Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Análise genética de vacas em relação à sua idade no último parto através de modelos de sobrevivência

Eduarda Godoy Afonso

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Análise genética de vacas em relação à sua idade no último parto através de modelos de sobrevivência

Eduarda Godoy Afonso
Orientadora: Daiane Aparecida Zuanetti
Coorientadora: Sabrina Luzia Caetano

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Dezembro de 2025

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

Genetic analysis of cows in relation to their age at last calving through survival models

Eduarda Godoy Afonso
Advisor: Daiane Aparecida Zuanetti
Co-advisor: Sabrina Luzia Caetano

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
June 2025

Eduarda Godoy Afonso

Análise genética de vacas em relação à sua idade no último parto através de modelos de sobrevivência

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Eduarda Godoy Afonso e aprovado pela banca examinadora.

Aprovado em 03 de dezembro de 2025

Banca Examinadora:

- Daiane Aparecida Zuanetti
- Andressa Cerqueira
- Teresa Cristina Martins Dias

Resumo

No Brasil, o estudo sobre a longevidade bovina é de extrema importância devido à posição em que o país se encontra na exportação e produção de carne nesse setor e, por isso, existe o desejo de alcançar a maior rentabilidade nesta atividade. A longevidade econômica das vacas é totalmente relacionada com o desempenho reprodutivo, sendo a característica relacionada à idade da vaca ao último parto (IVUP) uma variável muito importante nesse contexto. Como essa variável não pode ser perfeitamente medida em alguns cenários, geralmente aplica-se um critério de censura a essa característica, que tem como base a diferença entre a data do último parto observado da vaca e o último parto da fazenda. Caso essa diferença seja maior ou igual a 36 meses, a vaca é considerada como não estando mais no período reprodutivo, pois está há muito tempo sem parir e será descartada. Caso contrário, a vaca ainda é considerada como reprodutiva e tem a sua idade no último parto censurada, já que ainda existe a possibilidade dela ter uma nova gestação. Portanto, um dos objetivos desse estudo é analisar geneticamente a variável IVUP utilizando modelos de fragilidade aplicados à análise de sobrevivência. A utilização desse tipo de modelo se justifica pela inclusão de efeitos aleatórios, permitindo predizer os fatores genéticos associados a cada animal e também considerar associação entre eles. Outro objetivo do estudo é identificar variáveis genéticas que influenciam na longevidade bovina, e isso será feito por meio de um estudo de associação genômica ampla (GWAS). A análise é feita utilizando uma amostra de 862 vacas, filhas de 47 touros diferentes. Por questão de confidencialidade, a origem dos dados é mantida em sigilo.

Palavras-chave: *censura, GWAS, herdabilidade.*

Abstract

In Brazil, the study of bovine longevity is extremely important due to the country's position in the export and production of this sector and, for this reason, there is a strong interest in achieving maximum profitability in this activity. The economic longevity of cows is totally related to reproductive performance, with the trait known as the age of the cow at last calving (ACLC) being a very important variable in this context. Since this variable cannot always be perfectly measured in some scenarios, a censoring criterion is generally applied to this characteristic, which is based on the difference between the date of the cow's last observed calving and the date of the last calving on the farm. If this difference is greater than or equal to 36 months, the cow failed, as she has gone too long without calving and will be discarded. Otherwise, the cow is still considered reproductive and her age at last calving is censored, since there is still a possibility of her having another pregnancy. Therefore, one of the objectives of this study is to genetically analyze the IVUP variable using frailty models applied to survival analysis. The use of this type of model is justified by the inclusion of random effects, allowing the prediction of genetic factors associated with each animal and also considering associations between them. Another objective of the study is to identify genetic variables that influence bovine longevity, and this will be done through a Genome-Wide Association Study (GWAS). The analysis is done using a sample of 862 cows, daughters of 47 different bulls. For confidentiality reasons, the origin of the data is kept secret.

Keywords: *censorship, GWAS, heritability.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

O Brasil é o maior exportador e o segundo maior produtor de carne do mundo, segundo o Ministério da Agricultura e Pecuária (2024). Nesse contexto, o estudo genético sobre a longevidade bovina, que representa o tempo de permanência do animal no rebanho, tem se tornado muito importante para que este tipo de atividade seja cada vez mais produtiva. Além disso, o custo da produção de proteína bovina é diretamente relacionado com o desempenho reprodutivo das vacas. Para que elas produzam um número maior de bezerros, elas devem ser precoces sexualmente, com um intervalo pequeno entre um parto e outro.

Nesse sentido, uma variável que é relevante para análise é a idade da vaca ao último parto (IVUP), que é considerada como variável resposta neste estudo. Como essa variável não pode ser perfeitamente medida em alguns cenários, geralmente aplica-se um critério de censura que considera a diferença entre a data do último parto da vaca em questão e a data do último parto da fazenda. Se essa diferença for maior ou igual a 36 meses, a vaca não está mais no período reprodutivo, já que está há muito tempo sem parir e, portanto, será descartada. Caso contrário, a vaca ainda pode ser reprodutiva e tem a sua idade censurada, pois existe a possibilidade dela ter uma nova gestação. Este critério foi apresentado e utilizado em estudos genéticos analisando a herdabilidade da IVUP por Caetano (2011), usando alguns modelos de sobrevivência que têm sido pouco utilizados na literatura para análises genéticas e melhoramento genético de animais. Assumindo que uma vaca, em geral, pode parir a cada 12 meses (9 meses de gestação e 3 meses para recuperação e nova fecundação), a escolha de 36 meses sem parir para considerar a idade reprodutiva da vaca como censura ou não censura é para termos um tempo de espera caso algum imprevisto tenha acontecido e uma vaca tenha demorado mais que o normal para

parir antes de realmente se tornar não reprodutiva.

Devido à característica da variável resposta e à necessidade de se incluir efeitos aleatórios no modelo que considerem o grau de parentesco entre as vacas da amostra analisada e meçam individualmente o efeito genético em cada uma delas, são usados modelos de sobrevivência mistos (Caetano, 2011), também conhecidos como modelos de fragilidade na área de análise de sobrevivência e confiabilidade. Nesse trabalho, em particular, utilizamos o modelo de Cox com efeitos aleatórios como aplicado em Minguillo (2016).

Além de estimar a herdabilidade e predizer os efeitos aleatórios para identificar animais que geneticamente são mais longevos, um dos desafios genômicos mais atuais é selecionar variáveis (regiões) genômicas que estão mais associadas a uma característica de interesse via estudo GWAS (*Genome-Wide Association Study)*, que têm sido pouco explorado em modelos de sobrevivência.

A importância do GWAS (análise de associação genômica ampla) em estudos relacionados à genética é muito grande. Os GWAS são estudos (que podem envolver diferentes metodologias estatísticas) capazes de nos permitir detectar associações entre variações genéticas e uma característica de interesse específica (Nayara, 2019). A partir do GWAS, é possível detectar genes ou regiões genômicas que explicam determinadas relações ou acontecimentos em animais de uma amostra e, consequentemente, de uma população (Nayara, 2019). São testadas várias variantes genéticas em diversos genomas, sendo eles a sequência completa de DNA de um indivíduo que fornece todas as suas informações hereditárias (Santos, 2025). Isso é feito para identificar as variantes que são estatisticamente associadas a uma característica de interesse (Uffelmann *et al.*, 2021). Para esse estudo, o uso do GWAS é muito importante para entender a relação genética entre vacas e o quanto esse fator impacta na variável de interesse IVUP.

O trabalho está organizado como segue. No Capítulo 2, apresentamos conceitos fundamentais de modelos de análise de sobrevivência e, mais especificamente, o modelo de fragilidade de Cox que é utilizado nesse estudo. No Capítulo 3, são apresentados os conceitos mais importantes sobre GWAS, que são utilizados para a análise de associação entre regiões genéticas e IVUP. O Capítulo 4 contém a descrição sobre o banco de dados utilizado nas análises e, finalmente, nos Capítulos 5 e 6, apresentamos os resultados encontrados a partir de todas as análises feitas e também concluímos e discutimos todo o estudo feito e possíveis análises futuras.

Capítulo 2

Análise de Sobrevivência

A análise de sobrevivência é um conjunto de métodos na estatística usados para a análise de dados em que a variável de interesse, ou seja, a variável resposta é o tempo até a ocorrência de um (ou mais) evento de interesse. Esse tempo pode ser observado em dias, horas, semanas, meses, etc. Um dos interesses desse tipo de metodologia está em modelar a variável resposta (tempo) na presença ou não de covariáveis (Pinheiro, 2022).

Usualmente, as unidades amostrais (indivíduos) são acompanhadas por um período de tempo que é pré-fixado e os tempos das unidades que sofreram o evento de interesse são registrados. Outra situação é quando o estudo continua até que uma quantidade fixada de unidades amostrais sofra o evento de interesse.

Dessa maneira, para cada unidade amostral, o tempo de ocorrência do evento pode ser de dois tipos de registro:

- completo: sabe-se o tempo exato da ocorrência do evento, pois ocorreu no período de estudo;
- incompleto: não acontece o evento de interesse no período de estudo ou não se sabe o tempo exato de ocorrência.

Um dos conceitos mais importantes da análise de sobrevivência é, portanto, a censura, que acontece nos casos em que se tem informações incompletas sobre o tempo. Existem diferentes tipos de censuras (Kalbfleisch e Prentice, 2002), são eles:

- à direita: nesse caso, o evento não ocorreu até o final do estudo, ou seja, o tempo de ocorrência deste evento está à direita do tempo registrado. Para esse tipo, existem 3 mecanismos distintos:

  - Tipo I: o estudo foi conduzido até um tempo pré-estabelecido, as unidades amostrais que não sofreram o evento são chamadas de censuradas;
  - Tipo II: o estudo termina após o evento ocorrer para um número pré-estabelecido de unidades amostrais, as que não sofreram o evento são chamadas de censuradas;
  - Aleatória: o indivíduo é retirado do estudo sem sofrer o evento;

- intervalar: o evento ocorre entre dois tempos conhecidos, ou seja, o tempo exato de ocorrência não é conhecido, apenas seu intervalo;
- à esquerda: o evento ocorreu antes da unidade amostral ser observada no estudo, ou seja, a ocorrência do evento foi anterior ao início do estudo. Para este tipo de censura, também temos os mecanismos do tipo I e tipo II;
- dados truncados: as unidades amostrais são excluídas do estudo por motivos relacionados à ocorrência do evento.

Para o estudo em questão, o único tipo de censura de interesse é a censura à direita.

Existem algumas funções importantes na análise de sobrevivência, explicadas a seguir (Colosimo e Giolo, 2021). A função densidade de probabilidade $f(t)$ para o tempo de ocorrência T (variável aleatória contínua não negativa) é definida como:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

e representa o limite da probabilidade do evento de interesse acontecer no intervalo de tempo $[t; t + \Delta t]$, sendo $f(t) \geq 0$ para todo t e $\int_0^\infty f(t)\, dt = 1$.

A função de sobrevivência $S(t)$ é a probabilidade de uma unidade amostral sobreviver mais do que determinado tempo t, o que é complementar à função de distribuição acumulada ($F(t)$). Sendo $F(t) = P(T \leq t)$ a probabilidade do evento ocorrer antes do tempo t, a função de sobrevivência é definida como:

$$S(t) = P(T \geq t) = 1 - F(t).$$

A função de risco, por sua vez, é o potencial de um indivíduo sofrer o evento de interesse em um intervalo de tempo, dado que ele sobreviveu até o tempo t, ou seja, é uma taxa instantânea de falha durante um intervalo de tempo muito pequeno. Ela é definida como:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

Outro conceito interessante em análise de sobrevivência são as curvas de Kaplan-Meier, que são um método que resume os dados de sobrevivência, dividindo o período de acompanhamento em vários pequenos intervalos de tempo, evidenciando para cada um deles o número de casos acompanhados e o número de ocorrências do evento de interesse. Essas curvas apresentam a probabilidade de sobrevivência representada em função do tempo, em que cada degrau representa um evento de interesse. Em resumo, essas curvas indicam a probabilidade de um evento não ocorrer ao longo do tempo (ScienceDirect, 2025a).

2.1 Modelo de Cox

Um dos modelos estatísticos mais utilizados para analisar o tempo de ocorrência em uma amostra de unidades independentes é o modelo de Cox ou modelo de riscos proporcionais de Cox (Cox, 1972). A partir dele, conseguimos avaliar simultaneamente o efeito de diversos fatores na sobrevivência de cada unidade amostral. Ou seja, com ele, conseguimos identificar como fatores específicos influenciam a taxa de ocorrência de um evento específico (Kassambara, 2025). A suposição de riscos proporcionais é um conceito que afirma que a razão de risco entre dois indivíduos permanece constante ao longo do tempo, por consequência, as curvas de sobrevivência das duas unidades amostrais não se cruzam.

O modelo de Cox é expresso da seguinte forma:

$$h(t; \mathbf{x}) = h_0(t) \exp(\mathbf{x}'\boldsymbol{\beta}),$$

em que:

- $\boldsymbol{\beta}$: vetor de coeficientes de regressão de tamanho $p \times 1$, sendo p é o número de variáveis explicativas;
- $h_0(t)$: é a função de risco base, ou seja, é o risco de um indivíduo sofrer o evento de interesse no tempo t sem levar em consideração as covariáveis, assumindo valores não negativos no tempo;

- $\mathbf{x}$: vetor de covariáveis de tamanho $p \times 1$.

