador ALGA0026242 (Tabela F do apêndice). Isso possivelmente indica que a posição de tal marcador contém informações genéticas relacionadas com o ganho de peso em um curto período de tempo, sendo assim, deve ser melhor explorada em relação a possíveis genes que atuam sobre este processo. As Tabelas C, D, E e F do apêndice apresentam os marcadores com maiores efeitos e suas respectivas posições nos cromossomos com base na seleção das 120 marcas de maior efeito nos quatro modelos avaliados.

Observa-se que assim como nas figuras apresentadas acima, o marcador ALGA0026242 foi concordante nas três análises como o marcador de maior efeito utilizando o modelo misto de Cox nas situações com e sem censura e o de segundo maior efeito utilizando o modelo linear misto. Vale notar também efeitos expressivos de marcadores no cromossomo 7 nas posições 100,6055 cM, 120,8850 cM e 133,2550 cM, conforme Tabela F do anexo, os quais concordam com o trabalho de Edwards et al. (2008) que encontraram efeito de QTL na posição 113,9 cM para a variável ganho de peso no intervalo de 70 a 154 dias.

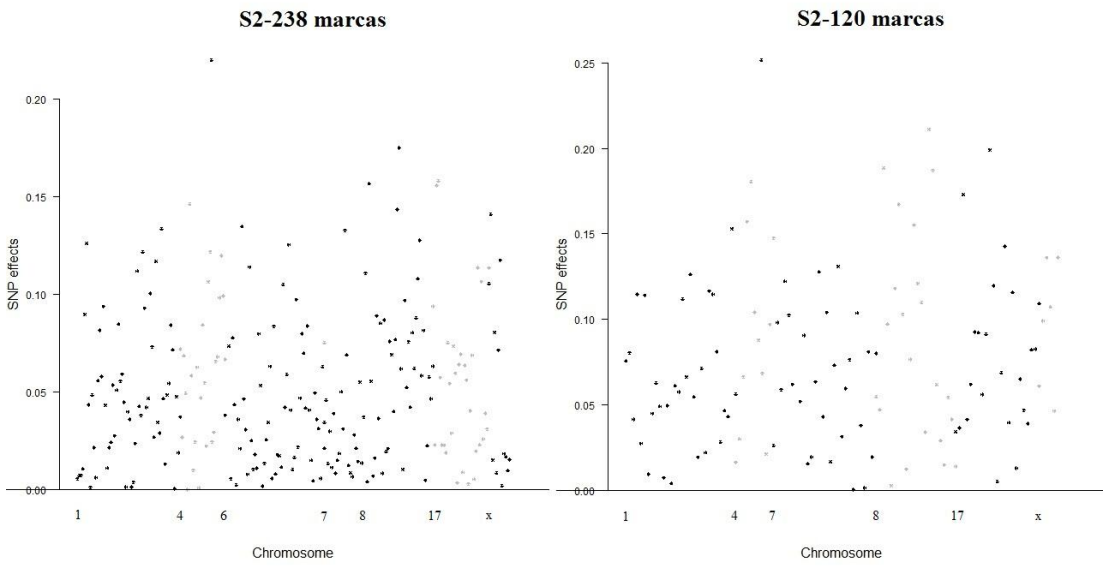

Figura 4: *Manhattan plot* dos efeitos de marcadores padronizados considerando o modelo de fragilidade de Cox sem censura e com todas as marcas e as 120 marcas com maiores efeitos.

Observa-se de maneira geral que foram encontrados marcadores com efeito significativo principalmente nos cromossomos 4, 7, 8 e 17.

4. CONCLUSÕES

Na situação de dados não censurados e normalidade, a análise por meio do modelo de Cox com efeito aleatório com distribuição normal e do modelo linear misto apresentou resultados concordantes na predição dos valores genéticos genômicos e na estimação dos efeitos de marcadores.

No entanto, ao considerar a censura, esses resultados foram discordantes, o que indica o modelo de Cox com efeito aleatório normal como o mais apropriado nessas situações. Uma recomendação futura é comparar o modelo de fragilidade de Cox com o modelo *Tobit* na presença de efeitos aleatórios e dados censurados.

O marcador ALGA0026242 se destacou com efeito expressivo nos três modelos, indicando uma possível presença de QTL no cromossomo 4.

A seleção de marcas permitiu um aumento nas correlações entre os postos dos valores genéticos genômicos preditos pelos modelos linear e de Cox com efeito aleatório normalmente distribuído com os valores fenotípicos corrigidos, sendo que para a característica analisada, 120 marcadores foram suficientes para maximizar a capacidade preditiva.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, C. A.; DUFFY, D. L.; MARTIN, N. G.; VISSCHER, P. M. Estimation of variance components for age menarche in twin families. **Behavior genetics**, v. 37, n. 5, p. 668 – 677, 2007.

AZEVEDO, C. F. **Métodos de redução de dimensionalidade aplicados na seleção genômica para características de carcaça em suínos**. 2012, 59 f. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2012.

BAND, G. D. O.; GUIMARÃES, S. E. F.; LOPES, P. S.; PEIXOTO, J. D. O.; FARIA, D. A.; PIRES, A. V.; FIGUEIREDO, F. C.; NASCIMENTO, C. S.; GOMIDE, L. A. M. Relationship between the Porcine Stress Syndrome gene and carcass and performance traits in F2 pigs resulting from divergent crosses. **Genetics and Molecular Biology**, v. 28, p. 92 – 96, 2005.

BRANDT, H.; VON BREVERN, N.; GLODEK, P. Factors affecting survival rate of crossbred sows in weaner production. **Livestock Production Science**, v. 57, p. 127 – 135, 1999.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American Statistical Associat**ion, v. 88, p. 9 – 25, 1993.

