LUCAS FERREIRA ROSA

COMPARANDO FORMAS DE ANÁLISE PARA DADOS CENSURADOS POR RAZÕES PRÁTICAS EM PROGRAMAS DE MELHORAMENTO VEGETAL

LAVRAS – MG

2023

LUCAS FERREIRA ROSA

COMPARANDO FORMAS DE ANÁLISE PARA DADOS CENSURADOS POR RAZÕES PRÁTICAS EM PROGRAMAS DE MELHORAMENTO VEGETAL

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Professor Júlio Sílvio de Sousa Bueno Filho
Orientador

LAVRAS – MG
2023

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Rosa, Lucas Ferreira.
Comparando formas de análise para dados censurados por razões práticas em programas de melhoramento vegetal / Lucas Ferreira Rosa. - 2023.
101 p. : il.

Orientador(a): Júlio Sílvio de Sousa Bueno Filho.

Dissertação (mestrado acadêmico) - Universidade Federal de Lavras, 2023.
Bibliografia.

1. Análise Bayesiana. 2. Dados censurados. 3. Melhoramento vegetal. I. Bueno Filho, Júlio Sílvio de Sousa. II. Título.

LUCAS FERREIRA ROSA

COMPARANDO FORMAS DE ANÁLISE PARA DADOS CENSURADOS POR RAZÕES PRÁTICAS EM PROGRAMAS DE MELHORAMENTO VEGETAL
COMPARING FORMS OF ANALYSIS FOR DATA CENSORED FOR PRACTICAL REASONS IN PLANT BREEDING PROGRAMS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 16 de janeiro de 2023.

Prof. Dr. José Airton Rodrigues Nunes	UFLA
Prof. Dr. Diógenes Ferreira Filho	UFRRJ
Prof. Dr. João Domingos Scalon	UFLA

Documento assinado digitalmente
gov.br
JULIO SILVIO DE SOUSA BUENO FILHO
Data: 20/03/2023 17:48:47-0300
Verifique em https://validar.iti.gov.br

Professor Júlio Sílvio de Sousa Bueno Filho
Orientador

LAVRAS – MG
2023

Dedico este trabalho em primeiro lugar a Deus que iluminou o meu caminho durante está caminhada. Aos meus pais, irmãos e a minha companheira Isadora.

AGRADECIMENTOS

Aos meus pais José e Isilene, irmãos Fábio e Bruno e a minha namorada Isadora que, com muito carinho e apoio, não mediram esforços para que eu chegasse até está etapa de minha vida. Ao professor Júlio por sua orientação, dedicação e disponibilidade, e por tudo que me ensinou durante a realização deste trabalho. Aos professores José Airton, Diógenes e Scalon por participarem da banca examinadora. A todos meus colegas do curso, pelos momentos compartilhados. A todos os professores do DES-UFLA que durante todo o curso tiveram um papel importante na minha formação. Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela bolsa de estudo concedida.

Comece fazendo o que é necessário, depois o que é possível, e de repente estará fazendo o impossível.
São Francisco de Assis

RESUMO

Experimentos de diversas áreas estão sujeitos à ocorrência de dados censurados por razões práticas, tais como limiares de detecção dados por aparelhos além dos quais não se registram valores. Nestes casos, pode-se supor que há uma variável aleatória latente contínua que contém estes limiares. Isto permite investigar as consequências da truncagem informada de dados para ajustar as distribuições associadas às populações experimentais. Este trabalho foi inspirado em um experimento de melhoramento genético de batata-doce e tem como objetivo desenvolver um método de análise para dados com censura à esquerda implementando um algoritmo para a sua previsão condicional usando a amostragem Gibbs e verificar suas propriedades em comparação a métodos usuais de análise de ensaios deste tipo em um exemplo simulado com propriedades semelhantes. Para tanto, simulamos um experimento em blocos incompletos parcialmente balanceados (PBIB) organizados em látice quadrado ($v = k^2, r = 3, k = 11, b = 33, \lambda_1 = 1$ e $\lambda_2 = 0$). Foram realizadas as seguintes análises: a análise sem censura dos dados completos "DC", as análises usuais Censura Zero ("C0", considerando zero para as censuras) e de Censura à esquerda ("CE", considerando observações perdidas na censura) e a análise proposta de Previsão Condicional ("PC", com imputação condicional dos dados censurados). As três formas de análise que lidam com censuras foram comparadas às referências em dois cenários: 1) taxa de censura moderada ($\sim 30\%$) e 2) taxa de censura alta ($\sim 50\%$). Foram verificadas a acurácia, a precisão e o viés na estimação dos parâmetros genéticos (componentes de variância e herdabilidades). Foram também calculadas as correlações entre os valores originalmente simulados com as observações que foram censuradas. Por fim, em cada análise, foram calculadas as correlações de Pearson e Spearman entre os valores genéticos preditos e respectivos valores paramétricos. Em ambos os cenários a análise "PC" foi precisa para fins de seleção, apresentando correlações entre os efeitos de tratamentos e valores paramétricos próximas da análise sem censura. Além disso, pelo método proposto, os valores simulados são frequentes nas respectivas distribuição marginais *a posteriori*. As formas usuais de análise ("C0" e "CE") têm correlação nula entre os valores tomados como zero e os valores paramétricos. A análise "CE" foi ruim em ambos os cenários quanto a estimação de parâmetros genéticos (especialmente variâncias e herdabilidades) e por apresentar baixa correlação, mas para seleção dos genótipos de elite foi melhor que a análise "C0" no cenário 2). Para o cenário 1), menor taxa de censura, a análise "C0" parece ser uma alternativa interessante, mas com uma piora considerável com o aumento da censura. Embora a análise "PC" tenha produzido declarações mais difíceis de interpretar por superestimar as herdabilidades, foi a análise mais indicada para se tomar decisões sobre seleção.

Palavras-chave: Análise Bayesiana. Componentes da variância. Dados censurados. Melhora mento vegetal. Modelos de limiar.

ABSTRACT

Experiments in several areas are subject to occurrences of censored data for practical reasons, such as detection thresholds given by devices beyond which no values recorded. In these cases one can assume that a continuous latent random variable contains these thresholds. Thus allows for investigating the consequences informed truncation of data to fit distributions associated with experimental populations. Present work, inspired by a sweet potato genetic improvement experiment, aims to develop an analysis method for data with left censoring by implementing an algorithm for its conditional prediction using Gibbs sampling and verifying its properties. We compared the analysis on a simulated example with similar properties. Simulated experiment was an incomplete block design partially balanced (PBIB) (square lattice with $v = k^2, r = 3, k = 11, b = 33, \lambda_1 = 1$ e $\lambda_2 = 0$). The methods carried out were: the uncensored analysis of the complete "DC"data, the usual Zero Censorship ("C0" considering zero for the censorships) and Left Censorship ("CE" considering missing observations in the censorship) and the proposed analysis of Conditional Prediction ("PC" with conditional imputation of censored data). Competing analysis with censored data were compared to references in two scenarios: 1) moderate or high censoring ($\sim 30\%$ $\sim 50\%$). We evaluated selective accuracy, precision and bias in the estimates of genetic parameters (variance components and heritabilities). We also obtained correlations of simulated and predicted the censored observations . Finally, in each analysis, the Pearson and Spearman correlations between predicted genetic values and respective parametic values were calculad. "PC" analysis was sensible and accurate for selection purposes, showing correlations between treatment effects and parametric values close to the uncensored case. The proposed method has the simulated values very likely in the respective marginal *a posteriori* distributions. The usual forms of analysis ("C0" and "CE") have zero correlation between the values taken as zero and the parametric values. The "CE" analysis was bad in both scenarios regarding the estimation of genetic parameters (especially variances and heritabilities) and for presenting low correlation, but for the selection of the elite genotypes it was better than the "C0" analysis in the scenario two). For scenario 1) the "C0" analysis seems to be an promising alternative, but has shown a considerable worsening with increasing censorship. Although the "PC" analysis produced statements that were more difficult to interpret because it overestimated heritabilities, it was the most indicated to make selection decisions.