Para que esse modelo possa ser ajustado, é necessário que algumas suposições sejam satisfeitas. A primeira delas é garantir que a taxa de falha entre dois indivíduos distintos seja proporcional para quaisquer indivíduos da amostra, independente do tempo, ou seja, que as taxas sejam constantes ao longo do tempo. Além disso, a independência entre os indivíduos também é uma suposição a ser satisfeita para o modelo de riscos proporcionais de Cox. Observando o modelo definido acima e sendo m e k dois indivíduos independentes, temos a seguinte taxa de riscos proporcionais:

$$\begin{aligned}\frac{h_m(t;\mathbf{x})}{h_k(t;\mathbf{x})} &= \frac{h_0(t)\exp(\mathbf{x}'_m\boldsymbol{\beta})}{h_0(t)\exp(\mathbf{x}'_k\boldsymbol{\beta})} \\ &= \frac{\exp(\mathbf{x}'_m\boldsymbol{\beta})}{\exp(\mathbf{x}'_k\boldsymbol{\beta})} \\ &= \exp(\mathbf{x}'_m\boldsymbol{\beta} - \mathbf{x}'_k\boldsymbol{\beta}) \\ &= \exp\{(\mathbf{x}'_m - \mathbf{x}'_k)\boldsymbol{\beta}\}\end{aligned}$$

que não depende do tempo.

2.2 Modelos de Cox com Fragilidade

O uso do modelo de Cox (Cox, 1972) pode ser limitado para situações em que os dados possuem correlação, pois nele existe a suposição de independência entre os indivíduos e isso nem sempre acontece. Para o estudo em questão, a correlação genética entre as vacas com grau de parentesco é muito importante, fazendo com que o uso do modelo de Cox apresentado anteriormente não seja adequado. Nesse caso, uma das opções é utilizar o modelo de Cox com fragilidade do indivíduo, que é útil nos casos em que os dados são correlacionados (Matuda, 2005). Nesse tipo de modelo, são incluídos efeitos aleatórios (por exemplo, efeitos genéticos), que são chamados de fragilidade (Caetano, 2011).

O componente de fragilidade foi introduzido por Vaupel *et al.* (1979) em modelos de sobrevivência univariados, sendo que esse termo é definido como uma quantidade aleatória não observada. A notação v_m é utilizada para o termo da fragilidade, em que m indica o animal e $m = 1, \ldots, n$. Nesse modelo, é possível descrever as características não observáveis (genéticas) que atuam sobre o risco de cada vaca não ser mais reprodutiva

em relação à IVUP. Segundo Caetano (2011), adotando o modelo de Cox, o modelo de fragilidade pode ser definido, resumidamente, como:

$$h_m(t;\mathbf{x}) = h_0(t)v_m\exp(\mathbf{x}'_m\boldsymbol{\beta})$$

sendo:

- $h_m(t;\mathbf{x})$: função de risco da vaca m falhar no tempo t em relação à IVUP e tendo as covariáveis $\mathbf{x}_m$;
- $h_0(t)$: função de risco base no tempo t;
- $\boldsymbol{\beta}$: vetor de efeitos fixos (relacionados às covariáveis ambientais e de comportamento disponíveis no estudo e também podendo considerar variáveis genéticas) de tamanho $p\times 1$;
- $\mathbf{x}_m$: vetor com os valores observados para a vaca m das covariáveis (vetor de incidência dos efeitos) de tamanho $p\times 1$.

Usualmente, a distribuição Gama é assumida para o termo de fragilidade (Vaupel *et al.*, 1979), mas no caso de análise genética, outras distribuições podem ser assumidas. Se aplicarmos a transformação $v_m = \exp(s_m)$ ou $s_m = \log(v_m)$, o modelo pode ser definido como:

$$h_m(t;\mathbf{x}) = h_0(t)\exp(s_m + \mathbf{x}'_m\boldsymbol{\beta})$$

e podemos assumir uma distribuição log-gama para o efeito aleatório s_m. Estudos que geralmente consideram a distribuição Gama ou log-gama para os efeitos aleatórios ou de fragilidade também consideram independência entre eles, que não é o caso aqui, pois temos dependência entre as vacas. Na análise de dados genéticos, portanto, é comum assumir uma distribuição normal multivariada para o vetor de todos os efeitos de fragilidade $\mathbf{s}$ e relacioná-los através da matriz de relacionamento entre as vacas. A matriz de relacionamento, construída através de informações familiares entre os indivíduos da amostra, ou a matriz genômica de relacionamento (GRM) construída a partir de informações genômicas, em situações que informações familiares são escassas ou incompletas, compõem a matriz de variância e covariância da distribuição normal multivariada assumida para os efeitos aleatórios, como será matematicamente definido a seguir.

2.2.1 Estimação

Para que seja possível realizar inferências e obter conclusões a partir do modelo, é necessário estimar seus parâmetros e predizer os efeitos aleatórios. Para os termos de fragilidade em $\mathbf{s} = (s_1, s_2, ..., s_n)'$, assumimos uma distribuição normal com média $\mathbf{0}$ e variância $\sigma_s^2 A$, em que A é a matriz de relacionamento entre as vacas.

Os parâmetros que são estimados são: $\boldsymbol{\beta}$ e σ_s^2. Se tratando da estimação destes parâmetros, Cox propôs uma solução alternativa ao método de máxima verossimilhança usual, já que ele não é adequado nesse caso pela presença do componente não paramétrico ($h_0(t)$) (Cox, 1972). Para isso, utiliza-se a função de verossimilhança parcial. Ela é considerada como parcial, pois não aborda a parte não paramétrica do modelo, isto é, a taxa de risco base, para construir o método de estimação dos parâmetros. Com isso, assumindo $\mathbf{s}$ como conhecido, temos a seguinte função de verossimilhança parcial:

$$L(\boldsymbol{\beta} \mid \mathbf{x}, \mathbf{s}, \mathbf{t}, \boldsymbol{\delta}) = \prod_{m=1}^{n} \left(\frac{\exp(s_m + \mathbf{x}_m' \boldsymbol{\beta})}{\sum_{k \in R(t_m)} \exp(s_k + \mathbf{x}_k' \boldsymbol{\beta})} \right)^{\delta_m}$$

em que $R(t_m)$ é o conjunto das unidades amostrais em risco no tempo t_m, isto é, o conjunto das unidades amostrais que não sofreram o evento e não foram censuradas até o tempo imediatamente anterior a t_m. Além disso, δ_m é uma variável indicadora que assume o valor 0 quando ocorre censura e o valor 1 quando o tempo é completo, ou seja, quando o evento ocorre, sendo $\mathbf{t} = (t_1, t_2, ..., t_n)$ e $\boldsymbol{\delta} = (\delta_1, \delta_2, ..., \delta_n)$.

Como assumimos que $\mathbf{s} = (s_1, s_2, ..., s_n)' \sim N_n(\mathbf{0}, \sigma_s^2 A)$ e de acordo com Therneau (2024a), conseguimos estimar o modelo de efeitos aleatórios via *REML (Restrited Maximum Likelihood)*, a partir da seguinte expressão:

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma_s^2 \mid \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}, \mathbf{s}) &= L(\boldsymbol{\beta} \mid \mathbf{x}, \mathbf{s}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(\mathbf{s} \mid \sigma_s^2) \\ &= \prod_{m=1}^{n} \left(\frac{\exp(s_m + \mathbf{x}_m' \boldsymbol{\beta})}{\sum_{k \in R(t_m)} \exp(s_k + \mathbf{x}_k' \boldsymbol{\beta})} \right)^{\delta_m} \cdot \frac{1}{(2\pi)^{n/2} |\sigma_s^2 A|^{1/2}} \exp\left(-\frac{1}{2\sigma_s^2} (\mathbf{s} - \mathbf{0})' A^{-1} (\mathbf{s} - \mathbf{0}) \right) \\ &= \prod_{m=1}^{n} \left(\frac{\exp(s_m + \mathbf{x}_m' \boldsymbol{\beta})}{\sum_{k \in R(t_m)} \exp(s_k + \mathbf{x}_k' \boldsymbol{\beta})} \right)^{\delta_m} \cdot \frac{1}{(2\pi)^{n/2} |\sigma_s^2 A|^{1/2}} \exp\left(-\frac{1}{2\sigma_s^2} \mathbf{s}' A^{-1} \mathbf{s} \right), \end{aligned}$$

em que $p(\mathbf{s} \mid \sigma_s^2)$ é a função de probabilidade para $\mathbf{s}$.

Primeiramente, essa expressão é integrada em relação aos efeitos aleatórios ($\mathbf{s}$) para encontrarmos a estimativa de $\boldsymbol{\beta}$ e σ_s^2, sendo IPL a verossimilhança parcial integrada, então:

$$IPL(\boldsymbol{\beta}, \sigma_s^2 \mid \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) = \frac{1}{(2\pi)^{n/2}|\sigma_s^2 A|^{1/2}} \int_{\mathbf{s}} \left[\prod_{m=1}^{n} \left(\frac{\exp(s_m + \mathbf{x}'_m\boldsymbol{\beta})}{\sum_{k \in R(t_m)} \exp(s_k + \mathbf{x}'_k\boldsymbol{\beta})} \right)^{\delta_m} \cdot \exp\left(-\frac{1}{2\sigma_s^2} \mathbf{s}' A^{-1} \mathbf{s} \right) \right] d\mathbf{s}.$$

Com isso, a estimativa de $\boldsymbol{\beta}$ e σ_s^2 é obtida pela maximização do IPL (Therneau, 2024a). Como a integral IPL não é tratável, o passo-chave na sua computação é a substituição da log-verossimilhança parcial pela sua expansão em séries de Taylor truncadas na segunda ordem.

Outra maneira de estimar σ_s^2 é via $REML$. Ela é feita a partir da integração do IPL em relação aos efeitos fixos ($\boldsymbol{\beta}$). Sendo assim:

$$REML(\sigma_s^2 \mid \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) = \frac{1}{(2\pi)^{n/2}|\sigma_s^2 A|^{1/2}} \int_{\boldsymbol{\beta}} \int_{\mathbf{s}} \left[\prod_{m=1}^{n} \left(\frac{\exp(s_m + \mathbf{x}'_m\boldsymbol{\beta})}{\sum_{k \in R(t_m)} \exp(s_k + \mathbf{x}'_k\boldsymbol{\beta})} \right)^{\delta_m} \cdot \exp\left(-\frac{1}{2\sigma_s^2} \mathbf{s}' A^{-1} \mathbf{s} \right) \right] d\mathbf{s}\, d\boldsymbol{\beta}$$

e nesse caso, a estimativa é obtida via maximização desta expressão em relação à σ_s^2. Apesar de alguns autores recomendarem usar a estimativa $REML$ para σ_s^2, o estimador de máxima verossimilhança parcial é o mais implementado atualmente.

Além disso, h_0, que é a função de risco base, não possui distribuição especificada e pode ser estimada de maneira não-paramétrica utilizando o método proposto por Breslow (1972). A função de risco base acumulada é estimada como:

$$\hat{H}_0(t) = \sum_{m:t_m<t} \left(\frac{d_m}{\sum_{k \in R(t_m)} \exp(\mathbf{x}'_k \hat{\boldsymbol{\beta}})} \right),$$

em que d_m representa o número de ocorrências do evento em t_m, sendo t_m o tempo de ocorrência do m-ésimo evento e $R(t_m)$ representa o conjunto das unidades amostrais em risco imediatamente antes do tempo t_m.

Mais detalhes técnicos desse processo de estimação não são fornecidos aqui, mas este modelo pode ser estimado via biblioteca *coxme* do R (Therneau, 2024b,c).

Testes de hipóteses para verificar a significância dos efeitos fixos e identificar os fatores mais relevantes também podem ser realizados. No pacote *coxme* do R, os p-valores associados aos efeitos fixos são encontrados a partir do teste de Wald (Therneau e Grambsch, 2000), realizado através da razão entre o quadrado da diferença do estimador do parâmetro e o seu valor sob H_0 e a variância do estimador. A estatística do teste de Wald (W), sob H_0 e com n suficientemente grande, tem distribuição χ^2_1 (Freitas, 2025) e é dada por:

$$W = (\frac{(\hat{\beta}-\beta_0)^2}{Var(\hat{\beta})}) \sim \chi^2_1$$

Sejam as hipóteses dadas por:

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0, \end{cases}$$

rejeitamos a hipótese nula H_0 se o valor da estatística de teste encontrada for maior que o valor crítico da distribuição χ^2_1 a um nível de significância α, geralmente, $\alpha = 5\%$. Outra forma de rejeitar H_0 é quando o p-valor associado ao valor da estatística na amostra for menor que o nível de significância α.

2.3 Herdabilidade

A herdabilidade é um parâmetro genético que, em modelos lineares, é definida pela razão da variância genética e da variabilidade total de uma variável da população. Segundo Caetano (2011), a variação genética é conhecida como variação aditiva (ou genotípica), enquanto a variação total é conhecida como variação fenotípica. Essa métrica consegue medir a importância da genética na determinação de uma característica em uma população (Hill, 2013). Portanto, a expressão da herdabilidade em modelos lineares é expressa da seguinte forma:

$$h^2 = \frac{\sigma^2_a}{\sigma^2_p}.$$

O fenótipo de um indivíduo se dá pela soma entre a contribuição ambiental na qual esse ser vive e seu genótipo, que representa a parte genética. Com isso, a expressão encontrada acima, usando a notação matemática adotada nesse trabalho e considerando σ^2 como a variabilidade total da IVUP, a herdabilidade pode ser escrita da seguinte forma:

$$h^2 = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2}.$$

Nesse sentido, a herdabilidade fornece a informação do quanto a variância de uma característica de interesse de um indivíduo, a IVUP no estudo em questão, é atribuída às suas características genéticas. Além dos fatores genéticos, os fatores ambientais também influenciam na variabilidade, com isso, quanto maior for o efeito dos fatores ambientais, menor será o efeito da genética, ou seja, eles possuem relação inversa.