BUENGER, A.; DUCROCQ, V.; SWALVE, H. H. Analysis of survival in dairy cows with supplementary data on type scores and housing systems from a region of northwest germany, **Journal Dairy Science**, v. 84, p. 1531 – 1541, 2001.

CATALAN, G. **Estimativa de parâmetros genéticos e fenotípicos em suínos Landrace, Large White e Duroc, nas fases de crescimento e terminação**. Dissertação (Mestrado em Zootecnia) - Universidade Federal de Viçosa, 129p., 1986.

COHEN, J. A coeficient of agreement for nominal scales. **Educ. Psychol. Meas.**, v. 20, p. 37 - 46, 1960.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência**. São Paulo: Edgard Blücher, 2006. 369 p.

CONGALTON, R. G.; GREEN, K. **Assessing the accuracy of remotely sensed data:** principles and practices. New York: Lewis Publishers, 1998. 137p.

DUCROCQ, V. **An analysis of productive life in dairy cattle**. Ph.D. Dissertation, Cornell University, Ithaca, New York, 1987.

EDWARDS, D. B.; ERNST, C. W.; TEMPELMAN, R. J.; ROSA, G. J. M.; RANEY, N. E.; HOGE, M. D.; BATES, R. O. Quantitative trait loci mapping in an F2 Duroc x Pietrain resource population: I. Growth traits. **Journal Animal Science**, v. 86, n. 2, p. 241 – 253, 2008.

ENDELMAN, J.B. Ridge regression and other kernels for genomic selection with R package rrBLUP. **Plant Genome**, v. 4, p. 250 – 255, 2011.

FERNANDO, R. L.; HABIER, D.; STRICKER, C.; DEKKERS, J. C. M.; TOTTIR, L. R. Genomic selection. **Acta Agriculturae Scandinavica**, Section A – Animal Science, v. 57, n. 4, p. 192 - 195, 2007.

FLEISS, J. L.; LEVIN, B.; PAIK, M. C. The measurement of interrater agreement. **Statistical methods for rates and proportions**, v. 2, p. 212 – 236, 1981.

FONSECA, R.; SILVA, P.; SILVA, R. Acordo inter-juízes: O caso do coeficiente kappa. **Laboratório de Psicologia**, v. 5, n. 1, p. 81 – 90, 2007.

GALPARSORO, L. U.; FERNÁNDEZ, S. P. Medidas de concordancia: el índice de Kappa. **CAD Aten Primaria**, n. 6, p. 169 – 171, 2001.

GIOLO, S. R.; DEMÉTRIO, C. G. B. A frailty modeling approach for parental effects in animal breeding. **Journal of Applied Statistics**, v. 38, n. 3, p. 619 – 629, 2011.

GIOLO, S. R.; HENDERSON, R.; DEMÉTRIO, C. G. B. Um critério para a seleção de touros nelore usando modelos de sobrevivência. **Revista Brasileira de Biometria**, v. 21, n. 3, p. 115 – 223, 2003.

GODDARD, M. E. Genomic selection: prediction of accuracy and maximization of long term response. **Genetica**, v. 136, n. 2, p. 245 – 257, 2009.

GODDARD, M. E.; HAYES, B. J.; MEUWISSEN, T. H. E. Using the genomic relationship matrix to predict the accuracy of genomic selection. **Journal of Animal Breeding and Genetics**, v. 128, n. 6, p. 409 – 421, 2011.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. The impact of genetic relationship information on genome-assisted breeding values. **Genetics**, v. 177, n. 4, p. 2389 – 2397, 2007.

HAYES, B. J.; VISSCHER, P. M.; GODDARD, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. **Genetics Research**, v. 91, p. 47 – 60, 2009.

HOU, Y.; MADSEN, P.; LABOURIAU, R.; ZHANG, Y.; LUND, M. S.; SU, G. Genetic analysis of days from calving to first insemination and days open in Danish Holsteins using different models and censoring scenarios. **Journal Dairy Science**, v. 92, n. 3, p. 1229 – 1239, 2009.

HOUSTON, R. D.; HALEY, C. S.; ARCHIBALD, A. L.; RANCE, K. A. A QTL affecting daily feed intake maps to Chromosome 2 in pigs. **Mammalian genome**, v. 16, n. 6, p. 464 – 470, 2005.

HOUSTON, R. D.; HALEY, C. S.; ARCHIBALD, A. L.; CAMERON, N. D.; PLASTOW, G. S.; RANCE, K. A. A polymorphism in the 5`unstranslated region of the porcine cholecystokinin type a receptor gene affects feed intake and growth. **Genetics**, v. 174, n. 3, p. 1555 – 1563, 2006.

KONING, D. J.; PONG-WONG, R.; VARONA, L.; EVANS, G. J.; GIUFFRA, E.; SANCHEZ, A.; PLASTOW, G.; NOGUERA, J. L.; ANDERSSON, L.; HALEY, C. S. Full pedigree quantitative trait locus analysis in commercial pigs using variance components. **Journal Animal Science**, v. 81, n. 9, p. 2155 – 2163, 2003.

LANDIS, J.; KOCH, G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p.159 – 74, 1977.

MARKLUND, L.; NYSTROM, P. E.; STERN, S.; ANDERSSON-EKLUND, L.; ANDERSSON, L. Confirmed quantitative trait loci for fatness and growth on pig chromosome 4. **Heredity**, v. 82, n. 2, p. 134 – 141, 1999.

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

MÉSZÁROS, G.; PÁLOS, J.; DUCROCQ, V.; SOLKNER, J. Heritability of longevity in Large White and Landrace sows using continuous time and grouped data models. **Genetics Selection Evolution**, v. 42, p. 1 – 13, 2010.

NEJATI-JAVAREMI, A.; SMITH, C.; GIBSON, J. P. Effect of total allelic relationship on accuracy of evaluation and response to selection. **Journal of Animal Science**, v. 75, n. 7, p. 1738 - 1745,1997.