Keywords: Bayesian analysis. Censored data. Plant breeding. Variance components. Threshold models.

LISTA DE FIGURAS

Figura 2.1 – Representação das categorias de censura: esquerda, direita e intervalar. 19

Figura 2.2 – Ilustração da Censura à esquerda com limite de detecção em τ. 21

Figura 3.1 – Gráfico de dispersão entre os valores simulados e sua esperança condicional aos valores de efeitos genéticos e de controle local. A correlação é de $69,49\%$. 33

Figura 3.2 – Ilustração de uma censura à esquerda com os dados censurados substituídos por zeros. 35

Figura 4.1 – Transformação de Box e Cox aplicada a $\mathbf{y}$, com $\lambda = 0$ (Cenário: DC). 45

Figura 4.2 – Gráfico de dispersão das predições dos efeitos de tratamentos para o caso "DC" (Cenário 1) e seus respectivos valores paramétricos ($\hat{\rho} = 69,29\%$). 47

Figura 4.3 – Transformação de Box e Cox aplicada a $\mathbf{y} + \mathbf{0,3}$, com $\lambda = 0$ (Cenário 1: C0). . 49

Figura 4.4 – Gráfico de dispersão das predições dos efeitos de tratamentos para o caso "C0" e seus respectivos valores paramétricos ($\hat{\rho} = 69,06\%$). 50

Figura 4.5 – Transformação de Box e Cox aplicada a $\mathbf{y} + \mathbf{0,2}$, com $\lambda \neq 0$ (Cenário 1: CE). . 52

Figura 4.6 – Gráfico de dispersão das predições dos efeitos de tratamentos "CE" e seus respectivos valores paramétricos ($\hat{\rho} = 52,78\%$). 55

Figura 4.7 – Intervalos de credibilidade (95%) dos 107 valores de produção estimados em relação aos valores conhecidos da variável produção gerados na simulação, antes do processo de censura. 57

Figura 4.8 – Gráfico de dispersão das médias *a posteriori* dos 107 valores de produção estimados em relação aos seus respectivos valores simulados antes da censura ($r = 39,07\%$). 57

Figura 4.9 – Gráfico de dispersão das predições dos efeitos de tratamentos "PC" (Cenário 1) e seus respectivos valores paramétricos ($\hat{\rho} = 66,79\%$). 60

Figura 4.10 – Transformação de Box e Cox aplicada a $\mathbf{y} + \mathbf{0,001}$, com $\lambda = 0$ (Cenário 2: C0). 61

Figura 4.11 – Gráfico de dispersão das predições dos efeitos de tratamentos "C0" e seus respectivos valores paramétricos ($\hat{\rho} = 58,66\%$). 64

Figura 4.12 – Transformação de Box e Cox aplicada a $\mathbf{y} + \mathbf{0,1}$, com $\lambda \neq 0$ (Cenário 2: CE). . 65

Figura 4.13 – Gráfico de dispersão das predições dos efeitos de tratamentos "CE" e seus respectivos valores paramétricos ($\hat{\rho} = 44,68\%$). 68

Figura 4.14 – Intervalos de credibilidade (95%) dos 179 valores de produção estimados em relação aos valores conhecidos da variável produção gerados na simulação, antes do processo de censura. 69
Figura 4.15 – Gráfico de dispersão das médias *a posteriori* dos 179 valores de produção estimados em relação aos seus respectivos valores simulados antes da censura ($r = 20,99\%$). 70
Figura 4.16 – Gráfico de dispersão das predições dos efeitos de tratamentos ("PC") e seus respectivos valores paramétricos (correlação estimada: $\hat{\rho} = 62,72\%$). 73
Figura 4.17 – Gráfico de dispersão e coeficientes de correlação entre as 12 maiores predições de efeitos de tratamentos e respectivos valores paramétricos entre eles. Análises: DC ($\hat{\rho} = 0,54$), C0 ($\hat{\rho} = 0,54$), CE ($\hat{\rho} = 0,51$) e PC ($\hat{\rho} = 0,44$). . . 75
Figura 4.18 – Gráfico de dispersão e coeficientes de correlação entre as 12 maiores predições de efeitos de tratamentos e respectivos valores paramétricos entre eles. Análises: DC ($\hat{\rho} = 0,54$), C0 ($\hat{\rho} = -0,40$), CE ($\hat{\rho} = 0,59$) e PC ($\hat{\rho} = 0,52$). . 77

LISTA DE TABELAS

Tabela 4.1 – Estimativas das componentes de variância e herdabilidades do caso "DC", com seus respectivos IC (95%) e valores paramétricos usando os dados transformados. 46

Tabela 4.2 – Correlações de Pearson e Sperman dos efeitos de tratamentos com seus respectivos efeitos paramétricos para o caso "DC". Intervalo de confiança (95%) usando a aproximação *t* de Student para correlações Pearson e Sperman. . . . 47

Tabela 4.3 – BLUP para os efeitos de tratamentos considerando todas as observações conhecidas, e seus respectivos erros-padrões, IC (95%) e valores paramétricos. . 48

Tabela 4.4 – Estimativas das componentes de variância e herdabilidades do caso "C0" (Cenário 1) , com seus respectivos IC (95%) e valores paramétricos usando os dados transformados. 49

Tabela 4.5 – Correlações de Pearson e Sperman dos efeitos de tratamentos com seus respectivos efeitos paramétricos para o caso "C0" (Cenário 1). Intervalo de confiança (95%) usando a aproximação *t* de Student para correlações Pearson e Sperman. 50

Tabela 4.6 – BLUP para os efeitos de tratamentos com a perda dos dados censurados, e seus respectivos erros-padrões, IC (95%) e valores paramétricos. 51

Tabela 4.7 – Estimativas das componentes de variância e herdabilidades do caso "CE" (Cenário 1) , com seus respectivos IC (95%) e valores paramétricos usando os dados transformados. 53

Tabela 4.8 – Correlações de Pearson e Sperman dos efeitos de tratamentos com seus respectivos efeitos paramétricos para o caso "CE". Intervalo de confiança (95%) usando a aproximação *t* de Student para correlações Pearson e Sperman. . . . 53

Tabela 4.9 – BLUP para os efeitos de tratamentos com a perda dos dados censurados, e seus respectivos erros-padrões, IC (95%) e valores paramétricos. 54