A herdabilidade assume valores entre 0 e 1 e, segundo Aidar de Queiroz (2025), temos o seguinte contexto:

- $h^2 < 0.2$, representa baixa herdabilidade (os fatores ambientais são mais importantes na variabilidade, a característica é pouco herdável). Nessa situação, geralmente não compensa investir em estudos de melhoramento genético para aprimorar a característica dos indivíduos e sim investir em melhores condições ambientais;

- $0.2 < h^2 < 0.4$, representa média herdabilidade (as diferenças genéticas e fatores ambientais têm a mesma influência, característica moderadamente herdável);

- $h^2 > 0.4$, representa alta herdabilidade (diferenças genéticas são muito importantes, característica bastante herdável). Nessa situação, geralmente compensa investir em estudos de melhoramento genético para aprimorar a característica dos indivíduos pelas condições genéticas.

Estimada a variância total da característica em análise e o σ_s^2 pelo modelo de fragilidade de Cox, conseguimos também estimar a herdabilidade associada à característica em análise. Apesar do modelo de Cox não ser um modelo linear, a herdabilidade através dele é calculada de maneira semelhante à dos modelos lineares.

Capítulo 3

GWAS - *Genome-Wide Association Study*

O GWAS (Estudo de Associação Genômica Ampla) é um estudo utilizado para identificar associações entre informações genéticas e uma característica de interesse específica. Ou seja, ele detecta a relação entre genótipos (composição genética de um indivíduo, que são todos os seus genes ou marcadores genéticos) e fenótipos (características observáveis de um indivíduo) (Santos, 2024).

Nesse sentido, podemos entender que GWAS são estudos baseados em métodos estatísticos que buscam encontrar as associações entre regiões genômicas e características de interesse, a fim de identificar possíveis regiões que possuem maior efeito sobre um determinado fenótipo (Nayara, 2019). Identificar quais regiões genéticas impactam com maior relevância a longevidade reprodutiva de bovinos é particularmente importante nesse estudo.

O GWAS pode ser feito a partir dos seguintes passos (Cerqueira *et al.*, 2024):

- 1 - Coletar os dados de DNA e do fenótipo de um grupo de indivíduos;
- 2 - Realizar a genotipagem dos marcadores genéticos ou SNPs (polimorfismos de nucleotídeo único, do inglês *Single nucleotide polymorphisms*), explicados melhor a seguir;
- 3 - Conduzir uma análise de ancestralidade se os indivíduos forem originários de regiões e ancestrais muito diferentes, que não é o caso aqui;
- 4 - Fazer a análise estatística de associação. Para este estudo, verificamos a asso-

ciação entre cada SNP e a variável IVUP através do teste de significância do efeito fixo de cada SNP no modelo de Cox com fragilidade para a IVUP;

- 5 - Análise dos resultados.

O DNA (Ácido Desoxirribonucleico) é uma molécula que carrega a informação genética dos organismos e fica dentro de cromossomos presentes no núcleo de todas as células dos seres vivos. É formado por uma fita dupla em forma de espiral (dupla hélice), composta por nucleotídeos. Sua estrutura contém quatro bases nitrogenadas: Adenina (A), Timina (T), Citosina (C) e Guanina (G) (Magalhães, s.d.).

Um SNP corresponde a uma variação em uma única posição na sequência de DNA. Um possível exemplo é quando uma posição específica na maior parte da população é ocupada por um C, mas em alguma parte dos indivíduos aparece a base T, então essa posição genética é um SNP. As duas possíveis variações, nesse exemplo C e T, são denominadas alelo mais frequente e alelo menos frequente, respectivamente, e codificados como:

- Alelo mais frequente: 0;
- Alelo menos frequente: 1.

Dados genotípicos, por sua vez, representam a codificação dos alelos do SNP no par de cromossomos observados, sendo representados da seguinte forma (Cerqueira *et al.*, 2024):

- 0: no SNP em questão, se o indivíduo possui ambos os alelos de maior frequência;
- 1: no SNP em questão, se o indivíduo possui um dos alelos como o de maior frequência e o outro como o de menor frequência;
- 2: no SNP em questão, se o indivíduo possui ambos os alelos de menor frequência.

Dessa maneira, os SNPs são utilizados como covariáveis genéticas para conseguirmos identificar áreas do genoma que são influentes no fenótipo em estudo.

3.1 Controle de Qualidade do Banco de Dados com SNPs

Antes de analisarmos o banco de dados de fato, é preciso realizar um controle de qualidade desse banco a fim de remover quaisquer inconsistências que podem aparecer nos dados ocasionadas pela etapa de genotipagem.

A seguir são apresentados as análises e filtros utilizados no controle de qualidade para filtrar os SNPs.

- Cálculo da frequência do alelo menor - MAF

  A frequência do alelo menor (MAF - *Minor Allele Frequency*) representa a proporção do alelo menos frequente, definido como alelo menor. Alelos com uma frequência muito pequena são raros e, por isso, dificultam a possibilidade de identificar qualquer associação entre fenótipo e genótipo. Por esse motivo, os SNPs cuja frequência de menor alelo (MAF) é muito baixa são retirados da análise de GWAS. Os SNPs com MAF $< 5\%$ são geralmente removidos das bases (Cerqueira *et al.*, 2024).

  Para calcularmos a frequência do alelo menor, utilizamos as seguintes expressões:

$$m_0 = 2n_0 + n_1,$$
$$m_1 = 2n_2 + n_1,$$

  sendo: n_0, n_1 e n_2 o número de vezes que os valores 0, 1 e 2 são encontrados na população, respectivamente. A partir dessas fórmulas, o cálculo para encontrar o MAF é da seguinte forma:

$$\min\{\frac{m_0}{m_0 + m_1}, \frac{m_1}{m_0 + m_1}\}.$$

- Cálculo do equilíbrio de Hardy-Weinberg - EHW

  É necessário testar se o SNP está em equilíbrio de Hardy-Weinberg, conceito que afirma que as frequências de alelo em uma população vão permanecer constantes de geração em geração, na ausência de fatores recorrentes da evolução. Para isso, é feito um teste de hipóteses utilizando o teste qui-quadrado, sendo a hipótese nula de que o SNP está em equilíbrio. Ou seja:

$$\begin{cases} H_0 : \text{o genótipo está em EHW}, \\ H_1 : \text{o genótipo não está em EHW}. \end{cases}$$

  A estatística de teste utilizada é da seguinte forma:

$$Q = \sum_{k=1}^{K} \frac{(o_k - e_k)^2}{e_k},$$

em que o_k e e_k são as frequências observadas e esperadas sob H_0 do k-ésimo genótipo, respectivamente, além disso, $\sum_{k=1}^{K} o_k = n, e_k = nP_k$ e P_k é a probabilidade do k-ésimo genótipo sob H_0, assumindo que o alelo paterno e materno são herdados de maneira independente (Zuanetti, 2025).

Valores altos de Q indicam que as frequências esperadas e observadas não possuem compatibilidade, com isso, a hipótese nula H_0 deve ser rejeitada. A distribuição de Q sob H_0 não é exata, mas para grandes amostras, ela se aproxima de χ^2_{k-1}.

Os SNPs que desviam do equilíbrio de Hardy-Weinberg devem ser removidos da análise, pois eles podem indicar erro de genotipagem. Em caso de amostras que já são resultados de melhoramento genético, quando se cruzam apenas indivíduos selecionados como os melhores geneticamente, precisamos ter atenção com esse filtro.

Capítulo 4

O Conjunto de Dados

O conjunto de dados usado contém informações de 862 vacas da raça Nelore, filhas de 47 touros diferentes. Por questões éticas, a origem do banco de dados é mantida em sigilo, a fim de preservar a identidade dos órgãos envolvidos. Em média, cada touro tem aproximadamente 20 filhas e cada vaca tem aproximadamente 3 partos.

As seguintes covariáveis estão presentes no banco de dados:

- Identificador do animal (da vaca);
- Identificador do pai;
- Identificador da mãe;
- Data de nascimento da vaca (contendo dia, mês e ano);
- IPP: Idade da vaca ao Primeiro Parto (em meses);
- Grupo contemporâneo da idade da vaca ao primeiro parto. Um grupo contemporâneo se refere a um grupo de animais da mesma linhagem genética e condição ambiental, sendo muito útil para avaliações genéticas. A partir das informações desse grupo, conseguimos comparar as vacas que pertencem a um mesmo grupo contemporâneo;
- Data do parto da vaca (contendo dia, mês e ano).

A base possui todos os partos de cada uma das vacas, cada um em uma linha. Como apenas o último parto dela é necessário para o estudo, ordenamos a data dos partos de forma decrescente. Posteriormente, apenas a primeira ocorrência de cada vaca foi

selecionada, deixando na base apenas o parto mais recente de cada uma delas. Para calcularmos os valores de algumas covariáveis que são necessárias, identificamos os grupos contemporâneos de cada uma das vacas e selecionamos a data de parto mais recente deste grupo.

A variável que fornece a data de nascimento da vaca possui informações de um período de 10 anos, indo de 03 de abril de 2010 até 29 de dezembro de 2020. Já as datas dos partos variam de 11 de março de 2013 até 22 de novembro de 2023, também abrangendo um período de 10 anos.

A partir dessas informações, as seguintes covariáveis foram criadas:

- Data do parto mais recente do grupo contemporâneo: obtida juntando os grupos contemporâneos e selecionando o parto mais recente de cada um (é utilizada para comparações entre as vacas de um mesmo grupo contemporâneo);
- Idade da Vaca no Último Parto (IVUP): calculada pela diferença entre a data do último parto da vaca e sua data de nascimento (em meses);
- Diferença: subtração entre a data do parto mais recente do grupo contemporâneo da vaca e o seu próprio último parto (em meses), é muito importante para definir censura;
- Censura: se a diferença for maior ou igual a 36 meses, a vaca será descartada, portanto, a censura recebe o valor 1, caso contrário, a vaca terá o seu tempo censurado e esta variável recebe o valor 0;
- Quantidade de partos de cada vaca: calculada pela quantidade de vezes que cada vaca aparece em uma linha da base inicial, indicando o número de partos que cada uma teve;
- Ano de nascimento: encontrada a partir da covariável data de nascimento, extraindo apenas o ano;
- Estação de parto: obtida através da data do parto da vaca, se ela estava entre outubro e março, a estação é considerada de águas (assumindo o valor 1), caso contrário, partos entre abril e setembro, a estação é de secas (assumindo o valor 0). Esta variável foi atribuída ao estudo, pois foi sugerida e definida por especialistas da área.

4.1 Análise Descritiva

A variável que indica a estação de parto da vaca está distribuída da seguinte forma: 362 vacas nasceram na estação de secas (igual a 0) e 500 delas têm a estação de nascimento na época de águas (igual a 1). Observando a variável que indica a diferença entre a data do último parto do grupo contemporâneo da vaca e a de seu último parto e também as censuras, encontramos 771 vacas com diferença menor do que 36 meses (para a variável censura, elas assumem o valor 0) e 91 têm diferença maior ou igual a 36 meses (censura igual a 1). Essa distribuição já indica um cenário desafiador para a modelagem estatística com uma porcentagem muito grande de censura (89.4%).

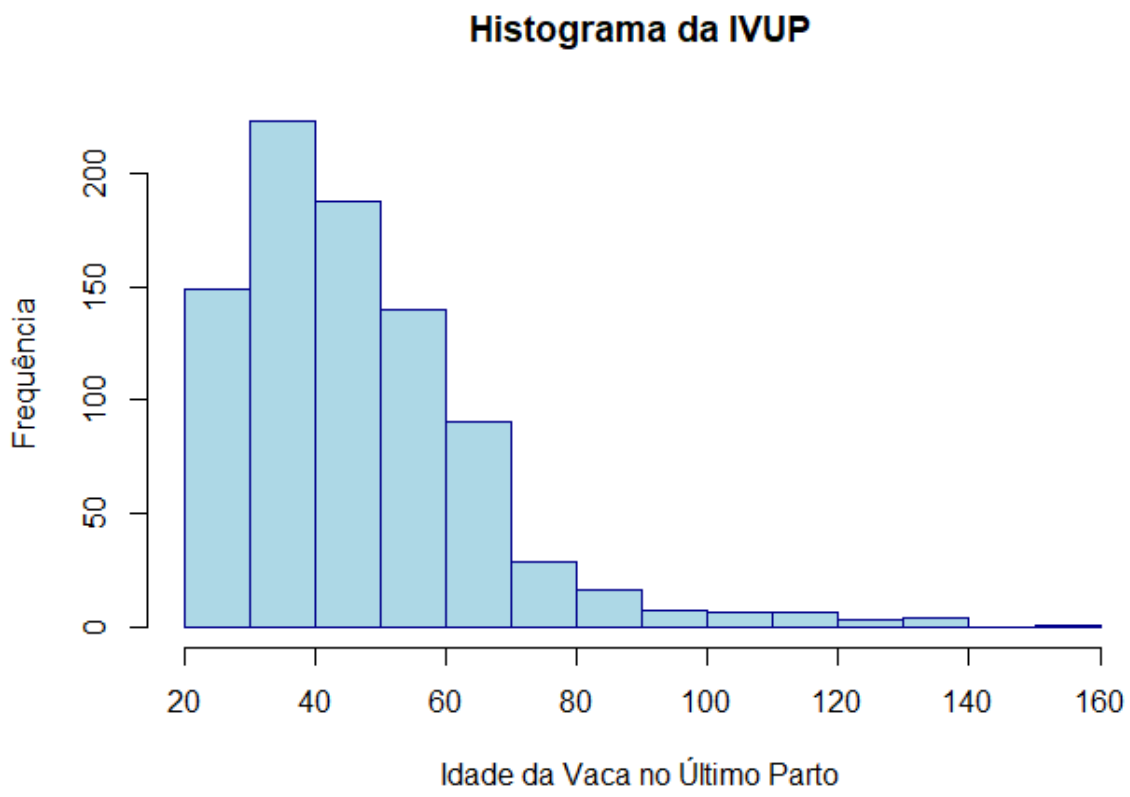

Figura 4.1: Histograma da IVUP

A variável IVUP assume valores entre 21.50 e 156.80 meses, sendo a mediana de 47.03 meses. Ou seja, 50% das vacas deste estudo têm a idade no último parto maior que 47.03 meses e 50% delas possuem a idade no último parto menor do que esses 47.03. Observando a Figura 4.1, nota-se que a maior concentração de idade das vacas no último parto está em torno de aproximadamente 30 a 60 meses e que poucas delas têm idade maior do que 80 meses em seu parto mais recente. Essa distribuição provavelmente está concentrada em

idades mais baixas do último parto por ser um rebanho jovem e muitas vacas ainda não chegaram no final da sua vida reprodutiva, como aponta a alta porcentagem de censuras.