OJANGO, J. M. K.; DUCROCQ, V.; POLLOTT, G. E. Survival analysis of factors affecting culling early in the productive life of Holstein–Friesian cattle in Kenya, **Livestock Production Science**, v. 92, n. 3, p. 317 – 322, 2005.

PANKRATZ, V. S.; ANDRADE, M.; THERNEAU, T. M. Random-effects Cox proportional hazards model: general variance components methods for time-to-event data. **Genetic Epidemiology**, v. 28, n. 2, p. 97 – 109, 2005.

PEIXOTO, J. O.; GUIMARAES, S. E. F.; LOPES, P. S.; SOARES, M. A. M.; PIRES, A. V.; SILVA, M. V.; TORRES, R. A.; SILVA, M. A. E. Associations of leptin gene polymorphisms with production traits in pigs. **Journal of Animal Breeding and Genetics**, v. 123, n. 6, p. 378 – 383, 2006.

QUENOUILLE, M. H. Notes on bias in estimation. **Biometrika**, v. 43, n. 3, p. 353 – 360, 1956.

QUINTANILLA, R.; MILAN, D.; BIDANEL, J. P. A further look at quantitative trait loci affecting growth and fatness in across between Meishan and Large White pig populations. **Genetics Selection Evolution**, v. 34, n. 2, p. 193 – 210, 2002.

RAMOS, A. M.; BASTIAANSEN, J. W. M.; PLASTOW, G. S.; ROTHSCHILD, M. F. Genes located on a SSC17 meat quality QTL region are associated with growth in outbred pig populations. **Animal genetics**, v. 40, n. 5, p. 774 – 778, 2009.

RESENDE, M. D. V. **Genética biométrica e estatística no melhoramento de plantas perenes**. Brasília: Embrapa Informação Tecnológica, 2002. 975p.

RESENDE, M. D. V. Seleção genômica ampla (GWS) e modelos lineares mistos. In: **Matemática e estatística na análise de experimentos e no melhoramento genético**. Colombo: Embrapa Florestas, 2007. p. 517 - 534.

RESENDE, M. D. V. **Genômica quantitativa e seleção no melhoramento de plantas perenes e animais**. Colombo: Embrapa Florestas, 2008. 330 p.

RESENDE, M. D. V.; RESENDE JUNIOR, M. F. R.; AGUIAR, A. M.; ABAD, J. I. M.; MISSIAGIA, A. A.; SANSALONI, C.; PETROLI, C.; GRATTAPALIA, D. **Computação da seleção genômica ampla (GWS)**. Colombo: Embrapa Florestas, 79p, 2010.

RESENDE, M. D. V.; SILVA, F. F.; LOPES, P. S.; AZEVEDO, C. F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial**. Viçosa: Universidade Federal de Viçosa/Departamento de Estatística. 2012. 291 p. <http://www.det.ufv.br/ppestbio/corpo_docente.php>. Acesso em: 07 de jan. 2013.

RESENDE JR., M.F.R.; MUÑOZ, P.; RESENDE, M.D.V.; GARRICK, D.J.; FERNANDO, R.L.; DAVIS, J.M.; JOKELA, E.J.; MARTIN, T.A.; PETER, G.F.; KIRST, M. Accuracy of Genomic Selection Methods in a Standard Dataset of Loblolly Pine (Pinus taeda L.). **Genetics**, v. 190, n. 1, p. 1503 - 1510, 2012.

RIPATTI S, PALMGREN J. Estimation of multivariate frailty models using penalized partial likelihood. **Biometrics**, v. 56 n. 4, p. 1016 – 1022, 2000.

ROXSTROM, A.; STRANDBERG, E.; BERGLUND, B.; EMANUELSON, U.; PHILIPSSON, J. Genetic and environmental correlations among female fertility traits and milk production in different parities of Swedish Red and White dairy cattle. **Acta Agriculture Scandinava**, v. 51, n. 1, p. 7 – 14, 2001.

ROCHA, G. S. **Métodos estatísticos na seleção genômica ampla para curvas de crescimento em animais**. 2011, 56 f. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2011.

SCHNEIDER, M. D. P.; STRANDBERG, E.; DUCROCQ, V.; ROTH, A. Survival analysis applied to genetic evaluation for female fertility in dairy cattle. **Journal Dairy Science**, v. 88, n. 6, p. 2253 – 2259, 2005.

SERENIUS, T.; STALDER, K. J. Genetics of length of productive life and lifetime prolificacy in the Finnish Landrace and Large White pig populations. **Journal Animal Science**, v. 82, n. 11, p. 3111 - 3117, 2004.

SU, G.; CHRISTENSEN, O. F.; OSTERSEN, T.; HENRYON, M.; LUND, M. S. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. **PLoS ONE**, v. 7, n. 9: e45293, 2012.

TARRES, J.; BIDANEL, J. P.; HOFER, A.; DUCROCQ, V. Analysis of longevity and exterior traits on Large White sows in Switzerland. **Journal Animal Science**, v. 84, n. 11, p. 2914 - 2924, 2006.

THERNEAU, T. M. **On mixed-effect Cox models, sparse matrices, and modeling data from large pedigrees**, Tech. Rep., Mayo Foundation, Rochester, MN, USA, 2007. Disponível em < http://mayoresearch.mayo.edu/biostat/upload/kinship.pdf. > Acesso em: 10 de out. 2012.

THERNEAU, T. M. Coxme: Mixed Effects Cox Models.. R package version 2.2-3, 2012. Disponível em <http://CRAN.R-project.org/package=coxme> Acesso em: 10 de fev. 2013.