Tabela 4.10 – Resumos da distribuição *a posteriori* obtida pela amostragem Gibbs em comparação aos valores paramétricos gerados na simulação. 56

Tabela 4.11 – Resumos das estimativas (distribuição *a posteriori*) das componentes de variância e herdabilidades em comparação aos seus respectivos valores paramétricos. 58

Tabela 4.12 – Resumos da distribuição *a posteriori* obtida pela amostragem Gibbs dos efeitos de tratamentos em comparação aos valores paramétricos gerados na simulação. 59

Tabela 4.13 – Correlações de Pearson e Sperman dos efeitos de tratamentos com seus respectivos efeitos paramétricos para o caso "PC" (Cenário 1). Intervalo de confiança (95%) usando a aproximação *t* de Student para correlações Pearson e Sperman. 60
Tabela 4.14 – Estimativas das componentes de variância e herdabilidades do caso "C0" (Cenário 2) , com seus respectivos IC (95%) e valores paramétricos usando os dados transformados. 61
Tabela 4.15 – Correlações de Pearson e Sperman dos efeitos de tratamentos com seus respectivos efeitos paramétricos para o caso "C0" (Cenário 2). Intervalo de confiança (95%) usando a aproximação *t* de Student para correlações Pearson e Sperman. 62
Tabela 4.16 – BLUP para os efeitos de tratamentos com a perda dos dados censurados, e seus respectivos erros-padrões, IC (95%) e valores paramétricos. (continua) 62
Tabela 4.18 – Estimativas das componentes de variância e herdabilidades do caso "CE" (Cenário 2) , com seus respectivos IC (95%) e valores paramétricos usando os dados transformados. 65
Tabela 4.19 – Correlações de Pearson e Sperman dos efeitos de tratamentos com seus respectivos efeitos paramétricos para o caso "CE" (Cenário 2). Intervalo de confiança (95%) usando a aproximação *t* de Student para correlações Pearson e Sperman. 66
Tabela 4.20 – BLUP para os efeitos de tratamentos com a perda dos dados censurados, e seus respectivos erros-padrões, IC (95%) e valores paramétricos. (continua) 66
Tabela 4.22 – Resumos da distribuição *a posteriori* obtida pela amostragem Gibbs em comparação aos valores paramétricos gerados na simulação. (continua) 68
Tabela 4.24 – Resumos das estimativas (distribuição *a posteriori*) das componentes de variância e herdabilidades em comparação aos seus respectivos valores paramétricos. 71
Tabela 4.25 – Resumos da distribuição *a posteriori* obtida pela amostragem Gibbs dos efeitos de tratamentos em comparação aos valores paramétricos gerados na simulação. (continua) 71
Tabela 4.27 – Correlações de Pearson e Sperman dos efeitos de tratamentos com seus respectivos efeitos paramétricos para o caso "PC" (Cenário 2). Intervalo de confiança (95%) usando a aproximação *t* de Student para correlações Pearson e Sperman. 73

Tabela 4.28 – Quadro-resumo das estimativas das herdabilidades, com seus respectivos IC com (95%) de probabilidade e seus valores paramétricos. 74
Tabela 4.29 – Quadro-resumo das porcentagens de recuperação da informação genética e correlações (Pearson) dos efeitos paramétricos de tratamentos, com suas respectivas predições pelos diferentes métodos de análise. Intervalo de confiança (95%) usando a aproximação *t* de Student. 74
Tabela 4.30 – Quadro-resumo das estimativas da herdabilidade e herdabilidade experimental, com seus respectivos IC com (95%) de probabilidade e seus valores paramétricos. 76
Tabela 4.31 – Quadro-resumo das porcentagens de recuperação da informação genética e correlações (Pearson) dos efeitos paramétricos de tratamentos, com suas respectivas predições pelos diferentes métodos de análise. Intervalo de confiança (95%) usando a aproximação *t* de Student. 77

SUMÁRIO

1 INTRODUÇÃO 16
2 REFERENCIAL TEÓRICO 18
2.1 Dados censurados 18
2.1.1 Dados censurados à esquerda 20
2.2 Métodos para análise estatística de conjuntos de dados censurados 21
2.3 Inferência Bayesiana 23
2.3.1 Distribuição *a priori* não informativa 24
2.3.2 Distribuição *a priori* informativa 24
2.3.3 Teorema de Bayes 24
2.3.4 Amostrador de Gibbs 26
2.4 Teoria de modelos mistos 28
2.5 Métodos misto e estimação pelo método da máxima verossimilhança restrita (REML) 29
3 MATERIAL E MÉTODOS 32
3.0.1 Modelo da simulação do experimento 32
3.1 Modelo da análise e suas pressuposições 34
3.1.1 Metodologia empregada para se simular a censura à esquerda 35
3.2 Estimação de componentes da variância e valores genéticos 35
3.3 Previsão condicional em modelagem hierárquica bayesiana 36
3.3.1 Conjunto de parâmetros 37
3.3.2 Verossimilhança 38
3.3.3 Especificação das *prioris* 38
3.3.4 Distribuição *a posteriori* conjunta 39
3.3.5 Amostra condicional do vetor η 40
3.3.6 Distribuições condicionais completas para os demais parâmetros 41
3.4 Análises 43
3.5 Implementação 44
4 RESULTADOS E DISCUSSÃO 45
4.1 Cenário: supondo que se conheça os dados completos ("DC"): 45

4.2 Cenário 1: substituindo censuras por 0 ("C0") 48
4.3 Cenário 1: eliminando censuras à esquerda ("CE") 52
4.4 Cenário 1: previsão condicional dos dados censurados ("PC") 55
4.5 Cenário 2: substituindo censuras por 0 ("C0") 61
4.6 Cenário 2: eliminando censuras à esquerda ("CE") 64
4.7 Cenário 2: previsão condicional dos dados censurados ("PC") 68
4.8 Síntese do Cenário 1 73
4.9 Síntese do Cenário 2 75
4.10 Discussão Geral 77
5 CONCLUSÕES 79
REFERÊNCIAS 80
APENDICE A – Simulação do experimento 84
APENDICE B – Análise REML 87
APENDICE C – Inferência Bayesiana 97

É comum que nas análises estatísticas em diversas áreas do conhecimento haja dados censurados (*censored data*). No contexto da análise de sobrevivência, onde originalmente se tratou do problema da censura, estas ocorrem porque o estudo terminou e não se conseguiu observar o tempo de ocorrência do evento de interesse. Dependendo da aplicação, tal censura à direita é referida como a análise de sobrevivência, também referida como confiabilidade, tempo de vida, tempo para eventos ou histórico de eventos (LEUNG; ELASHOFF; AFIFI, 1997)).

O tipo de censura pode interferir diretamente nos resultados da análise, provocando potenciais alterações nas interpretações e conclusões do estudo de interesse. Intervalos de confiança, testes de hipóteses, ajuste de distribuições de probabilidade, correlações e parâmetros estatísticos básicos como média e variância, podem sofrer alterações em razão da censura. Desse modo, dados censurados são aqueles valores não observados que possuem significado na análise estatística e, se observados, teriam potencial de alterar a análise. Em outras palavras, informações ausentes de forma incompleta, que implicam em informação relevante para a análise (LITTLE; RUBIN, 2019). O problema dos dados censurados está relacionado ao dos dados perdidos completamente ao acaso (ausentes), no entanto há motivos entre as causas de variação estudadas para que se incorra em maior ou menor probabilidade de perda de dados. Ou seja, as análises usuais de dados perdidos podem não ser adequadas se houver censura.