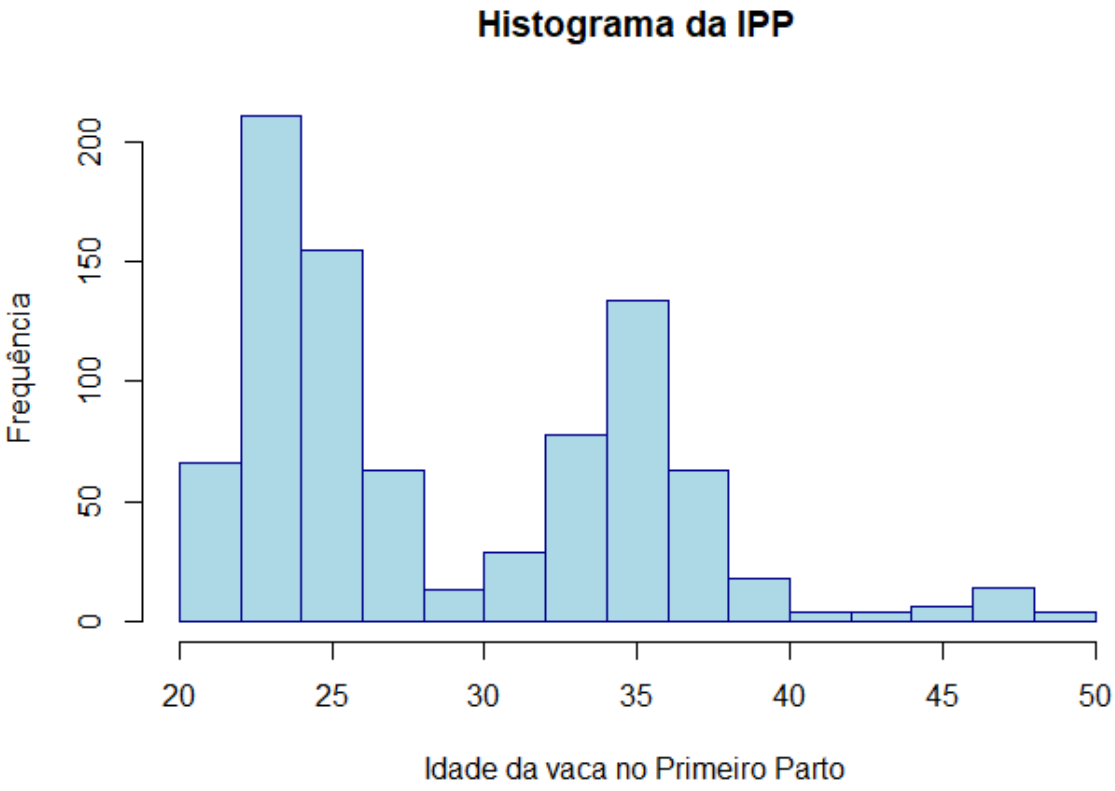

Figura 4.2: Histograma da IPP

Analisando a idade da vaca no primeiro parto, encontramos que o seu valor mínimo é de 21 meses e máximo de 49 meses, com uma mediana de 26 meses. A partir da Figura 4.2, nota-se que uma parte das vacas tinha idade aproximada entre 22 e 27 meses no seu primeiro parto e outra parte entre 32 e 37 meses. Existem poucos casos de vacas com mais de 40 meses em sua primeira gestação, o que evidencia um início de vida reprodutiva aceitável e não tão tardio.

Pelos valores e Figura 4.3 observados, nota-se que as maiores concentrações de quantidade de partos estão entre 1 e 3 partos por vaca, sendo 1 parto o caso com o maior número de ocorrências (310 vacas). Poucas vacas tiveram mais do que 6 partos, o que reforça a suposição de que provavelmente a base de dados se refere a um rebanho jovem.

As curvas de Kaplan-Meier para a Idade do Primeiro Parto da vaca (IPP) encontradas na Figura 4.4 foram divididas em quatro grupos, o primeiro deles é composto pelo primeiro quartil (valores menores que 24 meses) que contém 215 vacas. O segundo grupo engloba as 214 vacas que estão no segundo quartil (valores entre 24 e 27 meses) e assim por diante.

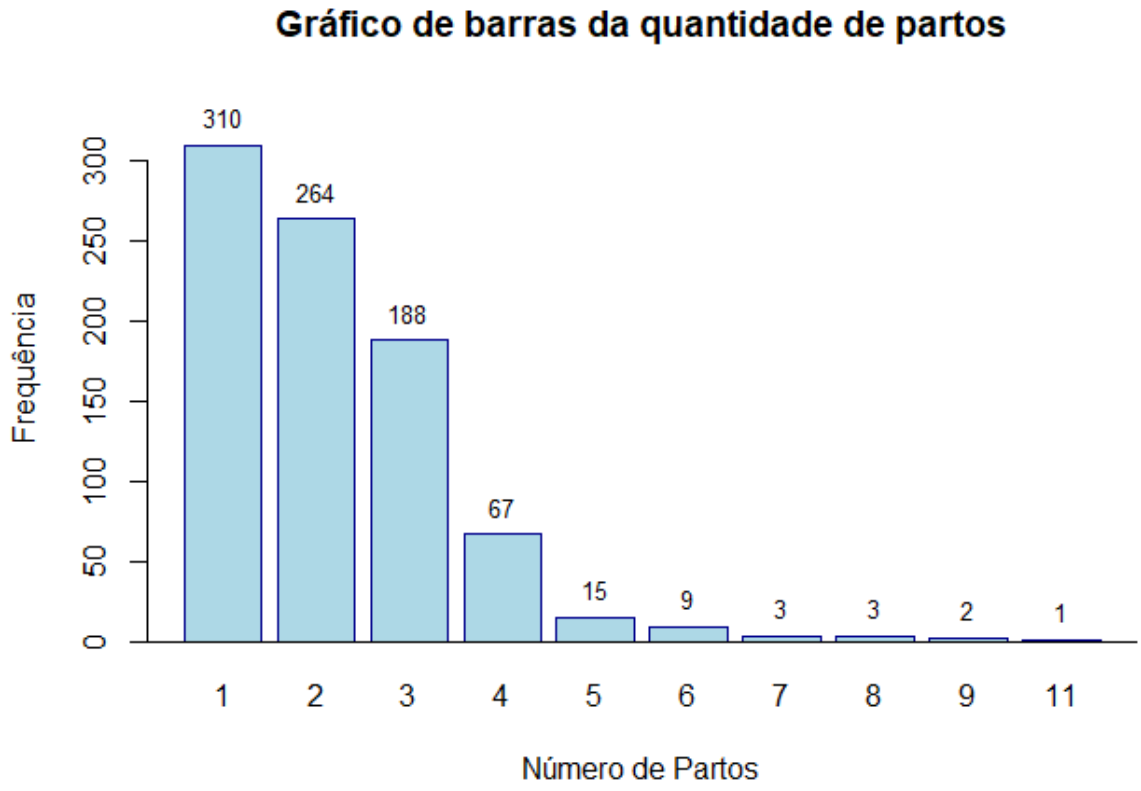

Figura 4.3: Gráfico de barras da quantidade de partos

O terceiro e quarto quartis contêm 214 vacas cada um e assumem os valores menores e maiores que 35 meses, respectivamente.

No início desse gráfico (até próximo de 50 meses), observa-se que vacas que tiveram seu primeiro parto mais velhas, ou seja, dos últimos grupos, sobrevivem mais quando comparadas às vacas que pariram mais cedo, que sobrevivem menos. Quando olhamos a segunda parte desse gráfico (aproximadamente a partir de 70 meses), esse comportamento se inverte. Nesse caso, vacas pertencentes ao primeiro e segundo grupos, ou seja, que tiveram o primeiro parto em idades mais novas, apresentam maior probabilidade de permanecer reprodutivamente ativas ao longo do tempo.

Além disso, as vacas pertencentes ao segundo grupo, que tiveram o primeiro parto entre 24 e 27 meses, são as que apresentaram, em média, a maior probabilidade de sobrevivência reprodutiva, indicando que esse é um intervalo de idade interessante para se ter o primeiro parto da vaca. Esses resultados sugerem que a precocidade reprodutiva está associada à longevidade na fertilidade, evidenciando que vacas que iniciam a vida reprodutiva mais cedo tendem a permanecer por mais tempo no rebanho e alcançar maior quantidade de partos ao longo de sua vida fértil.

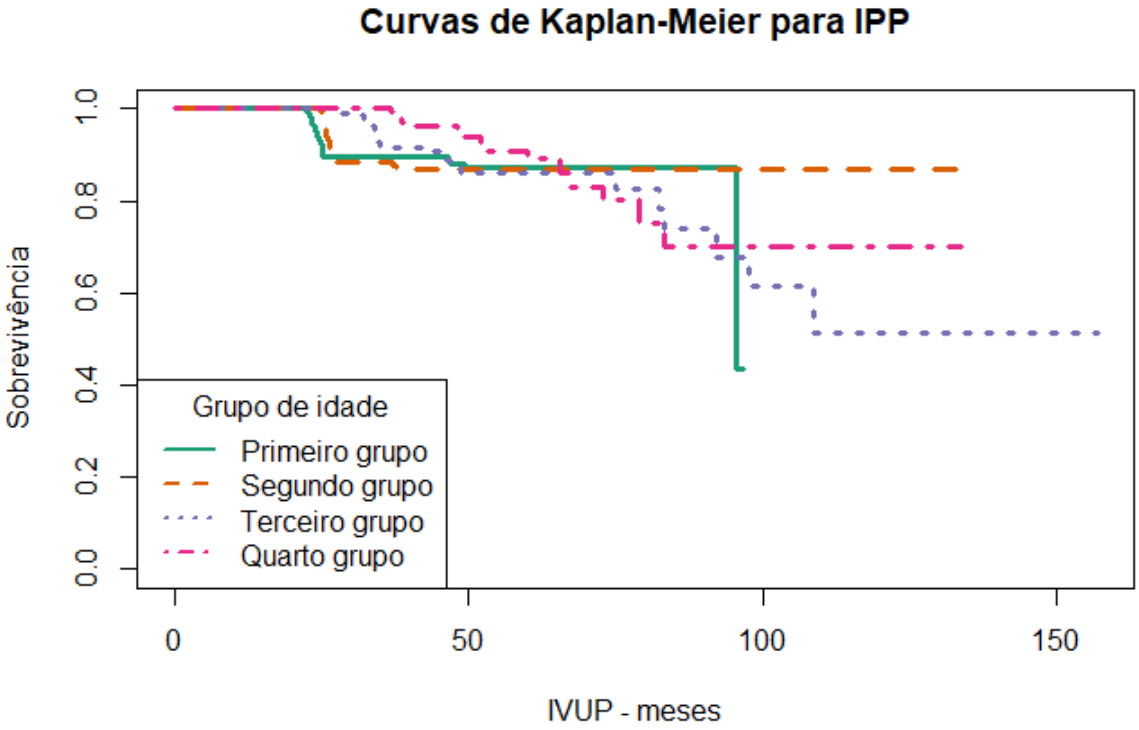

Figura 4.4: Curvas de Kaplan-Meier para a variável IPP

Para as curvas de Kaplan-Meier da variável quantidade de partos, foi utilizada a mesma lógica de divisão em quartis aplicada para IPP, sendo o primeiro quartil 1 parto, o segundo 2 partos e o terceiro 3 partos. O primeiro grupo possui 215 vacas e os outros três possuem 214 vacas cada um.

Os resultados encontrados na Figura 4.5 indicam que vacas com maior número de partos (quarto grupo) tendem a apresentar maior constância reprodutiva, mantendo ciclos gestacionais por mais tempo e atingindo o último parto em idades mais avançadas. Por outro lado, vacas com menor número de partos, agrupadas no primeiro grupo, tiveram menor longevidade na fertilidade, encerrando precocemente sua vida reprodutiva.