TORRES JÚNIOR, R. A. A.; SILVA, M. A.; LOPES, P. S.; REGAZZI, A. J.; EUCLYDES, R. F. Estimativas de componentes de (co)variância para características produtivas de suínos Landrace e Large White pelo método da máxima verossimilhança restrita. **Revista Brasileira de Zootecnia**, v.27, n. 2, p. 283 – 291, 1998.

TORRES FILHO, R. A.; TORRES, R. A.; LOPES, P. S.; EUCLYDES, R. F.; ARAÚJO, C. V.; PEREIRA, C. S.; SILVA, M. A. Avaliação de Modelos para Estimação de Componentes de (co)Variância em Características de Desempenho e Reprodutivas em Suínos. **Revista Brasileira de Zootecnia**, v. 33, n. 2, p. 350 - 357, 2004.

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

VAN RADEN, P. M.; VAN TASSELL, C. P.; WIGGANS, G. R.; SONSTEGARD, T. S.; SCHNABEL, R. D.; TAYLOR, J. F.; SCHENKEL, F. S. Invited review: Reliability of genomic predictions for North American Holstein bulls. **Journal of Dairy Science**, v. 92, n. 1, p. 16 – 24, 2009.

YAZDI, M. H.; RYDHMER, L.; RINGMAR-CEDERBERG, E.; LUNDEHEIM, N.; JOHANSSON, K. Genetic study of longevity in Swedish Landrace sows. **Livestock Production Science**, v. 63, n. 3, p. 255 – 264, 2000.

YAZDI, M. H.; VISSHER, P. M.; DUCROCQ, V. THOMPSON, R. Heritability, reability of genetic evaluations and response to selection in proportional hazard models. **Journal Dairy Science**. v. 85, p. 1563 – 1577, 2002.

ZHANG, Z.; LIU, J.; DING, X.; BIJMA, P.; KONING, D. J.; ZHANG, Q. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. **PLoS ONE**, v. 5, n. 9: e12648, 2010.

ZHANG, Z., DING, X.; LIU, J.; ZHANG, Q.; KONING, D. J. Accuracy of genomic prediction using low-density marker panels. **Journal Dairy Science**, v. 94, n. 7, p. 3642 – 3650, 2011.

APÊNDICE

Tabela A: Concordância entre os 10% maiores valores genéticos genômicos preditos pelo modelo misto de Cox sem censura (S2) e os 10% menores preditos pelo modelo linear misto (L1) no método jacknife de validação cruzada.

Ordem[1]	Indivíduo	$\widehat{GBV}^2$	$\widehat{GBV}^3$	Ordem[1]	Indivíduo	$\widehat{GBV}^2$	$\widehat{GBV}^3$
1	499	0,6730	-4,6639	13	1013	0,4695	-6,3481
2	412	0,6634	-5,4424	14	1012	0,4675	-5,9650
3	940	0,6000	-5,4611	15	1125	0,4548	-3,6740
4	519	0,5780	-5,6131	16	939	0,4515	-3,7360
5	1008	0,5744	-7,1052	17	795	0,4446	-4,1623
6	792	0,5712	-6,2177	18	937	0,4386	-4,4337
7	490	0,5354	-6,8951	19	407	0,4335	-5,3971
8	482	0,5301	-5,9839	20	807	0,4331	-4,3976
9	550	0,5079	-6,3670	21	1014	0,4305	-4,5854
10	941	0,5073	-4,5635	22	660	0,4080	-6,4669
11	956	0,5038	-6,3833	23	780	0,3761	-3,9909
12	1023	0,4964	-5,6229	24	714	0,3555	-3,6570

Taxa de concordância: 71%

1: Em ordem decrescente com base no modelo misto de Cox; 2: $\widehat{GBV}$ estimado pelo modelo misto de Cox; 3: $\widehat{GBV}$ estimado pelo modelo linear misto.

Tabela B: Concordância entre os 10% maiores valores genéticos genômicos preditos pelo modelo misto de Cox com censura (S1) e os 10% menores preditos pelo modelo linear misto (L2) no método jacknife de validação cruzada.

Ordem[1]	Indivíduo	$\widehat{GBV}^2$	$\widehat{GBV}^3$	Ordem[1]	Indivíduo	$\widehat{GBV}^2$	$\widehat{GBV}^3$
1	713	0,8045	-2,3769	7	807	0,5389	-3,2356
2	440	0,7139	-3,2291	8	490	0,5235	-2,8701
3	519	0,6331	-2,8629	9	780	0,5108	-3,7016
4	660	0,6016	-2,3742	10	714	0,4653	-2,9062
5	794	0,5548	-2,5305	11	777	0,4577	-4,9803
6	792	0,5409	-2,3772				

Taxa de concordância: 32%

1: Em ordem decrescente com base no modelo misto de Cox; 2: $\widehat{GBV}$ estimado pelo modelo misto de Cox com censura; 3: $\widehat{GBV}$ estimado pelo modelo linear misto com dados imputados.

Tabela C: Marcadores com maiores efeitos obtidos por meio do modelo linear misto (L1) para a característica idade ao abate em suínos, com base em 120 marcadores.

Marcador	Efeito	Cromossomo	Posição (cM)
ALGA0049546	-0,8470	SSC8	60,0367
ALGA0026242	0,7271	SSC4	80,1968
ALGA0047440	0,7237	SSC8	15,0429
ALGA0047444	-0,6578	SSC8	15,1855
ALGA0045009	0,6194	SSC7	120,8850
ALGA0025374	-0,5981	SSC4	60,2156
ALGA0046005	-0,5761	SSC7	133,2550
ALGA0049550	0,5464	SSC8	60,0694
ASGA0080454	-0,5440	SSCX	0,0394
ALGA0027862	0,5228	SSC4	105,0238
ALGA0099785	0,5204	SSCX	35,1721
ALGA0048133	-0,5003	SSC8	35,0374

Tabela D: Marcadores com maiores efeitos obtidos por meio do modelo linear misto com observações imputadas (L2), para a característica idade ao abate em suínos, com base em 120 marcadores.