Frequentemente em ciências agrárias (e outras áreas de aplicação) alguns caracteres são medidos apenas acima de limiares de detecção indicados por limitações técnicas ou por necessidades práticas. Como exemplo, em estágios iniciais do melhoramento da cultura de batata-doce, o descarte realizado é importante, pois evita que sejam selecionados genótipos com má formação inicial ou com crescimento lento e deficitário. Desse modo, centenas de parcelas com pequena produtividade de tubérculos são simplesmente desprezadas (nem mesmo medidas), ocasionando assim por razões práticas a censura à esquerda desses dados. Tal censura é, na verdade, uma forma de seleção de quais fenótipos serão observados.

Este problema ocorre nos dados obtidos por Silva (2019) em experimento de melhoramento de batata-doce conduzido na área experimental do Setor de Olericultura da Universidade Federal de Lavras (UFLA), no município de Lavras, MG. O limiar de censura adotado foi de 1,2Kg de tubérculos por parcela, sendo que foram registradas medidas apenas para as demais parcelas (po-

tencialmente com genótipos mais promissores). No momento da colheita foram censurados 858 ($\sim 50\%$) dados dos 1.606 genótipos do delineamento inicial (SILVA, 2019).

Uma primeira opção de análise destes dados é supor perda inteiramente casual. Outras opções possíveis envolvem substituir os valores faltantes por zero ou por um valor associado ao limite de detecção. Em ambos os casos, a hipótese é de que haverá consequências negativas na estimação de parâmetros genéticos de interesse. Uma opção de modelagem mais consistente com o mecanismo de censura é modelar de forma hierárquica o experimento, supondo que exista uma distribuição subjacente contínua contendo os valores que teriam sido observados. Neste caso as estimativas resultantes poderiam ser realistas, mas devemos investigar como se comportam em cenários com baixa e alta taxa de censura. Os primeiros são mais frequentes em experimentos finalizadores, com genótipos superiores, os outros ocorrem na seleção inicial de genótipos (onde evidentemente, se espera maior porcentagem de genótipos ruins e descarte de parcelas).

O objetivo geral do presente trabalho foi apresentar uma forma de análise para dados com censura à esquerda utilizando a modelagem hierárquica bayesiana. Os objetivos específicos foram: implementar uma rotina **R** para a modelagem bayesiana do problema de censura à esquerda, comparar as propriedades deste modelo às formas usuais de análise prática para ensaios deste tipo em exemplos simulados e analisar os modelos em cenários com moderado (30%) e alto (50%) grau de censura.

2 REFERENCIAL TEÓRICO

Na experimentação agronômica, em diversas situações de pesquisa, não se pode medir com precisão uma variável de interesse. Seus valores não podem então ser expressos como números, mas muitas vezes podem ser representados por intervalos de incerteza em que os valores reais se encontram. Por exemplo, no cultivo de batata-doce tubérculos que apresentam pouca produtividade não são avaliados. Diante disso, não é possível de se dizer com precisão, a produtividade de tubérculos das parcelas com peso abaixo de um determinado valor, mas é possível de se dizer que seus valores reais encontram-se abaixo de um certo limite (peso). Tal observação não é um valor exato, mas apenas um intervalo. Esses dados são um exemplo dos chamados de dados censurados (ONOFRI; PIEPHO; KOZAK, 2019).

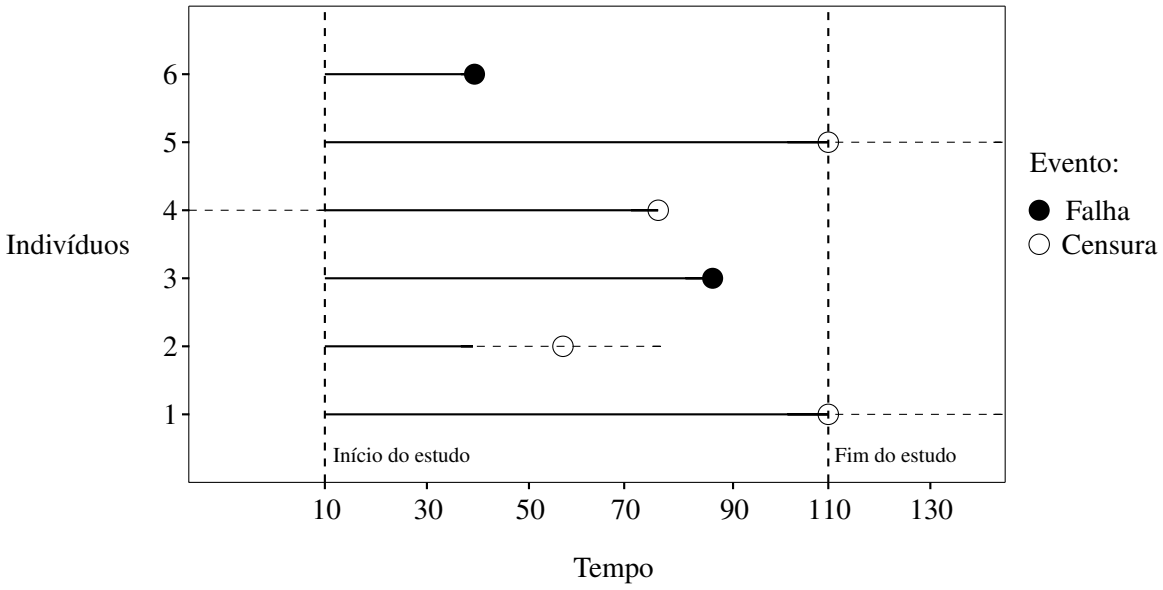
2.1 Dados censurados

Dados censurados ocorrem frequentemente na agricultura, mas geralmente são analisados no contexto da análise de sobrevivência. A análise de sobrevivência é um ramo da estatística que possui ferramentas estatísticas e computacionais, cujo objetivo é analisar e modelar dados onde o resultado é o tempo até que um evento de interesse ocorra, evento esse denominado tempo de falha ou de sobrevivência (COLOSIMO; GIOLO, 2006; GUO, 2010).

Nos estudos de tempo até a ocorrência de um evento, os dados obtidos podem apresentar-se de diferentes maneiras, criando problemas típicos na análise de tais dados. A censura é uma característica peculiar que ocorre frequentemente e por diversas razões nos dados de tempo de vida. Sumariamente falando, a censura acontece quando se sabe que alguns eventos ocorreram apenas dentro de certos intervalos, com o restante dos eventos conhecidos exatamente (KLEIN; MOESCHBERGER, 2003; LIU, 2012; REID; COX, 2018).

As categorias mais comuns de censura são apresentadas na Figura 2.1 a seguir (COLOSIMO; GIOLO, 2006; KLEINBAUM; KLEIN et al., 2012; SMITH, 2017):

Figura 2.1 – Representação das categorias de censura: esquerda, direita e intervalar.