As curvas de Kaplan-Meier evidenciam que animais mais produtivos se mantêm reprodutivamente ativos por um tempo mais longo. Essa associação faz sentido, pois evidencia que vacas com maior número números de partos são mais reprodutivas. O contrário também é válido, pois vacas com menos partos tiveram quedas mais acentuadas e mais adiantadas nas curvas de sobrevivência, o que indica maior probabilidade de descarte em menor tempo. Essa situação sugere que vacas com menor número de partos são menos reprodutivas, portanto, serão descartadas precocemente, o que também faz sentido. Vale ressaltar que aqui estamos analisando apenas associação e não causa e efeito. Se a in-

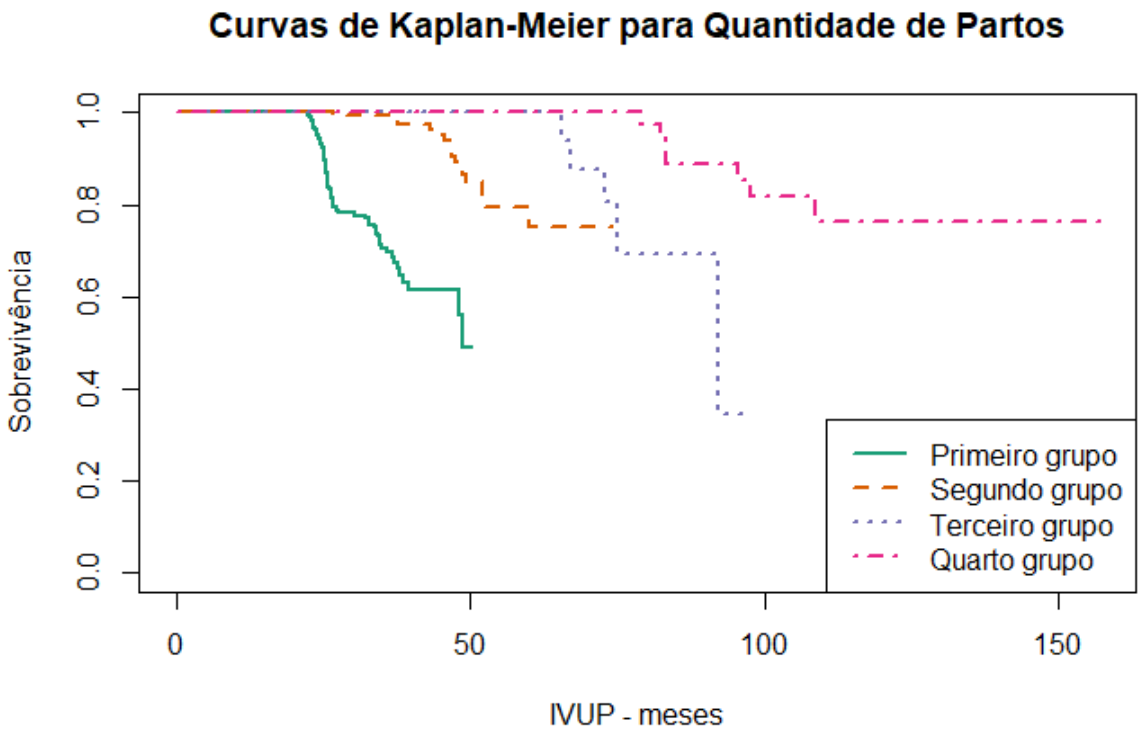

Figura 4.5: Curvas de Kaplan-Meier para a variável quantidade de partos

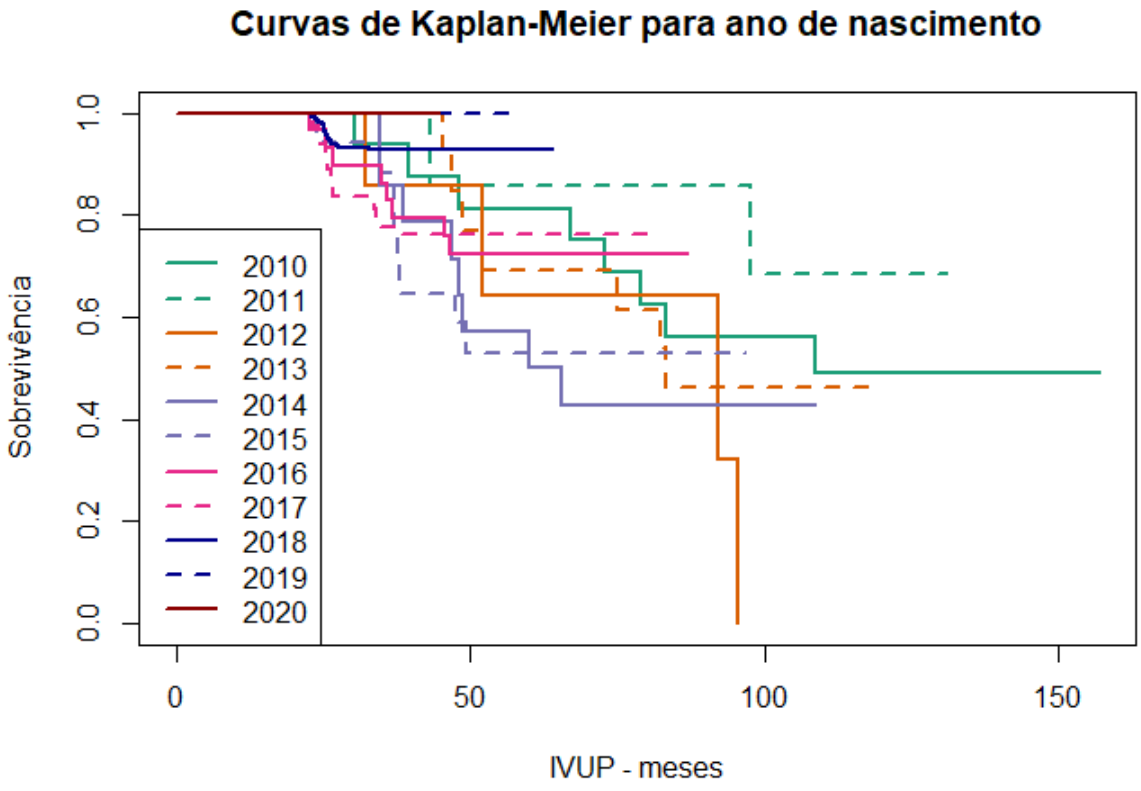

Figura 4.6: Curvas de Kaplan-Meier para a variável ano de nascimento

tenção fosse analisar causa e efeito, provavelmente, o fato de ser menos ou mais reprodutiva geneticamente causa o efeito de menos ou mais partos, respectivamente.

Observando as curvas de Kaplan-Meier para a variável ano de nascimento na Figura 4.6, percebe-se que vacas que são mais novas, ou seja, que têm anos de nascimento mais recentes como 2019 e 2020, por exemplo, possuem probabilidade maior de continuarem ativas no rebanho, ou seja, elas têm menor risco de sair da base. Vacas que nasceram no ano de 2012 apresentaram a menor probabilidade de permanecerem reprodutivamente ativas ao longo do tempo, seguidas de vacas que nasceram em 2011, 2013, 2014 e 2015, que possuem probabilidade semelhante. O comportamento das vacas que nasceram nos outros anos é bem parecido, não possuindo probabilidade alta de se manterem ativas por muito tempo.

Em resumo, isso indica que as vacas mais novas, ou seja, que nasceram em anos mais recentes, possuem menor risco de sair da base. Essa situação pode ser resultado de um possível melhoramento genético que já vem sendo aplicado no rebanho e vacas mais jovens são frutos de cruzamentos mais focados em longevidade reprodutiva.

A distribuição das vacas no ano de nascimento é apresentada na Tabela 4.1.

Ano	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Nº de vacas	16	7	7	13	14	17	29	80	387	150	137

Tabela 4.1: Número de vacas por ano de nascimento

4.2 Matriz de Relacionamento

A matriz de relacionamento (denotada por matriz A) é muito utilizada em análise genética, pois ela representa a covariância genética entre indivíduos de uma amostra. Primeiramente, com base na informação do pai e da mãe de cada vaca, calculamos a matriz de parentesco. Cada valor dessa matriz contém um coeficiente que varia de 0 a 0.5, esses valores refletem o grau de compartilhamento genético entre indivíduos, sendo:

- 0: nenhum parentesco;
- 0.125: relação entre avós e netos ou meio-irmãs;
- 0.25: entre pais e filhos ou irmãos completos (de mesmo pai e mesma mãe);
- 0.5: o animal com ele mesmo ou irmãos gêmeos idênticos.

Essa matriz é uma matriz simétrica, cujos valores da diagonal são 0.5 (relação entre a vaca e ela mesma) e os outros são referentes ao parentesco dos animais. Ela foi calculada no R a partir da biblioteca *kinship2*, o resultado fornecido deve ser multiplicado por 2 para obtermos a matriz de relacionamento, que assume valores de 0 a 1, sendo 1 na diagonal principal e os outros respectivos graus de relacionamento entre as vacas. Nessa situação, vacas que são irmãs completas ou relação entre pais e filhas, assumem o valor 0.5; vacas que são meio-irmãs ou avós e netas têm o valor 0.25.

Como o estudo é baseado em dados de pedigree provenientes de um rebanho controlado, foi possível identificar com precisão as relações familiares entre todos os indivíduos. Por isso, a matriz utilizada corresponde à matriz de relacionamento de pedigree, construída a partir das informações de pai e mãe de cada vaca.

4.3 Bases de Marcadores Genéticos e seu Mapa

Além da base mencionada inicialmente, outras duas são necessárias para a análise genética dos animais, são elas:

- Base de genótipos, que contém o número de alelos menor para os marcadores genéticos (SNPs) de cada vaca;

- Base do mapa, que contém todas as informações necessárias dos SNPs.

A primeira tinha 65.436 marcadores genéticos espalhados em 29 cromossomos. Após ter sido feito o controle de qualidade, realizando a exclusão dos marcadores com MAF $< 5\%$, sobrou um total de 33.129 marcadores genéticos espalhados em 29 cromossomos. Todos os marcadores da base inicial já estavam em equilíbrio de Hardy-Weinberg, portanto, nenhum outro foi excluído. A base de genótipos contém na primeira coluna o ID de cada vaca para as quais temos o genótipo dos marcadores e depois os 33.129 marcadores.

Já a segunda base citada contém o nome de cada um dos marcadores genéticos que são utilizados na análise, a posição em pares de base no cromossomo e o número do cromossomo associado.

A partir do ID de cada uma das vacas, foi feito um cruzamento com a base de genótipos dos marcadores para encontrarmos a informação genética das vacas do estudo. A base inicial das vacas tinha 862 vacas, porém 5 delas não estavam na base dos marcadores

genéticos e, por esse motivo, essas 5 vacas foram excluídas da análise e da matriz de relacionamento, resultando em uma base final com informação de 857 vacas distintas.

Capítulo 5

Resultados

Após todo o tratamento nas bases, construção das variáveis e definição das covariáveis relevantes seguindo a opinião de especialistas na área, um modelo basal (considerando apenas variáveis de controle) foi ajustado pelo R utilizando o pacote *coxme*. Após conversa com especialista e levando em consideração variáveis que foram significativas para o modelo, as variáveis de controle que foram colocadas no modelo são: IPP e ano de nascimento, ambas com coeficiente de regressão estimado negativo e p-valor significativo (menor do que 5%). Na Tabela 5.1 são apresentados os valores importantes do resultado do modelo.

Variável (efeito fixo)	Coeficiente de Regressão	P-Valor	Variância do efeito aleatório
IPP	-0.12362	8.99e-07	1.652822
Ano de nascimento	-0.35610	1.41e-07	

Tabela 5.1: Resultados do modelo base

Como estamos modelando o risco, a partir da Tabela 5.1 observamos que, quanto maior for o valor da variável associada ao coeficiente de regressão, menor será o risco da vaca ser precocemente inativa. Ou seja, no geral, vacas que tiveram seu primeiro parto com idades mais avançadas seriam mais longevas. Além disso, as vacas que nasceram mais recentemente, ou seja, as mais novas, também sobrevivem mais tempo reprodutivas. A relação IPP e reprodutividade estimada aqui talvez seja inversa ao que especialistas acreditam, mas é o comportamento refletido até a idade de 50 meses pelos gráficos de Kaplan-Meier mostrados anteriormente e, talvez, seja uma consequência da alta frequência de censuras na base.

Após esse modelo basal ter sido feito com apenas as variáveis de controle genético e de ambiente, cada marcador genético foi adicionado ao modelo, um a um, e foram

obtidos os seus p-valores, seus coeficientes e a variância dos efeitos aleatórios do modelo com essa inclusão. A partir dos valores de cada um dos modelos, podemos utilizar duas metodologias que são muito comuns e interessantes quando se trata da análise de dados em genética, são elas: gráfico de vulcão e gráfico de Manhattan. A seguir, são apresentadas breves descrições e análises.

O gráfico de vulcão é um diagrama de dispersão que possui no eixo x o coeficiente de regressão estimado para cada SNP em questão e no eixo y o $-log$ do p-valor. Esse tipo de gráfico nos permite identificar genes que são biologicamente significativos (BioStatsquid, 2025).

Em geral, nesse gráfico, quanto mais distante $\hat{\beta}$ está de 0, mais significativo é o marcador genético associado, pois, nesses casos, os p-valores relacionados são menores e, consequentemente, os valores de $-log_{10}$ deles serão maiores.

Comumente, é utilizado o limiar do p-valor de 5×10^{-8} para identificar marcadores genéticos que são significativos para o estudo (ScienceDirect, 2025b). Entretanto, nenhum dos SNPs estudados atingiu esse p-valor. Mas, a partir do gráfico de vulcão na Figura 5.1, podemos identificar 10 pontos mais discrepantes (em azul escuro) e que possuem os menores p-valores, consequentemente os maiores $-log_{10}$. Eles são apresentados na Tabela 5.2.