Marcador	Efeito	Cromossomo	Posição (cM)
ALGA0047440	0,2110	SSC8	15,0429
ALGA0047003	-0,1980	SSC8	10,1677
ALGA0099944	0,1958	SSCX	55,0226
ALGA0048131	-0,1793	SSC8	35,0249
ALGA0007897	-0,1722	SSC1	190,5937
ALGA0043769	-0,1672	SSC7	100,6632
ALGA0026242	0,1631	SSC4	80,1968
ALGA0048133	-0,1609	SSC8	35,0374
ALGA0048658	0,1583	SSC8	45,1115
MARC0051258	-0,1548	SSCX	112,2210
ALGA0024031	-0,1525	SSC4	20,2485
ALGA0025382	-0,1431	SSC4	60,3117

Tabela E: Marcadores com maiores efeitos obtidos por meio do o modelo de fragilidade de Cox com censura (S1) para a característica idade ao abate em suínos, com base em 120 marcadores.

Marcador	Efeito	Cromossomo	Posição (cM)
ALGA0026242	-0,0995	SSC4	80,1968
ALGA0096093	-0,0887	SSC17	50,2902
ALGA0047003	0,0796	SSC8	10,1677
ALGA0095323	-0,0756	SSC17	40,1282
ALGA0024446	-0,0735	SSC4	30,1924
ALGA0048843	0,0702	SSC8	50,0208
ALGA0099785	-0,0692	SSCX	35,1721
ALGA0094911	0,0687	SSC17	35,0202
ALGA0007897	0,0627	SSC1	190,5937
ALGA0007807	0,0627	SSC1	184,6198
ALGA0046005	0,0611	SSC7	133,2550
ALGA0037853	0,0589	SSC7	0,4704

Tabela F: Marcadores com maiores efeitos obtidos por meio do modelo de fragilidade de Cox sem censura (S2) para a característica idade ao abate em suínos, com base em 120 marcadores.

Marcador	Efeito	Cromossomo	Posição (cM)
ALGA0026242	-0,0937	SSC4	80,1968
ALGA0047440	-0,0787	SSC8	15,0429
ALGA0049546	0,0742	SSC8	60,0367
ALGA0043766	0,0702	SSC7	100,6055
ALGA0047444	0,0697	SSC8	15,1855
ALGA0024031	0,0673	SSC4	20,2485
ALGA0048133	0,0644	SSC8	35,0374
ALGA0045009	-0,0623	SSC7	120,8850
ALGA0023180	-0,0585	SSC4	10,0109
ALGA0046005	0,0578	SSC7	133,2550
ALGA0009321	0,0570	SSC1	225,1225
ALGA0025374	0,0549	SSC4	60,2156

Tabela G: Concordância entre os 10% maiores efeitos de marcas obtidas por meio do modelo linear misto (L1) e de fragilidade de Cox sem censura (S2) para a característica idade ao abate em suínos, com base em 120 marcadores na população de validação.

Marcador	Efeito (S2)	Efeito (L1)	Cromossomo	Posição (cM)
ALGA0026242	-0,0935	0,7262	SSC4	80,1968
ALGA0047440	-0,0784	0,7223	SSC8	15,0429
ALGA0049546	0,0740	-0,8455	SSC8	60,0367
ALGA0047444	0,0695	-0,6569	SSC8	15,1855
ALGA0048133	0,0642	-0,4994	SSC8	35,0374
ALGA0045009	-0,0621	0,6183	SSC7	120,8850
ALGA0046005	0,0576	-0,5748	SSC7	133,2550
ALGA0025374	0,0548	-0,5973	SSC4	60,2156
Taxa de Concordância: 67%				

ANEXO

Script das principais análises no *software* R

```
setwd("C:\\Users\\vinicius\\Dissertacao\\Analises")
library(rrBLUP)
library(coxme)
library(MASS)
```

#lendo fenótipos e efeitos fixos

```
dados=read.table("fenotipo_cov.txt", h=T)
```

#corrigindo fenótipo para efeitos fixos

```
ID=dados$ID
sexo=factor(dados$SEXO)
lote de manejo=factor(dados$lote)
y_adj=mean(dados$IDA)+ lm(dados$IDA~ sexo + lote)$residuals
```

#idade corrigida com e sem censura

```
adj_cen=cbind(ID,y_adj,cen)
adj_cen1=cbind(ID,y_adj,cen1)
```

#lendo arquivo marcadores SNPs

```
snp=read.table("snp_todos.txt",h=T)
```

#arquivo final: fenotipo_adj, censura e marcadores

```
snp_adj_cen=merge(adj_cen,snp,by=c("ID"))
snp_adj_cen1=merge(adj_cen1,snp,by=c("ID"))

write.table(snp_adj_cen,"data_final.txt",quote=FALSE,row.names
=FALSE)
write.table(snp_adj_cen1,"data_final_naocen.txt",quote=FALSE,r
ow.names=FALSE)
```

#lendo os dados finais

```
data=read.table("data_final.txt",h=T) #dados com censura
data1=read.table("data_final_naocen.txt",h=T) #dados sem censura
dados2= read.table("data_final_linear_cen.txt",h=T) #dados
imputados
```