Fonte: Pereira (2020), adaptado.

a) **censura à esquerda** - para o indivíduo 4 na Figura 2.1, tudo o que é conhecido é que o indivíduo experimentou o evento de interesse antes do início do estudo, ou seja, um dado que embora seja desconhecido está abaixo de um determinado valor;

b) **censura à direita** - para os indivíduos 1 e 5 na Figura 2.1, tudo o que se sabe é que esses indivíduos ainda estão vivos no fim do estudo, ou seja, dados que embora sejam desconhecidos estão acima de um determinado valor. A censura à direita de dados pode ser classificada em três tipos:

   - **tipo I:** ocorre se um experimento com um número definido de itens é interrompido em um tempo predeterminado, ponto no qual todos os itens restantes são censurados à direita;
   - **tipo II:** ocorre se um experimento tem um número definido de itens e para o experimento quando um número predeterminado é observado como tendo falhado; os itens restantes são então censurados à direita;
   - **censura aleatória (ou não informativa)** - ocorre quando cada sujeito tem um tempo de censura estatisticamente independente do tempo de falha. O valor observado é o mínimo dos tempos de censura e falha; assuntos cujo tempo de falha é maior do que seu tempo de censura são censurados à direita;

c) **censura de intervalo** - para o indivíduo 2 na Figura 2.1, a única informação é que o evento ocorre dentro de algum intervalo, ou seja, um dado que está em algum lugar em um intervalo entre dois valores. A censura de intervalo pode ocorrer quando a observação de um valor requer acompanhamentos ou inspeções. A censura à esquerda e à direita são casos especiais de censura de intervalo, com o início do intervalo em zero ou o fim no infinito, respectivamente (KLEINBAUM; KLEIN et al., 2012).

Para os indivíduos nomeados 3 e 6 na Figura 2.1 o tempo de falha foi observado, logo para esses indivíduos tem-se o tempo do início do estudo até a ocorrência do evento de interesse.

2.1.1 Dados censurados à esquerda

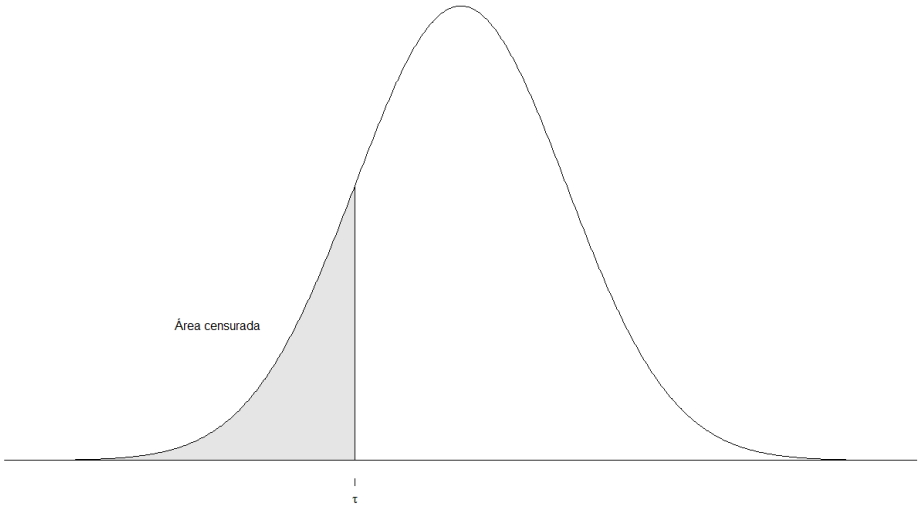
Na censura à esquerda, caso considerado neste trabalho, um indivíduo específico num estudo de sobrevivência é considerado ser censurado à esquerda se o evento de interesse for menor que um tempo de censura T, isto é, o evento de interesse ocorreu antes desse indivíduo ser observado no estudo. Para esses indivíduos, sabe-se que eles experienciaram o evento algum tempo antes do tempo T, mas o tempo exato do evento é desconhecido (KLEIN; MOESCHBERGER, 2003). O tempo exato de vida de um indivíduo será conhecido se, e somente se, seu tempo de vida for maior ou igual a um valor τ previamente estabelecido (LIU, 2012).

O mesmo ocorre no campo das ciências agrárias, já que dados de interesse menores que um limite são considerados censurados à esquerda. Para esses dados, sabe-se somente que seus valores encontram-se abaixo de um certo limite, mas esses valores não são exatamente conhecidos. Desse modo, além dos estudos de sobrevivência, pesquisadores na experimentação agronômica trabalham com dados censurados em diversas situações, podendo ser essas situações inerentes aos métodos analíticos de medição (limitações técnicas) ou questões práticas. A perda de informações desses dados não é meramente ao acaso, são dados desconhecidos (rejeitados) que se encontram fora de um determinado intervalo (ONOFRI; PIEPHO; KOZAK, 2019). Portanto, para o exemplo citado anteriormente, no cultivo de batata-doce parcelas com baixa produtividade de tubérculos são simplesmente desprezadas, ocasionando por razões práticas a censura à esquerda desses dados.

Caso ignorada, a censura à esquerda gera uma perda de informações, acarretando assim numa distorção potencialmente grave das inferências, o que pode se acentuar de acordo com o

grau de censura (SÖRENSEN; GIANOLA; KORSGAARD, 1998). Sob o olhar estatístico, dados com valores abaixo de um certo limite são denominados "censurados à esquerda" ou simplesmente "censurados". Outros termos aplicados para esse limite são: "valor crítico"; "limite de detecção do método" ou apenas "limite de detecção". Conforme a Figura 2.2, uma vez estabelecido, o limite crítico pode ser utilizado como um nível de censura para medições observadas abaixo desse limite descrito (CHRISTOFARO; LEÃO, 2014).

Figura 2.2 – Ilustração da Censura à esquerda com limite de detecção em τ.



Fonte: Do autor (2022).

2.2 Métodos para análise estatística de conjuntos de dados censurados

Experimentos controlados normalmente são cuidadosamente projetados, de modo a permitir a realização de análises estatísticas utilizando cálculos simples. Devido ao equilíbrio nos experimentos, estimativas, erros padrão e a tabela ANOVA correspondentes à maioria dos experimentos planejados são facilmente computados (LITTLE; RUBIN, 2019).

Uma grande quantidade de informação pode estar disponível nos dados censurados de um experimento e se um método eficiente for utilizado, informações validas podem ser extraídas destes dados de modo a validar as inferências. Desse modo os dados censurados devem ser considerados no estudo, pois distorções consideráveis podem ser geradas nas estimação dos parâmetros e na utilização dos testes estatísticos com a ausência desses dados. Técnicas específicas devem ser

utilizadas para minimizar a interferência negativa das observações censuradas (CHRISTOFARO; LEÃO, 2014).

Existem diversas maneiras de lidar com valores censurados em conjuntos de dados. Excluí-los ou ignorá-los é a abordagem mais simples e padrão. Substituir os valores censurados por zero é outra opção, mas esses métodos de tratamento possuem grandes desvantagens na medida que degradam a qualidade das estimativas. Remover algumas informações presentes ou, simplesmente substituir esses valores faltantes por zero, acabam influenciando nas estimativas obtidas dos parâmetros de interesse (MCGRORY; HOLIAN; MORRISON, 2020; VAN BUUREN, 2018).