Nome do SNP	Cromossomo	P-Valor	Coeficiente	Variância do efeito aleatório
ARS-BFGL-NGS-34864	21	0,000069	-0.7444408	1.370593
ARS-BFGL-NGS-34254	5	0.000118	-0.7913658	1.435089
BOVINEHD1500017338	15	0.000337	1.0168280	1.397761
BOVINEHD0500006624	5	0.000349	-1.0738290	1.329523
ARS-BFGL-NGS-104172	14	0.000352	-0.6891247	1.483208
HAPMAP43483-BTA-117844	3	0.000433	0.7926819	1.590156
ARS-BFGL-NGS-112542	5	0.000452	-0.9625104	1.308138
BOVINEHD0900029838	9	0.000484	-0.9825532	1.404125
BOVINEHD0300029114	3	0.000493	1.8059340	1.378822
BTA-91367-NO-RS	15	0.000580	0.9590682	1.443261

Tabela 5.2: Os dez marcadores genéticos mais significativos ordenados de forma crescente em relação ao p-valor

Observando a Figura 5.2, percebemos que, conforme $\hat{\beta}$ aumenta em módulo, a variância dos efeitos aleatórios fica mais heterocedástica, variando mais entre os valores de 1.3 a 1.9. Além disso, notamos que, quando incluímos no modelo SNPs com valores de $\hat{\beta}$ próximos de zero, o valor da variância dos efeitos aleatórios fica entre 1.6 e 1.7, o que condiz com o valor da variância para o modelo basal (de 1.65). Isso nos mostra que SNPs mais associados ao fenótipo em estudo, quando incluídos no modelo, têm a

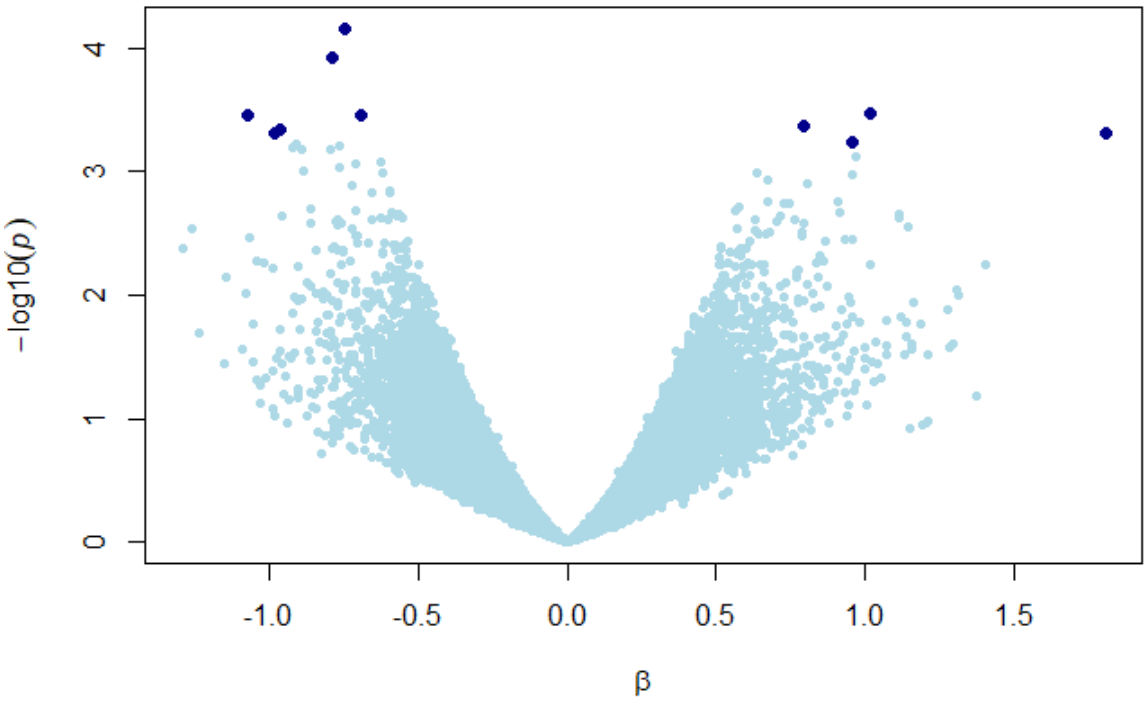

Figura 5.1: Gráfico vulcão

capacidade de reduzir a variância dos efeitos aleatórios genéticos porque explicam uma parte da variabilidade genética antes explicada pelos efeitos aleatórios e esse é o caso dos 10 SNPs mais relevantes destacados anteriormente, mas também a capacidade de aumentar a variância dos efeitos aleatórios. Nesse segundo caso, quando a inclusão de um SNP aumenta a variância dos efeitos aleatórios, seu efeito provavelmente deve reforçar uma tendência genética que é muito comum, e os indivíduos que não a seguem precisam de um efeito aleatório (específico) mais distante de zero para serem melhor modelados, aumentando a variância dos efeitos.

O gráfico de Manhattan possui no eixo y o $-log_{10}$ do p-valor e no eixo x a posição dos SNPs no DNA. Este gráfico representa o quanto a associação entre o $-log_{10}$ do p-valor e sua posição nos cromossomos é significativa estatisticamente (ScienceDirect, 2025b). Analogamente ao caso do gráfico de vulcão, no de Manhattan, quanto maior for o valor de um ponto no eixo y, mais significativa é aquela região no impacto do fenótipo em estudo.

A partir da Figura 5.3, podemos observar e confirmar as regiões e SNPs mais associados à IVUP, que estão de acordo com o que foi apresentado na Tabela 5.2. Os cromossomos

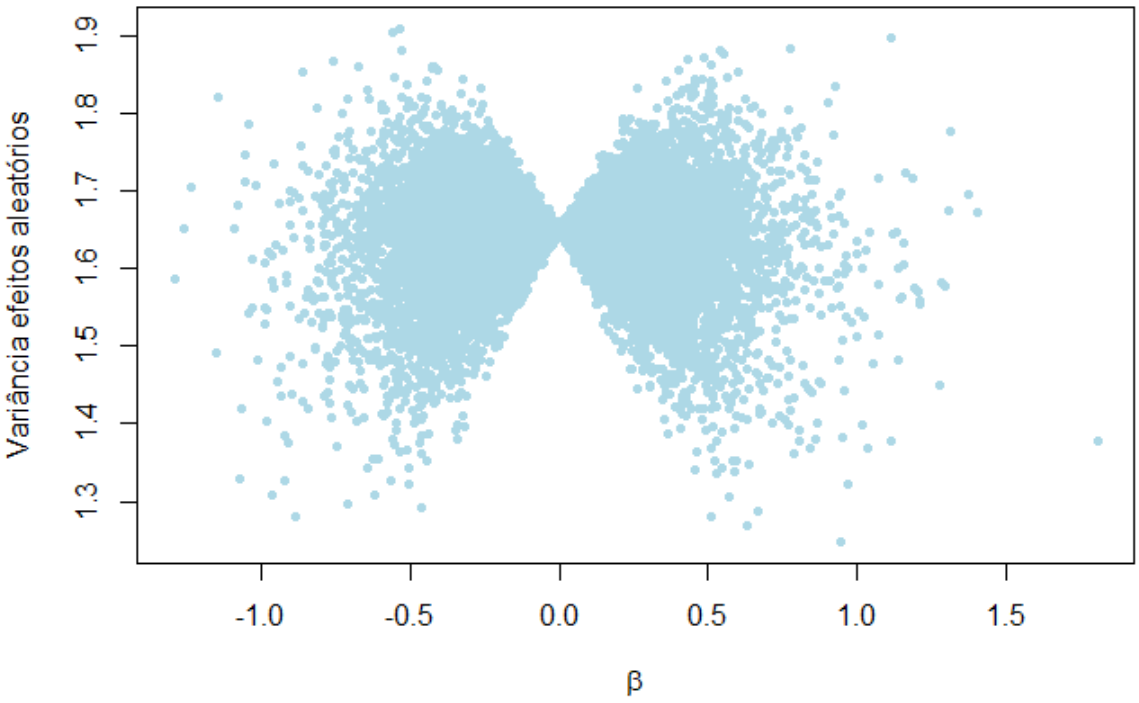

Figura 5.2: Gráfico de vulcão para a variância dos efeitos aleatórios com a inclusão de cada SNP no modelo

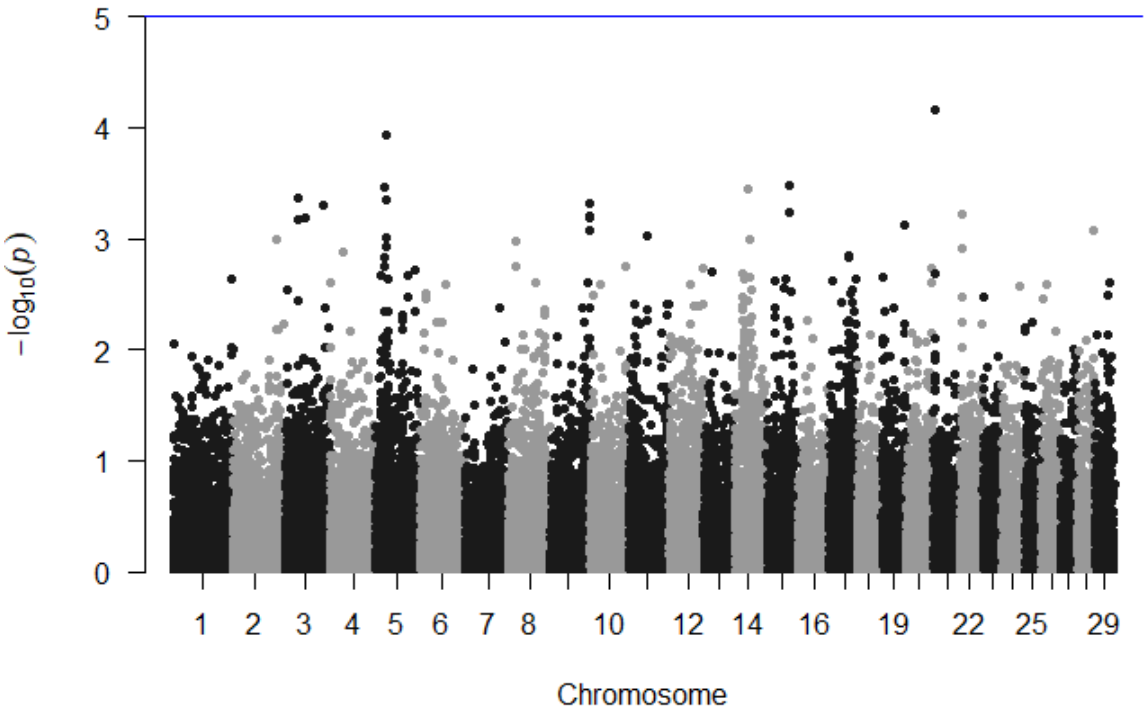

Figura 5.3: Gráfico Manhattan

5, 3 e 15 são os que possuem SNPs com maior associação à IVUP.

Portanto, apesar de não termos encontrado nenhum SNP com p-valor acima do limiar citado, encontramos alguns marcadores genéticos que se destacam mais em relação ao p-valor quando comparados a outros, o que pode representar possíveis candidatos a regiões associadas à longevidade bovina. Por isso, é interessante, do ponto de vista de melhoramento genético, buscar touros e vacas que tenham características específicas nesses marcadores genéticos para eles cruzarem e terem filhas que possam ter maior IVUP em média. Aqui reforçarmos novamente que a alta frequência de censura na base analisada pode ter impactado o cálculo do p-valor pois, muita censura causa uma incerteza maior em relação às estimativas pontuais dos coeficientes de regressão e, consequentemente, nas decisões a serem tomadas. Com mais incerteza, a tendência é sermos mais conservadores na decisão.

Para o cálculo da herdabilidade, encontramos a variabilidade total da variável IVUP na base, a qual resultou no valor de 342,2899. Após observarmos a variância dos efeitos aleatórios para o modelo basal (apenas com variáveis de controle) que foi de 1.652822, podemos calcular a herdabilidade como é citado pela literatura, fazendo a divisão entre a

variância dos efeitos aleatórios e a variabilidade total, da seguinte forma:

$$h^2 = \frac{\sigma_s^2}{\sigma^2}.$$

Com isso, encontramos uma herdabilidade amostral de 0,00483. Essa é uma herdabilidade bem baixa, mas como ela foi calculada apenas nas vacas filhas e não nos touros, talvez seja por esse motivo que a herdabilidade tenha sido tão baixa, considerando que na literatura a IVUP já foi registrada com herdabilidade acima de 0,30. Além disso, selecionamos uma amostra com um tamanho bem reduzido de vacas, o que também pode ter influenciado no cálculo da herdabilidade.

Capítulo 6

Conclusão e Discussão

Nesse trabalho, apresentamos modelos de análise de sobrevivência aplicados ao estudo da longevidade reprodutiva de vacas da raça Nelore utilizando o modelo de Cox com efeitos aleatórios, ajustado pelo pacote *coxme* do R. O modelo incorporou uma matriz de relacionamento construída a partir dos dados fornecidos para modelar a correlação genética entre indivíduos com grau de parentesco e também capturar efeitos genéticos não presentes em variáveis genéticas disponíveis na base.

A análise, realizada com uma amostra de 857 vacas, filhas de 47 touros diferentes, tem como objetivo estudar a Idade da Vaca ao Último Parto (IVUP), observando como ela é influenciada por características genéticas que podem ser potencializadas via melhoramento genético dos animais. Diante disso, as seguintes variáveis de controle foram utilizadas no modelo: IPP e ano de nascimento. Além dessas, todos os 33.129 marcadores genéticos foram adicionados no modelo, um a um, para entendermos e encontrarmos possíveis SNPs que fossem significativos para a variável em estudo.