#calculando matriz de parentesco genômica (G)

```
M=as.matrix(data[,-(1:3)])
dim(M)
M=M-1
A=A.mat(M)
n=nrow(M)
```

```
G=A+diag(n)*10^-6

# ajuste pelo modelo linear misto com dados completos – L1
(model1=lmekin(data$y_adj ~ (1|data$ID) ,
data=data,varlist=list(G)))
u_ml=random.effects(model1)$data.ID

#Herdabilidade
vg= 9.102893 #variância genética
ve=model1$sigma^2   #variância residual
h2_ml=vg/(vg+ve)    #herdabilidade
h2_ml

# ajuste pelo modelo linear misto com dados imputados – L2
(model2=lmekin(data2$y_adj ~ (1|data2$ID) ,
data=data2,varlist=list(G)))
u_ml1=random.effects(model2)$data2.ID

#Herdabilidade
vg= 0.595 #variância genética
ve=model2$sigma^2   #variância residual
h2_ml1=vg/(vg+ve)    # herdabilidade
h2_ml1

# ajuste pelo modelo misto de Cox com censura – S1
(model3=coxme(Surv(y_adj,cen) ~ (1|ID) , data=data,
varlist=list(G)))
u_cox=model3$frail$ID

#Herdabilidade calculada pela expressão 6.
vg=model3$vcoef$ID   #variância genética
c=0.561   #proporção de dados censurados
ve=1/(1-c)
h_s1=vg/(vg+ve)

# ajuste pelo modelo misto de Cox sem censura (S2)
(model4=coxme(Surv(y_adj,cen) ~ (1|ID) , data=data1,
varlist=list(G)))
u_cox1=model4$frail$ID

#Herdabilidade calculada pela expressão 6.
vg=model4$vcoef$ID   #variância genética
c=0   #proporção de dados censurados
ve=1/(1-c)
h_s2=vg/(vg+ve)


########## Validação com todas as marcas############################
#ajuste pelo modelo linear misto com dados completos – L1
gbv_ml=NULL
eff_ml=matrix(0,ncol(M),nrow(M))
system.time(
for(i in 1:nrow(M))
{
```

```
u_ml_vc=(lmekin(y_adj[-i] ~ (1|data$ID[-i]) , method="ML",
data=data,varlist=list(G[-i,-
i])))$coefficients$random$data.ID..i.
eff_ml[,i]=ginv(t(M[-i,])%*%M[-i,])%*%t(M[-i,])%*%u_ml_vc
gbv_ml[i]=M[i,]%*%eff_ml[,i]
}
)
write.table(gbv_ml,"gbv_gblup_ml.txt",row.names=FALSE,col.name
s=FALSE,quote=FALSE)
```

#Vetor de efeito de marcadores

```
mean_eff_ml=NULL
for(i in 1:nrow(eff_ml))
{
mean_eff_ml[i]=sum(eff_ml[i,])/ncol(eff_ml)
}
mean_eff_ml
mean_eff_ml=cbind(colnames(data[,-(1:3)]),mean_eff_ml)
colnames(mean_eff_ml)=c("marker","effect")
write.table(mean_eff_ml,"eff_gblup_ml.txt",row.names=FALSE,col
.names=TRUE,quote=FALSE)
```

#ajuste pelo modelo linear misto com dados imputados – L2

```
gbv_ml1=NULL
eff_ml1=matrix(0,ncol(M),nrow(M))
system.time(
for(i in 1:nrow(M))
{
u_ml1_vc=(lmekin(y_adj[-i] ~ (1|data2$ID[-i]) , method="ML",
data=data2,varlist=list(aux5[-i,-
i])))$coefficients$random$data2.ID..i.
eff_ml1[,i]=ginv(t(M[-i,])%*%M[-i,])%*%t(M[-i,])%*%u_ml1_vc
gbv_ml1[i]=M[i,]%*%eff_ml1[,i]
}
)
write.table(gbv_ml1,"gbv_gblup_ml1.txt",row.names=FALSE,col.na
mes=FALSE,quote=FALSE)
```

#Vetor de efeito de marcadores

```
mean_eff_ml1=NULL
for(i in 1:nrow(eff_ml1))
{
mean_eff_ml1[i]=sum(eff_ml1[i,])/ncol(eff_ml1)
}
mean_eff_ml1
mean_eff_ml1=cbind(colnames(data2[,-(1:3)]),mean_eff_ml1)
colnames(mean_eff_ml1)=c("marker","effect")
write.table(mean_eff_ml1,"eff_gblup_ml1.txt",row.names=FALSE,c
ol.names=TRUE,quote=FALSE)
```

#ajuste pelo modelo misto de Cox com censura – S1

```
gbv_cox_vc=NULL
eff_cox=matrix(0,ncol(M),nrow(M))
system.time(
```

```
for(i in 1:nrow(M))
{
u_cox_vc=(coxme(Surv(y_adj,cen)[-i] ~ (1|ID[-i]) , data=data,
varlist=list(G[-i,-i])))$frail$ID..i.
eff_cox[,i]=ginv(t(M[-i,])%*%M[-i,])%*%t(M[-i,])%*%u_cox_vc
gbv_cox_vc[i]=M[i,]%*%eff_cox[,i]
}
)

write.table(gbv_cox_vc,"gbv_cox_vc.txt",row.names=FALSE,col.na
mes=FALSE,quote=FALSE)
```

#Vetor de efeito de marcadores

```
gbv_cox_vc
mean_eff_cox=NULL
for(i in 1:nrow(eff_cox))
{
mean_eff_cox[i]=sum(eff_cox[i,])/ncol(eff_cox)
}
mean_eff_cox
mean_eff_cox=cbind(colnames(data[,-(1:3)]),mean_eff_cox)
colnames(mean_eff_cox)=c("marker","effect")
write.table(mean_eff_cox,"eff_cox_vc.txt",row.names=FALSE,col.
names=TRUE,quote=FALSE)
```