Um método atraente para se lidar com a censura é preencher os valores ausentes por um conjunto de valores plausíveis para restaurar a análise estatística, para posteriormente prosseguir com a análise padrão. Algumas vantagens de preencher valores ausentes em um experimento em vez de tentar analisar os dados reais observados, incluem:

a) especificar a estrutura de dados utilizando a terminologia do projeto experimental se torna menos complicado (por exemplo, um bloco incompleto balanceado);
b) menor complexidade no cálculo de resumos estatísticos;
c) fácil interpretação e exibições dos resultados, pois resumos padrões podem ser utilizados.

Em vista disso, uma alternativa possível para se trabalhar com dados censurados é gerar esses dados censurados com valores obtidos através de métodos que possibilitam que analises desses dados sejam feitas de modo a obter inferências válidas. Esses métodos são geralmente chamados de métodos de imputação. Esses métodos são uma espécie de tratamento para a incerteza e imprecisão existentes nos conjuntos de dados (AMIRI; JENSEN, 2016; ONOFRI; PIEPHO; KOZAK, 2019).

Desse modo, a imputação é um método geral e flexível para lidar com problemas de dados censurados, no entanto possui armadilhas. A ideia de imputação é sedutora porque pode influenciar o pesquisador a acreditar no final das contas, que os dados estão completos, e é perigosa, pois agrupa situações em que o problema é suficientemente pequeno para que possa ser tratado legitimamente dessa maneira (DEMPSTER; RUBIN, 1983).

As imputações devem ser conceituadas como retiradas de uma distribuição preditiva dos valores censurados e requerem um método para criar uma distribuição primitiva para a imputação com base nos dados observados. Vale ressaltar, que os valores imputados não são reais. A análise

de dados imputados individualmente, sem refletir a incerteza sobre o valores está incorreto e deve ser evitado na prática (RAGHUNATHAN, 2015).

2.3 Inferência Bayesiana

Os métodos Bayesianos constituem um conjunto de técnicas alternativas à análise clássica, que podem facilitar a interpretação dos resultados e que permitem a incorporação de informações sobre os parâmetros do modelo antes dos dados serem observados (BOLSTAD; CURRAN, 2016; BOX; TIAO, 1973). Na estatística quando se deseja efetuar uma análise dispõe-se basicamente de dois paradigmas alternativos, ou seja, duas linhas de pensamentos distintos: o paradigma tradicional (clássico) freqüentista que dominou as análises estatísticas por um longo tempo, e o paradigma Bayesiano que vem se destacando cada vez mais com o avanço computacional (KINAS; ANDRADE, 2021).

Na escola freqüentista, as probabilidades são medidas em infinitas repetições teóricas, sob condições idênticas, do experimento de interesse. A população é descrita por um modelo probabilístico que é função de constantes desconhecidas (parâmetros populacionais). Já na escola bayesiana cada observação é única e procura-se estabelecer distribuições de probabilidade para os parâmetros populacionais e, ou, outras variáveis aleatórias, como previsões de novas observações (GELMAN; CARLIN et al., 2013),

Do ponto de vista freqüentista, o parâmetro de interesse θ é um valor constante ou fixo, além de não levar em consideração informações prévias provenientes da experiência que o pesquisador possui em relação o objeto de estudo. Diferente da inferência freqüentista, na inferência bayesiana interpreta-se o parâmetro de interesse θ como variável aleatória, levando em consideração fontes de informações a amostra associada ao experimento e os conhecimentos prévios que o pesquisador possui do experimento.

Esse conhecimento prévio que o pesquisador acrescenta em sua análise sobre o parâmetro θ de interesse nada mais é na inferência bayesiana que sua *priori* em relação a θ, mais especificamente a densidade *a priori* de θ, já que nos métodos Bayesianos utiliza-se de probabilidade para quantificar a incerteza em inferências baseadas em análise de dados estatísticos. Desse modo $p(\theta)$

é a probabilidade a qual, antes da observação dos dados, contém a distribuição de probabilidade de θ, onde θ pode ser um escalar ou um vetor de parâmetros.

A distribuição *a priori* influencia na distribuição a *posteriori*, logo é parte fundamental da inferência bayesiana, pois representa as informações sobre o parâmetro θ desconhecido. Portanto, a combinação da distribuição de probabilidade da *priori* com a distribuição de dados produz a distribuição *a posteriori*, onde se retira inferências e decisões sobre o parâmetro θ (GELMAN, 2002).

2.3.1 Distribuição *a priori* não informativa

É comum o uso de distribuições *a priori* não-informativas quando não existe, ou se tem poucas informações sobre os parâmetros θ. Neste caso é comum se pensar em todos os possíveis valores de θ como equiprováveis, isto é, com uma distribuição *a priori* uniforme (BOX; TIAO, 1973):

$$p(\theta) \propto C$$

para θ variando em um subconjunto de R.

2.3.2 Distribuição *a priori* informativa

Quando o pesquisador possui informações prévias sobre o parâmetro desejado θ, utiliza-se de uma distribuição *a priori* informativa, ou seja, uma densidade de probabilidade própria $p(\theta)$, por sua vez, especificada com o auxílio de constantes chamadas de hiperparâmetros (parâmetros da distribuição dos parâmetros). Inicialmente, os hiperparâmetros são considerados conhecidos e traduzem a informação que se tem sobre o parâmetro, antes da realização da amostra (TURKMAN; PAULINO; MÜLLER, 2019).

2.3.3 Teorema de Bayes

Para se fazer afirmações sobre a probabilidade de θ dado que tem-se $\mathbf{Y}$, deve-se começar com um modelo que forneça uma distribuição conjunta para o parâmetro e a amostra. A junção

das informações que os dados (amostra) possuem com o conhecimento pré-existente (subjetivo) do pesquisador a respeito dos parâmetros só é possível através do Teorema de Bayes, daí o termo inferência bayesiana (BOX; TIAO, 1973).