Apesar de nenhum marcador genético ter alcançado o limiar de associação geralmente utilizado na literatura, identificamos quais foram os SNPs mais significativos e os seus cromossomos associados, que podem ser regiões genômicas relacionadas à longevidade reprodutiva e estudadas com mais detalhes pelos geneticistas.

A partir desses resultados, podemos incentivar a busca por touros e vacas que tenham características interessantes nos SNPs identificados como mais significativos. Com isso, podemos aplicar o melhoramento genético e cruzar esses animais entre si, o que pode resultar na fecundação de vacas que, em média, poderão ter maior IVUP.

Além dos pontos genéticos a serem observados para uma maior IVUP, a significância das variáveis de controle reforça a importância de fatores ambientais e de manejo ao longo

da vida reprodutiva da vaca, para que ela alcance a maior longevidade em sua vida fértil.

Como na base analisada a frequência de censuras é muito alta, provavelmente conseguiríamos resultados mais precisos e evidentes se o rebanho continuasse a ser monitorado e analisado para, com mais certeza, termos a IVUP de um número bem maior de vacas.

No entanto, os resultados apresentados neste trabalho sugerem que o melhoramento genético voltado para maior longevidade reprodutiva é possível e esperado. Assim como a identificação de touros e vacas portadores de características nos SNPs evidenciados, pois isso pode contribuir para aumentar a vida útil produtiva do rebanho ao longo das gerações.

Por fim, recomenda-se que estudos futuros ampliem o número de animais e incorporem variáveis ambientais mais detalhadas, ajustando modelos que não assumam riscos proporcionais entre os indivíduos, além de realizarem investigações mais profundas sobre as regiões genéticas identificadas neste trabalho. Isso pode ajudar a esclarecer os mecanismos biológicos envolvidos e incentivar programas de seleção. Assim, os resultados apresentados contribuem para o entendimento da longevidade reprodutiva e fornecem bases para iniciativas de manejo e melhoramento genético mais eficientes nos rebanhos.

Referências Bibliográficas

Aidar de Queiroz, S. (2025). Herdabilidade. Material de aula (Genética Quantitativa I). Acesso em 23 jun. 2025 via https://www.fcav.unesp.br/Home/departamentos/zootecnia/SANDRAAIDARDEQUEIROZ/herdabilidade.pdf.

BioStatsquid (2025). How to interpret a volcano plot. Acessado em: 13 nov. 2025.

Breslow, N. E. (1972). Discussion of professor Cox's paper. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **34**, 216.

Caetano, S. L. (2011). *Estudo da idade da vaca ao último parto para avaliar longevidade em rebanhos da raça nelore por análise de sobrevivência*. Tese de doutorado, Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências Agrárias e Veterinárias, Campus de Jaboticabal.

Cerqueira, A., Zuanetti, D. A. e Soler, J. P. (2024). Minicurso: Análise estatística de dados genômicos. Minicurso apresentado no 68ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria, realizada na ESALQ/USP em Piracicaba.

Colosimo, E. A. e Giolo, S. R. (2021). *Análise de sobrevivência aplicada*. Editora Blucher.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**(2), 187–202.

Freitas, L. A. C. d. (2025). Testes de Wald para restrições em parâmetros — msc page. https://lineu96.github.io/msc_page/2_wald_test.html#teste_wald_para_um_%C3%BAnico_par%C3%A2metro. Acessado em: 10 de outubro de 2025.

Hill, W. G. (2013). Heritability. Em S. W. Maloy e K. A. Hughes, editors, *Brenner's Encyclopedia of Genetics (Second Edition)*, páginas 695–699. Academic Press. ISBN 9780123749840. Capítulo sobre hereditariedade.

Kalbfleisch, J. D. e Prentice, R. L. (2002). *The statistical analysis of failure time data*. John Wiley & Sons.

Kassambara, A. (2025). Cox proportional hazards model. https://www.sthda.com/english/wiki/cox-proportional-hazards-model. Acesso em: 1 jun. 2025.

Magalhães, L. (s.d.). Dna: o que é e sua estrutura explicada. https://www.todamateria.com.br/dna/.

Matuda, N. d. S. (2005). *Fragilidade gama e variância robusta: extensões do modelo semiparamétrico de Cox*. Tese de doutorado, Universidade de São Paulo.

Minguillo, M. (2016). *Introdução aos modelos de fragilidade: Uma maneira de analisar dados correlacionados de sobrevivência*. Trabalho de conclusão de curso (graduação), Universidade Federal do Rio Grande do Sul, Porto Alegre.

Ministério da Agricultura e Pecuária (2024). Carne bovina é um dos principais produtos pecuários nas exportações brasileiras. Acesso em: 14 abr. 2025.

Nayara, J. (2019). Gwas: Estudo de associação entre fenótipos e genótipos. https://www.laborgene.com.br/gwas-estudo-de-associacao-entre-fenotipos-e-genotipos/. Acesso em: 12 jun. 2025.

Pinheiro, N. M. (2022). Entenda o que é análise de sobrevivência e como utilizar essa técnica em projetos de data science. https://medium.com/data-hackers/entenda-o-que-%C3%A9-an%C3%A1lise-de-sobreviv%C3%AAncia-e-como-utilizar-essa-t%C3%A9cnica-em-projetos-de-data-science-82bfaea8f546. Publicação em Medium/Data Hackers, acessado em 23 de junho de 2025.

Santos, V. S. d. (2024). O que é genótipo? https://brasilescola.uol.com.br/o-que-e/biologia/o-que-e-genotipo.htm. Acessado em: 12 nov. 2025.

Santos, V. S. d. (2025). O que é genoma. https://brasilescola.uol.com.br/o-what-is/genoma.htm. Acessado em 23 de junho de 2025.

ScienceDirect (2025a). Kaplan–meier method. https://www.sciencedirect.com/topics/medicine-and-dentistry/kaplan-meier-method. Acesso em: 17 dez. 2025.

ScienceDirect (2025b). Manhattan Plot. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/manhattan-plot. Acessado em: 13 nov. 2025.

Therneau, T. (2024a). *Mixed Effects Cox Models*. Mayo Clinic. Disponível como documentação do pacote `coxme`.

Therneau, T. M. (2024b). *coxme: Mixed Effects Cox Models*. CRAN. Pacote R para modelos de Cox com efeitos fixos e aleatórios.

Therneau, T. M. (2024c). *coxme: Mixed Effects Cox Models*. CRAN. Manual de referência do pacote; publicado em 23 de agosto de 2024.

Therneau, T. M. e Grambsch, P. M. (2000). Modeling survival data: extending the Cox model.

Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T. e Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, **1**(1), 59.

Vaupel, J. W., Manton, K. G. e Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**(3), 439–454.

Zuanetti, D. A. (2025). Unidade 3: Lei e equilíbrio de hardy-weinberg. Notas de aula da disciplina Tópicos em Estatística Genética - DEs (UFSCar).

Apêndice A

Tratamento e Descritiva do Banco de Dados

```
    library(foreign)

# Baixando a base de dados
dados
dim(dados)
head(dados)
paste("Dimensão do banco de dados inicialmente: ", dim(dados))

# Pegando apenas as colunas que serão necessárias
dados <- dados[ , c("G_ANIM", "G_PAI", "G_MAE", "DT_NASC", "IPP", "GCIPP",
"DT_PARTO")]
dados



# Agora, vamos analisar a base sem duplicadas de touro, a partir dela
# usaremos a informação "G_PAI"

# Quantos touros têm na base?
num_touros <- length(unique(dados$G_PAI))
paste("Número de touros na base: ", num_touros)
```

```
dados_touros <- unique(dados$G_PAI)
dados_touros

# Conferindo se pegou os dados corretamente
length(dados_touros)
sum(duplicated(dados_touros))

# Primeiramente, vamos fixar uma semente para garantir reprodutibilidade
set.seed(123)

# A partir dessa base, vamos sortear aleatoriamente 50 touros (sem reposição)
amostra <- sample(dados_touros, 50, replace = FALSE)
amostra



# Agora, vamos voltar à base original e retirar a duplicidade por vaca,
# além de deixar apenas o parto mais recente de cada uma

# Ordenando por data de parto (mais recente para mais antigo)
dados_dt_ord <- dados[order(dados$DT_PARTO, decreasing = TRUE), ]
dados_dt_ord

# Coonferindo se pegou o parto mais recente
summary(dados$DT_PARTO)
dados[dados$G_ANIM == 44067, ]
dados_dt_ord[dados_dt_ord$G_ANIM == 44067, ]
dados[dados$G_ANIM == 44068, ]
dados_dt_ord[dados_dt_ord$G_ANIM == 44068, ]

# Agora deixa apenas as primeiras ocorrências de cada vaca
dados_unicos_vacas <- dados_dt_ord[!duplicated(dados_dt_ord$G_ANIM), ]
dados_unicos_vacas
paste("Dimensão do banco de dados após remoção das vacas duplicadas: ",
```

```
dim(dados_unicos_vacas))


# Conferindo se pegou os dados corretamente
sum(duplicated(dados_unicos_vacas))
dados_unicos_vacas[dados_unicos_vacas$G_ANIM == 44067, ]
dados_unicos_vacas[dados_unicos_vacas$G_ANIM == 44068, ]


# A partir dessa base e da amostra (dos touros), vamos juntá-las por G_PAI
str(amostra)
amostra <- as.data.frame(amostra)
colnames(amostra) <- "G_PAI"
dados_juntos <- merge(amostra, dados_unicos_vacas, by = "G_PAI")
dados_juntos
head(dados_juntos)
table(dados_juntos$G_PAI)
dim(dados_juntos)

# Conferindo a junção
dados[dados$G_PAI == 26097, ]
dados_juntos[dados_juntos$G_PAI == 26097, ]



# Juntando agora com a base por GCIPP
# Tirar a duplicidade pela GCIPP
dados_unicos_GCIPP <- dados_dt_ord[!duplicated(dados_dt_ord$GCIPP), ]
dados_unicos_GCIPP
paste("Dimensão do banco de dados após remoção das GCIPP's duplicadas: ",
dim(dados_unicos_GCIPP))

# Conferindo se tirou as vacas com GCIPP duplicadas
```

```
sum(duplicated(dados_unicos_GCIPP$GCIPP))


# Pegando apenas as colunas que serão necessárias
dados_unicos_GCIPP <- dados_unicos_GCIPP[ , c( "GCIPP", "DT_PARTO")]
dados_unicos_GCIPP
colnames(dados_unicos_GCIPP)[2] <- "DT_PARTO_GCIPP"

# Juntando as bases a partir do GCIPP
dados_juntos_final <- merge(dados_juntos, dados_unicos_GCIPP, by = "GCIPP")
head(dados_juntos_final)
dim(dados_juntos_final)

# Conferindo se uniu corretamente
dados_unicos_GCIPP[dados_unicos_GCIPP$GCIPP == 36, ]
dados_juntos_final[dados_juntos_final$GCIPP == 36, ]
dados_juntos[dados_juntos$G_ANIM == 110230, ]

dados_unicos_GCIPP[dados_unicos_GCIPP$GCIPP == 38, ]
dados_juntos_final[dados_juntos_final$GCIPP == 38, ]
dados_juntos[dados_juntos$G_ANIM == 130077, ]


# Retirando os NA's
dados_juntos_final <- dados_juntos_final[!is.na(dados_juntos_final$GCIPP), ]
dim(dados_juntos_final)


# Primeira análise
summary(dados_juntos_final$GCIPP)
summary(dados_juntos_final)
```

```
# Criando a variável IVUP para cada vaca
dados_juntos_final$IVUP <- dados_juntos_final$DT_PARTO -
dados_juntos_final$DT_NASC
dados_juntos_final$IVUP


# A saída veio em dias, então transformo em meses (apenas dividindo por 30)
dados_juntos_final$IVUP <- dados_juntos_final$IVUP / 30
dados_juntos_final$IVUP

# Transformando em numérico
dados_juntos_final$IVUP <- as.numeric(dados_juntos_final$IVUP)
summary(dados_juntos_final$IVUP)

head(dados_juntos_final)


# Criando a variável que indica a diferença entre a data do último parto
para cada vaca
dados_juntos_final$DIFERENÇA <- dados_juntos_final$DT_PARTO_GCIPP -
dados_juntos_final$DT_PARTO
dados_juntos_final$DIFERENÇA


# A saída veio em dias, então transformo em meses (apenas dividindo por 30)
dados_juntos_final$DIFERENÇA <- dados_juntos_final$DIFERENÇA / 30
dados_juntos_final$DIFERENÇA
summary(dados_juntos_final$DIFERENÇA)

# Transformando em numérico
dados_juntos_final$DIFERENÇA <- as.numeric(dados_juntos_final$DIFERENÇA)
str(dados_juntos_final$DIFERENÇA)
```

```
paste("Número de vacas que tiveram uma diferença menor que 36 meses:",
sum(dados_juntos_final$DIFERENÇA < 36, na.rm = TRUE))
paste("Número de vacas que tiveram uma diferença maior que 36 meses:",
sum(dados_juntos_final$DIFERENÇA > 36, na.rm = TRUE))


# Calculando a censura
dados_juntos_final$Censura <- ifelse(dados_juntos_final$DIFERENÇA < 36, 0, 1)

head(dados_juntos_final)


# Adicionando uma nova coluna com a quantidade de vezes que cada vaca pariu
partos <- table(dados[, 1])
partos
dim(partos)

dados_juntos_final$QUANT_PARTOS <- partos[match(dados_juntos_final$G_ANIM,
names(partos))]
dados_juntos_final$QUANT_PARTOS


# Conferindo
dados[dados$G_ANIM == 110230, ]
dados[dados$G_ANIM == 127240, ]


str(dados_juntos_final)
summary(dados_juntos_final)


# Criando a variável estação de parto
```

```
dados_juntos_final$EST_PARTO <- ifelse(as.POSIXlt(dados_juntos_final$DT_PARTO)$mon
+ 1 >= 10 | as.POSIXlt(dados_juntos_final$DT_PARTO)$mon + 1 <= 3, 1, 0)
# 1 é águas e 0 secas

# Pegando apenas o ano de nascimento da vaca
dados_juntos_final$ANO_NASC <- as.POSIXlt(dados_juntos_final$DT_NASC)$year+ 1900



dados_juntos_final
summary(dados_juntos_final)
dim(dados_juntos_final)

# Informações das variáveis
table(dados_juntos_final$Censura)
table(dados_juntos_final$QUANT_PARTOS)
table(dados_juntos_final$EST_PARTO)
table(dados_final2$ANO_NASC)



# Transforma a base em data frame para colocar no pacote
dados_final2 <- as.data.frame(dados_final_final)
dados_final2
dim(dados_final2)

# Descritiva IVUP
hist(dados_final2$IVUP, col = "lightblue", ylab = "Frequência",
xlab = "Idade da Vaca ao Último Parto", main = "Histograma da IVUP")
summary(dados_final2$IVUP)

# Descritiva IPP
hist(dados_final2$IPP, col = "lightblue", ylab = "Frequência",
xlab = "Idade da Vaca ao Último Parto", main = "Histograma da IPP")
summary(dados_final2$IPP)
```

```
# Descritiva Quantidade de Partos
barplot(table(dados_final2$QUANT_PARTOS), col = "lightblue", ylab = "Frequência",
xlab = "Número de Partos", main = "Gráfico de barras da quantidade de partos")
summary(dados_final2$IPP)


# Baixando pacotes necessários
library(survival)
library(survminer)
library(dplyr)

# Para IPP (dividindo nos quartis):
dados_final2$grupo_idade <- ntile(dados_final2$IPP, 4)
dados_final2$grupo_idade <- factor(dados_final2$grupo_idade,
labels = c("Primeiro grupo", "Segundo grupo", "Terceiro grupo", "Quarto grupo"))
table(dados_final2$grupo_idade)
quantile(dados_final2$IPP, probs = c(0, 0.25, 0.5, 0.75, 1))


km <- survfit(Surv(IVUP, Censura) ~ grupo_idade, data = dados_final2)
summary(km)


plot(km,
     col = c("#1B9E77", "#D95F02", "#7570B3", "#E7298A"),
     lwd = 3, lty = 1:4,
     xlab = "IVUP - meses",
     ylab = "Sobrevivência",
     main = "Curvas de Kaplan-Meier para IPP")

# Adiciona a legenda no canto inferior direito
legend("bottomleft", legend = levels(dados_final2$grupo_idade),
```

```
       col = c("#1B9E77", "#D95F02", "#7570B3", "#E7298A"),
       lwd = 2,
       lty = 1:4,
       title = "Grupo de idade")



# Para quantidade de partos
dados_final2$grupo_quant_partos <- ntile(dados_final2$QUANT_PARTOS, 4)
dados_final2$grupo_quant_partos <- factor(dados_final2$grupo_quant_partos,
labels = c("Primeiro grupo", "Segundo grupo", "Terceiro grupo", "Quarto grupo"))

km <- survfit(Surv(IVUP, Censura) ~ grupo_quant_partos, data = dados_final2)
summary(km)

plot(km, col = c("#1B9E77", "#D95F02", "#7570B3", "#E7298A"), lwd = 2, lty = 1:4,
xlab = "IVUP - meses", ylab = "Sobrevivência",
     main = "Curvas de Kaplan-Meier para Quantidade de Partos")

legend("bottomright", legend = c("Primeiro grupo", "Segundo grupo",
"Terceiro grupo", "Quarto grupo"), col = c("#1B9E77", "#D95F02",
"#7570B3", "#E7298A"), lty = 1:4, lwd = 2)
table(dados_final2$grupo_quant_partos)
quantile(dados_final2$QUANT_PARTOS, probs = c(0, 0.25, 0.5, 0.75, 1))



# Para ano de nascimento:
km <- survfit(Surv(IVUP, Censura) ~ ANO_NASC, data = dados_final2)
summary(km)



cores <- c("#1B9E77", "#D95F02", "#7570B3", "#E7298A", "darkblue","darkred")
```

```
# Linhas: 1 = contínua, 2 = tracejada
linhas <- rep(c(1, 2), 5)

# Vetor de cores repetindo cada cor duas vezes (anos vizinhos)
colores <- rep(cores, each = 2)

plot(km, col = colores, lwd = 2, lty = linhas,
     xlab = "IVUP - meses",
     ylab = "Sobrevivência",
     main = "Curvas de Kaplan-Meier para ano de nascimento")

legend("bottomleft", legend = c("2010", "2011", "2012", "2013", "2014", "2015",
"2016", "2017", "2018", "2019", "2020"), col = colores, lwd = 2, lty = linhas)
table(dados_final2$ANO_NASC)
```