#ajuste pelo modelo misto de Cox sem censura – S2

```
gbv_cox_vc1=NULL
eff_cox1=matrix(0,ncol(M),nrow(M))
system.time(
for(i in 1:nrow(M))
{
u_cox_vc1=(coxme(Surv(y_adj,cen)[-i] ~ (1|ID[-i]) ,
data=data1, varlist=list(G[-i,-i])))$frail$ID..i.
eff_cox1[,i]=ginv(t(M[-i,])%*%M[-i,])%*%t(M[-i,])%*%u_cox_vc1
gbv_cox_vc1[i]=M[i,]%*%eff_cox1[,i]
}
)
write.table(gbv_cox_vc1,"gbv_cox_vc1.txt",row.names=FALSE,col.
names=FALSE,quote=FALSE)
```

#Vetor de efeito de marcadores

```
gbv_cox_vc1
mean_eff_cox1=NULL
for(i in 1:nrow(eff_cox1))
{
mean_eff_cox1[i]=sum(eff_cox1[i,])/ncol(eff_cox1)
}
mean_eff_cox1
mean_eff_cox1=cbind(colnames(data[,-(1:3)]),mean_eff_cox1)
colnames(mean_eff_cox1)=c("marcas","effect")
write.table(mean_eff_cox1,"eff_cox_vc1.txt",row.names=FALSE,co
l.names=TRUE,quote=FALSE)
```

#######GBLUP supervisionado - script para as 120 marcas de maior efeito####

modelo linear misto – L1

```
marcas=read.table("marcas.txt",h=T)  #arquivo com o nome de todas as
marcas
a_ml=abs(ginv(t(M)%*%M)%*%t(M)%*%u_ml)   #efeitos de todas as marcas
a_ml_fim=(cbind(marcas,a_ml))
top_a_ml=a_ml_fim[order(a_ml_fim[ ,2], decreasing = TRUE),
][1:120,]
colnames(top_a_ml)=c("marcas","ef.ml")

M_new120m=cbind(colnames(M),t(M))
colnames(M_new120m)=c("marcas",1:335)

snp_120m=merge(top_a_ml,M_new120m,
by=intersect("marcas","marcas"))
Z120m=t(snp_120m[,-(1:2)])
colnames(Z120m)=c(as.matrix(snp_120m[,1]))
write.table(Z120m,"Z120m.txt",row.names=FALSE,col.names=TRUE,q
uote=FALSE)

Z120m=read.table("Z120m.txt",h=T)
Z120m=as.matrix(Z120m)
dim(Z120m)
A_new=A.mat(Z120m)
n=335
G_120m=A_new+diag(n)*10^-6
head(G_120m)
head(Z120m)

# validação com as 120 marcas
gbv_ml_120=NULL
eff_ml_120=matrix(0,ncol(Z120m),nrow(Z120m))
system.time(
for(i in 1:nrow(Z120m))
{
u_ml_vc_120=(lmekin(y_adj[-i] ~ (1|data$ID[-i]) , method="ML",
data=data,varlist=list(G_120m[-i,-
i])))$coefficients$random$data.ID..i.
eff_ml_120[,i]=ginv(t(Z120m[-i,])%*%Z120m[-i,])%*%t(Z120m[-
i,])%*%u_ml_vc_120
gbv_ml_120[i]=Z120m[i,]%*%eff_ml_120[,i]
}
)
write.table(gbv_ml_120,"gbv_ml_120.txt",row.names=FALSE,col.na
mes=FALSE,quote=FALSE)
#Vetor de efeito de marcadores

mean_eff_ml_120=NULL
for(i in 1:nrow(eff_ml_120))
{
mean_eff_ml_120[i]=sum(eff_ml_120[i,])/ncol(eff_ml_120)
}
mean_eff_ml_120
mean_eff_ml_120=cbind(colnames(Z120m),mean_eff_ml_120)
colnames(mean_eff_ml_120)=c("marker","effect")
write.table(mean_eff_ml_120,"eff_gblup_ml_120.txt",row.names=F
ALSE,col.names=TRUE,quote=FALSE)
```

#modelo linear misto com dados imputados – L2

```
a_ml1=abs(ginv(t(M)%*%M)%*%t(M)%*%u_ml1)  #efeitos de todas as marcas
a_ml1_fim=(cbind(marcas,a_ml1))
top_a_ml1=a_ml1_fim[order(a_ml1_fim[ ,2], decreasing = TRUE),
][1:120,]
colnames(top_a_ml1)=c("marcas","ef.ml1")

M_new120m1=cbind(colnames(M),t(M))
colnames(M_new120m1)=c("marcas",1:335)

snp_120m1=merge(top_a_ml1,M_new120m1,
by=intersect("marcas","marcas"))
Z120m1=t(snp_120m1[,-(1:2)])
colnames(Z120m1)=c(as.matrix(snp_120m1[,1]))
write.table(Z120m1,"Z120m1.txt",row.names=FALSE,col.names=TRUE
,quote=FALSE)

Z120m1=read.table("Z120m1.txt",h=T)
Z120m1=as.matrix(Z120m1)
dim(Z120m1)
A_new=A.mat(Z120m1)
n=335
aux5_120m1=A_new+diag(n)*10^-6
head(aux5_120m1)
head(Z120m1)
```

#validação com as 120 marcas

```
gbv_ml1_120=NULL
eff_ml1_120=matrix(0,ncol(Z120m1),nrow(Z120m1))
system.time(
for(i in 1:nrow(Z120m1))
{
u_ml1_vc_120=(lmekin(y_adj[-i] ~ (1|data2$ID[-i]) ,
method="ML", data=data2,varlist=list(aux5_120m1[-i,-
i])))$coefficients$random$data2.ID..i.
eff_ml1_120[,i]=ginv(t(Z120m1[-i,])%*%Z120m1[-
i,])%*%t(Z120m1[-i,])%*%u_ml1_vc_120
gbv_ml1_120[i]=Z120m1[i,]%*%eff_ml1_120[,i]
}
)
write.table(gbv_ml1_120,"gbv_ml1_120.txt",row.names=FALSE,col.
names=FALSE,quote=FALSE)
```