Seguindo a definição de Box e Tiao (1973), suponha $\mathbf{Y} = (y_1, ..., y_n)$ um vetor de n observações cuja distribuição de probabilidade $p(\mathbf{Y}|\theta)$ depende dos k valores do parâmetro $\theta' = (\theta_1, ..., \theta_k)$. Suponha também que a distribuição de probabilidade de θ é $p(\theta)$. Então, $p(\mathbf{Y}|\theta)p(\theta) = p(\mathbf{Y}, \theta) = p(\theta|\mathbf{Y})p(\mathbf{Y})$.
Dado $\mathbf{Y}$ observado, a distribuição condicional de θ é:

$$p(\theta|\mathbf{Y}) = \frac{p(\mathbf{Y}|\theta)p(\theta)}{p(\mathbf{Y})} \tag{2.1}$$

onde, pode-se escrever

$$p(\mathbf{Y}) = E\left[p(\mathbf{Y}|\theta)\right] = c^{-1} = \begin{cases} \int p(\mathbf{Y}|\theta)p(\theta)d\theta & \text{se } \theta \text{ contínuo} \\ \sum p(\mathbf{Y}|\theta)p(\theta) & \text{se } \theta \text{ discreto} \end{cases}$$

dessa maneira a soma ou a integral é tomada sobre a amplitude admissível de θ, onde $E[f(\theta)]$ é a esperança matemática de $f(\theta)$ com relação a distribuição $p(\theta)$. Assim pode-se escrever 2.1 alternativamente como:

$$p(\theta|\mathbf{Y}) = cp(\mathbf{Y}|\theta)p(\theta). \tag{2.2}$$

A condição em 2.1, ou seu equivalente em 2.2 é referido como Teorema de Bayes. Nesta expressão, $p(\theta)$, diz o que é conhecido a respeito de θ sem o conhecimento dos dados, é chamada de distribuição *a priori* de θ, geralmente tratado como contínuo. Para um valor fixo de $\mathbf{Y}$, a função $l(\mathbf{Y};\theta) = p(\mathbf{Y}|\theta)$ fornece a plausibilidade ou verossimilhança de cada um dos possíveis valores de θ. Correspondentemente, a combinação da verossimilhança com *a priori* nos fornece $p(\theta|\mathbf{Y})$, o que é conhecido a respeito de θ dado o conhecimento dos dados, é chamada distribuição *a posteriori* de θ dado $\mathbf{Y}$. A quantidade c é simplesmente uma constante de "normalização" necessária para assegurar-se de que a distribuição *a posteriori* $p(\theta|\mathbf{Y})$ integrada ou somada (em caso discreto) é 1.

Estas duas fontes de informação, priori e verossimilhança, quando combinadas levam a distribuição *a posteriori* de θ, $p(\theta|\mathbf{Y})$. Assim, a forma usual do Teorema de Bayes é,

$$p(\theta|\mathbf{Y}) \propto l(\mathbf{Y};\theta)p(\theta). \tag{2.3}$$

Em outras palavras, o Teorema de Bayes diz que a distribuição de probabilidade de θ a *posteriori* para $\mathbf{Y}$ dado é proporcional ao produto da distribuição *a priori* para θ e a verossimilhança de θ dado $\mathbf{Y}$ (BOX; TIAO, 1973). Isto é,

Distribuição *a posteriori* $\propto$ verossimilhança $\times$ distribuição *a priori*

Observando-se as quantidades y_i com n repetições independentes dado θ, ou seja, $p(y_i|\theta)$ segue que

$$p(\theta|y_1, y_2, ..., y_n) \propto \prod_{i=1}^{n} l(y_i|\theta)p(\theta). \tag{2.4}$$

Obtida a distribuição *a posteriori* é possível fazer inferências sobre as distribuições de probabilidade conjunta dos parâmetros, mas na maioria dos casos, deseja-se encontrar uma distribuição para um parâmetro específico. A forma de encontrar tal distribuição, chamada distribuição marginal, está na integração da distribuição *posteriori* conjunta, obtendo uma função dessa integral para o parâmetro de interesse θ_i, isto é,

$$f(\theta_i|\mathbf{Y}) = \int ... \int f(\theta|\mathbf{Y}) d\theta_{-i} \tag{2.5}$$

em que θ_{-i} é o conjunto complementar de parâmetros para θ_i.

É comum em muitas situações a integração de 2.5 ser inviável, ou até mesmo impossível devido a complexidade da forma analítica das marginais. A amostragem Gibbs é uma boa ferramenta para obtenção de $f(\theta_i)$, tal método foi elaborado por Geman e Geman (1984), sendo inicialmente utilizado por Gelfand e Smith (1990), na estatística. O método numérico de Monte Carlo via Cadeias de Markov (MCMC), pode ser utilizado para gerar valores de uma distribuição condicional a posteriori para cada parâmetro.

2.3.4 Amostrador de Gibbs

O Amostrador de Gibbs (*Gibbs Sampling*) nada mais é, que um algoritmo particular de uma cadeia de Markov, ou seja, um esquema iterativo de amostragem, em que o núcleo de transição é

formado pelas distribuições condicionais completas (GELMAN; CARLIN et al., 1995). Trata-se de uma técnica para gerar variáveis aleatórias de uma distribuição marginal sem que se conheça a sua densidade. A ideia do método de Monte Carlo via Cadeia de Markov consiste em simular um passeio aleatório no espaço do parâmetro θ, o qual converge para uma distribuição estacionária, que é a distribuição a posteriori $P(\theta|\mathbf{Y}_v)$, em que $\mathbf{Y}_v$ é o vetor de observações. Essa distribuição é dita estacionária porque não sofre alterações caracterizadas por picos ou por tendências direcionais, constituindo, assim, uma base estável, ou contável, para o processo de estimação.

Considerando a existência da distribuição conjunta, pode-se pensar no Algoritmo de Gibbs como uma implementação prática do fato que o conhecimento das distribuições condicionais é suficiente para determinar a distribuição conjunta.

A essência do algoritmo de Gibbs é a amostragem de parâmetros condicionada à outros parâmetros. Suponha-se que o vetor de parâmetros θ seja dividido em p subvetores $P(\theta_1|\theta_2,...,\theta_p)$ e que as distribuições condicionais de cada parâmetro θ_i dado todos os outros sejam conhecidas. Estas distribuições são chamadas distribuições condicionais completas e denotadas por

$$P(\theta_1|\theta_2,\theta_3,...,\theta_p,\mathbf{Y}),P(\theta_2|\theta_1,\theta_3,...,\theta_p,\mathbf{Y}),...,P(\theta_p|\theta_1,\theta_2,...,\theta_{p-1},\mathbf{Y}).$$

O algoritmo de Gibbs pode ser descrito da seguinte maneira. Seja θ o vetor de parâmetros de interesse e $\mathbf{Y}$ o vetor de dados:

a) defina $\theta^{(0)} = \left(\theta_1^{(0)},\theta_2^{(0)},...,\theta_p^{(0)}\right)$ um valor arbitrário inicial para o vetor de parâmetros θ, do qual $P(\theta^0|\mathbf{Y}) > 0$. Geralmente amostramos de uma distribuição normal ou de uma distribuição especificada como a distribuição prévia de θ;

b) para $t = 1,2,...,L$, designe:

$$\begin{cases} \theta_1^{(1)} & \sim & P(\theta_1|\theta_2^{(0)},\theta_3^{(0)},...,\theta_p^{(0)},\mathbf{Y}) \\ \theta_2^{(1)} & \sim & P(\theta_2|\theta_1^{(1)},\theta_3^{(0)},...,\theta_p^{(0)},\mathbf{Y}) \\ \vdots & & \\ \theta_p^{(1)} & \sim & P(\theta_p|\theta_1^{(1)},\theta_2^{(1)},...,\theta_{p-1}^{(1)},\mathbf{Y}) \end{cases} ; \tag{2.6}$$

c) muda-se o contador t para $t+1$ e retorna-se ao passo 2. O ciclo é repetido L vezes, produzindo assim $\theta^0,\theta^1,\theta^2,...,\theta^L$.

Assumindo que a convergência foi atingida, à medida que o número t de iterações aumenta, a sequência de valores gerados se aproxima da distribuição de equilíbrio, assim, obtendo a densidade marginal desejada.