Apêndice B

Matriz de Parentesco

```
# Matriz de relacionamento
library(kinship2)

# Pegando apenas as colunas necessárias para calcular a matriz de parentesco
info_parentes <- dados_juntos_final[ , c("G_ANIM", "G_PAI", "G_MAE")]
dim(info_parentes)
info_parentes <- as.matrix(info_parentes)

# Tirando as duplicatas de pai e mãe
pais <- unique(info_parentes[ , 2])
length(pais)
maes <- unique(info_parentes[ , 3])
length(maes)

# Criando a matriz com 0's e 1's, primeiramente atribuindo 0 aos avós das vacas
familia <- matrix(0, nrow = length(pais) + length(maes), ncol = 3)

# Atribui os animais das primeiras linhas com os ID's dos pais
familia[1:length(pais), 1] <- pais

# Atribui os animais da segunda parte de linhas com os ID's das mães
familia[(length(pais) + 1):(length(pais) + length(maes))] <- maes
```

```
# Juntando agora com as informações dos parentes
familia <- rbind(familia, info_parentes)

# Ordena por G_ANIM
ordena <- order(familia[, 1])
familia <- familia[ordena, ]

# Criando a matriz x para usá-la no cálculo da matriz de parentesco
x <- matrix(0, nrow(familia), ncol(familia))
x[, 1] <- familia[, 1]
x[, 2] <- familia[, 2]
x[, 3] <- familia[, 3]

# A primeira linha é zero, então tira
x <- x[-1, ]
x

# Calculando a matriz de parentesco
matriz_parentesco <- kinship(id = x[, 1], dadid = x[, 2], momid = x[, 3])
summary(c(matriz_parentesco))
dim(matriz_parentesco)

# Ordenando as vacas da maneira que estão na base final
ordem_filhos <- info_parentes[, 1]
# Pegando as linhas das filhas
fica <- which(x[, 1] %in% ordem_filhos)
length(fica)
mat_paren_filhas <- matriz_parentesco[fica, fica]
dim(mat_paren_filhas)
table(mat_paren_filhas)

# Ordena a base dados_juntos_final pela mesma ordem em que as filhas aparecem
na matriz de parentesco
```

```
ordena2 <- order(dados_juntos_final$G_ANIM)
dados_final_final <- as.matrix(dados_juntos_final[ordena2, c(1, 2, 3, 4, 6, 7, 8,
11, 12, 13, 14, 15, 16)])
dim(dados_final_final)
head(dados_final_final)

# Multiplicando por 2
matriz_parentesco <- 2 * mat_paren_filhas
matriz_parentesco
table(matriz_parentesco)
dim(matriz_parentesco)

# Conferindo alguns casos
matriz_parentesco["113859", "129208"]
matriz_parentesco["113859", "113859"]
matriz_parentesco["116097", "83055"]
```

Apêndice C

Base de Genótipos e Marcadores

```
# Juntando com a base de genótipos e mapa
mapa
genot

dim(mapa)
head(mapa)

dim(genot)
head(genot)
str(genot)

# Identificando as vacas que estavam apenas na base inicial e removendo da base
id <- genot[, 1]
length(id)

dados_final_final[ , 3]

setdiff(dados_final_final[ , 3], id)

tirar <- which(dados_final_final[ , 3] %in% c(115024, 115278, 115513, 115622,
139827))
tirar
```

```
dados_final_final <- dados_final_final[-tirar, ]
matriz_parentesco <- matriz_parentesco[-tirar, -tirar]
dim(dados_final_final)
dim(matriz_parentesco)


# Cálculo da variância total da IVUP
var(dados_final_final[, "IVUP"])
```

Apêndice D

Ajuste do Modelo Basal

```
    # Aplicando os dados no coxme
library(coxme)
help(coxme)

# Cálculo do modelo base
modelo <- coxme(Surv(IVUP, Censura) ~ 1 + IPP + (1|G_ANIM) + ANO_NASC,
data = dados_final2, varlist = coxmeMlist(matriz_parentesco))
summary(modelo)



# Agora aplicando o modelo para cada um dos marcadores genéticos um a um
# Comando para rodar mais rápido
library(compiler)
enableJIT(3)

pvalores <- NULL
var_ef_aleat <- NULL
coef_gen <- NULL
cont <- 1

for (i in 2:ncol(genot))
{
  dados_final2$genotipo_atual <- genot[, i]
```

```
  modelo_gen <- coxme(Surv(IVUP, Censura) ~ 1 + IPP + (1|G_ANIM) + ANO_NASC
  + genotipo_atual,
                      data = dados_final2,
                      varlist = coxmeMlist(matriz_parentesco)
  )

  s <- summary(modelo_gen)
  # Guardando o p-valor
  pvalores[cont] <- s$coef[3, 5]

  # Guardando a variância e o coeficiente
  var_ef_aleat[cont] <- as.numeric(VarCorr(modelo_gen)$G_ANIM)
  coef_gen[cont] <- modelo_gen[[1]][3]

  print(paste("Modelo", cont))
  cont <- cont + 1
}

# Exibindo os resultados:
pvalores
var_ef_aleat
coef_gen
```

Apêndice E

GWAS

```
    # Juntando todos
pvalores <- c(pvalores_6500, pvalores_13001, pvalores_13002_20001, p
valores_20002_28001, pvalores_28002_33130)
head(pvalores)
length(pvalores)
str(pvalores)
pvalores <- as.numeric(pvalores)

coeficientes <- c(coef_gen_6500, coef_gen_13001,
coef_gen_13002_20001, coef_gen_20002_28001, coef_gen_28002_33130)
head(coeficientes)
length(coeficientes)
str(coeficientes)
coeficientes <- as.numeric(coeficientes)

var_ef_aleat <- c(var_ef_aleat_6500, var_ef_aleat_13001,
var_ef_aleat_13002_20001, var_ef_aleat_20002_28001, var_ef_aleat_28002_33130)
head(var_ef_aleat)
length(var_ef_aleat)
str(var_ef_aleat)
var_ef_aleat <- as.numeric(var_ef_aleat)

## Gráfico Vulcão
```

```
# Ele é um diagrama de dispersão em que os valores do eixo x são os coeficientes
# estimados e os valores do eixo y representam o logaritmo do p-valor.
plot(coeficientes, -log10(pvalores), main = "", pch = 20, xlab = expression(beta),
ylab = expression(-log10(italic(p))), col = "lightblue")
# Pegando os com menor p-valor
top10 <- order(pvalores)[1:10]

# Pegando as informações dos top 10 mais discrepantes
pvalores[c(26755, 7520, 21678, 7480, 20330, 4400, 7547, 14705, 5310, 21677)]
coeficientes[c(26755, 7520, 21678, 7480, 20330, 4400, 7547, 14705, 5310, 21677)]
var_ef_aleat[c(26755, 7520, 21678, 7480, 20330, 4400, 7547, 14705, 5310, 21677)]
mapa[c(26755, 7520, 21678, 7480, 20330, 4400, 7547, 14705, 5310, 21677), ]


# Destacar os pontos discrepantes
points(coeficientes[top10], -log10(pvalores[top10]), col = "darkblue", pch = 19)


## Gráfico Vulcão para variância dos efeitos aleatórios
# Ele é um diagrama de dispersão em que os valores do eixo x são os coeficientes
# estimados e os valores do eixo y representam o logaritmo do p-valor.
plot(coeficientes, var_ef_aleat, main = "", pch = 20, xlab = expression(beta),
ylab = "Variância efeitos aleatórios", col = "lightblue")



## Gráfico Manhattan
# CHR - número do cromossomo do SNP
# BP - Posição do cromossomo
# P - p-valor obtido no GWAS
# SNP - nome do SNP

#install.packages("qqman")
```

```
library("qqman")
mapa
head(mapa)
str(mapa)
# Atribuindo um número para cada SNP
mapa$indice <- 1:nrow(mapa)
mapa

datt <- data.frame(SNP = mapa$snp_id, CHR = as.numeric(mapa[mapa$indice, 2]),
BP = as.numeric(mapa[mapa$indice, 3]), P = pvalores)

manhattan(datt, chr="CHR", bp="BP", snp="SNP", p="P", highlight = top10_snps)
# Pegando os top 10 para conferir
top10_idx <- order(datt$P)[1:10]
top10_snps <- datt$SNP[top10_idx]

# Ver se tem significativo naquele limiar
significativos <- which(pvalores < 5e-8)
```