#Vetor de efeito de marcadores

```
mean_eff_ml1_120=NULL
for(i in 1:nrow(eff_ml1_120))
{
mean_eff_ml1_120[i]=sum(eff_ml1_120[i,])/ncol(eff_ml1_120)
}
mean_eff_ml1_120
mean_eff_ml1_120=cbind(colnames(Z120m1),mean_eff_ml1_120)
colnames(mean_eff_ml1_120)=c("marker","effect")
write.table(mean_eff_ml1_120,"eff_gblup_ml1_120.txt",row.names
=FALSE,col.names=TRUE,quote=FALSE)
```

#modelo de fragilidade de Cox com censura (S1)

```
a_cox=abs(ginv(t(M)%*%M)%*%t(M)%*%u_cox)  #efeitos de todas as marcas
a_cox_fim=cbind(marcas,a_cox)
top_a_cox=a_cox_fim[order(a_cox_fim[ ,2],decreasing = TRUE),
][1:120,]
colnames(top_a_cox)=c("marcas","ef.cox")

M_new120c=cbind(colnames(M),t(M))
colnames(M_new120c)=c("marcas",1:335)

snp_120c=merge(top_a_cox,M_new120c,
by=intersect("marcas","marcas"))
Z120c=t(snp_120c[,-(1:2)])
colnames(Z120c)=c(as.matrix(snp_120c[,1]))
write.table(Z120c,"Z120c.txt",row.names=FALSE,col.names=TRUE,q
uote=FALSE)

Z120c=read.table("Z120c.txt",h=T)
Z120c=as.matrix(Z120c)
dim(Z120c)
A_new=A.mat(Z120c)
n=335
G_120c=A_new+diag(n)*10^-6
head(G_120c)
```

validação com as 120 marcas

```
gbv_cox_vc_120=NULL
eff_cox_120=matrix(0,ncol(Z120c),nrow(Z120c))
system.time(
for(i in 1:nrow(Z120c))
{
u_cox_vc1_120=(coxme(Surv(y_adj,cen)[-i] ~ (1|ID[-i]) ,
data=data, varlist=list(G_120c[-i,-i])))$frail$ID..i.
eff_cox_120[,i]=ginv(t(Z120c[-i,])%*%Z120c[-i,])%*%t(Z120c[-
i,])%*%u_cox_vc_120
gbv_cox_vc_120[i]=Z120c[i,]%*%eff_cox_120[,i]
}
)
write.table(gbv_cox_vc_120,"gbv_cox_vc_120.txt",row.names=FALS
E,col.names=FALSE,quote=FALSE)
```

#Vetor de efeito de marcadores

```
gbv_cox_vc_120
mean_eff_cox_120=NULL
for(i in 1:nrow(eff_cox_120))
{
mean_eff_cox_120[i]=sum(eff_cox_120[i,])/ncol(eff_cox_120)
}
mean_eff_cox_120
mean_eff_cox_120=cbind(colnames(Z120c),mean_eff_cox_120)
colnames(mean_eff_cox_120)=c("marker","effect")
write.table(mean_eff_cox_120,"eff_cox_120.txt",row.names=FALSE
,col.names=TRUE,quote=FALSE)
```

#modelo de fragilidade de Cox sem censura (S2)

```
a_cox1=abs(ginv(t(M)%*%M)%*%t(M)%*%u_cox1)      #efeitos de todas as
marcas
a_cox_fim=cbind(marcas,a_cox1)
top_a_cox=a_cox_fim[order(a_cox_fim[ ,2],decreasing = TRUE),
][1:120,]
colnames(top_a_cox)=c("marcas","ef.cox")

M_new120c1=cbind(colnames(M),t(M))
colnames(M_new120c1)=c("marcas",1:335)

snp_120c1=merge(top_a_cox,M_new120c1,
by=intersect("marcas","marcas"))
Z120c1=t(snp_120c1[,-(1:2)])
colnames(Z120c1)=c(as.matrix(snp_120c1[,1]))
write.table(Z120c1,"Z120c1.txt",row.names=FALSE,col.names=TRUE
,quote=FALSE)

Z120c1=read.table("Z120c1.txt",h=T)
Z120c1=as.matrix(Z120c1)
dim(Z120c1)
A_new=A.mat(Z120c1)
n=335
G_120c1=A_new+diag(n)*10^-6
head(G_120c1)

# validação com as 120 marcas
gbv_cox_vc1_120=NULL
eff_cox1_120=matrix(0,ncol(Z120c1),nrow(Z120c1))
system.time(
for(i in 1:nrow(Z120c1))
{
u_cox_vc1_120=(coxme(Surv(y_adj,cen)[-i] ~ (1|ID[-i]) ,
data=data1, varlist=list(G_120c1[-i,-i])))$frail$ID..i.
eff_cox1_120[,i]=ginv(t(Z120c1[-i,])%*%Z120c1[-
i,])%*%t(Z120c1[-i,])%*%u_cox_vc1_120
gbv_cox_vc1_120[i]=Z120c1[i,]%*%eff_cox1_120[,i]
}
)
write.table(gbv_cox_vc1_120,"gbv_cox_vc1_120.txt",row.names=FA
LSE,col.names=FALSE,quote=FALSE)

#Vetor de efeito de marcadores
gbv_cox_vc1_120
mean_eff_cox1_120=NULL
for(i in 1:nrow(eff_cox1_120))
{
mean_eff_cox1_120[i]=sum(eff_cox1_120[i,])/ncol(eff_cox1_120)
}
mean_eff_cox1_120
mean_eff_cox1_120=cbind(colnames(Z120c1),mean_eff_cox1_120)
colnames(mean_eff_cox1_120)=c("marker","effect")
write.table(mean_eff_cox1_120,"eff_cox1_120.txt",row.names=FAL
SE,col.names=TRUE,quote=FALSE)
```