2.4 Teoria de modelos mistos

A teoria dos modelos mistos foi desenvolvida através do estudo de modelos lineares constituídos por efeitos fixos e aleatórios, com interesse de compreender relações funcionais entre variáveis, sendo elas de natureza quantitativa ou qualitativa. Estimação de componentes de variância e a predição de efeitos aleatórios, na presença de efeitos fixos, são alguns motivos que levam a adesão de um modelo linear misto.

Inicialmente, todo modelo linear que contenha média geral ou uma constante μ, tomada como fixa, e um termo referente ao erro, assumido como aleatório, é considerado como modelo linear misto. Entretanto, existe uma distinção quanto a classificação de modelos em relação ao seus efeitos. Modelos que apresentam todos seus efeitos fixos, com exceção feita ao erro experimental, considerado aleatório, é chamado de modelo fixo. Quando todos os efeitos são considerados aleatórios, com exceção de uma média geral μ, é classificado como modelo aleatório. Modelos que possuem tanto efeitos fixos quanto aleatórios, são denominados de modelo misto.

Uma suposição usual ao analisar experimentos aleatorizados é que as observações são retiradas de uma mesma população, sendo independentes e identicamente distribuídas. Nos dados de modelo misto tem-se uma estrutura mais complexa, multinível e hierárquica, ou seja, formas de dependência e covariâncias que dependem dos níveis dos fatores aleatórios. No processo de amostragem, se os tratamentos (ou níveis) representam a população inteira dos tratamentos para os quais as inferências serão aplicadas, então os tratamentos são considerados como efeitos fixos, ou seja, são aquelas cujas considerações são limitadas, isto é, restritas aos tratamentos. Já os efeitos aleatórios são aqueles que representam tratamentos de uma amostra oriunda de uma determinada população, desse modo, observações entre níveis são independentes, mas as observações dentro de cada nível são dependentes porque pertencem à mesma subpopulação (DEMIDENKO, 2013).

Para Hinkelmann e Kempthorne (2007) todos os efeitos podem ser analisados como fixos na modelagem de experimentos aleatorizados, mas pode haver vantagens am supor que alguns efeitos têm seus níveis associados a distribuições de probabilidade (e seriam, portanto, efeitos aleatórios).

Desse modo, a definição de um modelo como fixo, aleatório ou misto está relacionada à possibilidade de se predizer o comportamento de suas variáveis aleatórias ou de se estimar parâmetros do modelo para um determinado conjunto de dados. Convencionou-se chamar as estimativas de médias condicionais de efeitos aleatórios "predições" (do Inglês *prediction*), gerando BLUPs (*Best Linear Unbiased Predictor(s)*) ao invés de BLUEs (*Best Linear Unbiased Estimator(s)*) (ROBINSON, 1991).

Um exemplo de análise é a do modelo misto para um delineamento em blocos causalizados no melhoramento vegetal, em que se defina que os blocos são considerados fixos e os tratamentos aleatórios. Em geral, um modelo linear misto é descrito da seguinte forma (HENDERSON, 1984):

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon \tag{2.7}$$

em que:

- $\mathbf{Y}$: é o vetor de dimensão $(nx1)$; com n, o número de observações.
- $\mathbf{X}$: é a matriz (nxp) de delineamento dos efeitos fixos;
- β: é o vetor de dimensões $(px1)$ dos parâmetros de efeitos fixos, sendo p o número de parâmetros associados aos efeitos fixos;
- $\mathbf{Z}$: é a matriz (nxq) de delineamento dos efeitos aleatórios;
- $\mathbf{u}$: é o vetor de dimensões $(qx1)$ dos parâmetros de efeitos aleatórios, sendo q o número de parâmetros associados aos efeitos aleatórios, com $\mathbf{u} \sim N(\phi, \mathbf{G})$, onde $\mathbf{G}$ é a matriz de variâncias e covariâncias dos efeitos aleatórios e ϕ é o vetor nulo;
- ε: é o vetor de resíduos $(nx1)$, com $\varepsilon \sim N(\phi, \mathbf{R})$, onde $\mathbf{R}$ é a matriz de variâncias e covariâncias dos erros.

2.5 Métodos misto e estimação pelo método da máxima verossimilhança restrita (REML)

Em um modelo linear misto, o seguinte sistema de equações 2.8 deve ser solucionado para estimar os valores de β (efeitos fixos) e predizer os valores de $\mathbf{u}$ (efeitos aleatórios):

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{pmatrix}. \tag{2.8}$$

Os estimadores pontuais de efeitos fixos derivados destas equações são conhecidos como BLUE e os preditores de efeitos aleatórios, denominados de BLUP. As equações incluídas nesta relação matricial constituem o chamado Sistema de Equações do Modelo Misto de Henderson (HENDERSON et al., 1959).

Expressa em notação matricial e simplificada (para situações em que a variância genética e residual são homogêneas) em relação ao elemento $\mathbf{R}^{-1}$ (quando $\mathbf{R} = Var(\varepsilon) = \mathbf{I}\sigma_e^2$, $\mathbf{I}$ a matriz identidade, $\mathbf{G} = A\sigma_u^2$, $\alpha = \dfrac{\sigma_e^2}{\sigma_u^2}$ e σ_e^2 e σ_u^2 as variâncias residual e genética respectivamente) o Sistema de Equações do Modelo Misto de Henderson 2.8 se equivale a (HENDERSON, 1973):

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\alpha\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{pmatrix} \tag{2.9}$$

onde $\mathbf{A}$ é o numerador da matriz de parentesco entre os indivíduos avaliados. Assim, desde que σ_u^2 e σ_e^2 sejam conhecidos, β pode ser estimado e $\mathbf{u}$ predito a partir de 2.9.

O procedimento REML sob modelo misto realiza, simultaneamente, as operações de estimação (BLUE) dos efeitos fixos, estimação de componentes de variância por REML e de predição (BLUP) de valores genéticos. O método REML demanda procedimentos iterativos, realizados através de equações de modelo misto (RESENDE et al., 1996). Esse processo iterativo demanda de recursos computacionais, logo o método utilizado para estimar os componentes de variância foi a função *lme4* do software **R** v.4.2.0 (BATES et al., 2015; R CORE TEAM, 2022).

Considerando $\mathbf{A}$ a matriz identidade e partindo de um valor inicial (estimado ou previamente estabelecido) para estimativas da componente de variância σ_u^2 e para variância do erro σ_e^2, obtém-se as soluções de $\hat{\beta}$ e $\hat{\mathbf{u}}$ a partir da seguinte relação matricial:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{I}\dfrac{\sigma_e^2}{\sigma_u^2} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{pmatrix}. \tag{2.10}$$

Com as soluções de β e $\mathbf{u}$, os componentes de variância são reestimados da seguinte forma:

$$\hat{\sigma}_e^2 = \frac{Y'Y - \hat{\beta}'X'Y - \hat{\mathbf{u}}'Z'Y}{n - r(\mathbf{X})},$$
$$\hat{\sigma}_u^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}} + tr(C_{22})\sigma_e^2}{q}, \tag{2.11}$$

sendo $r(\mathbf{X})$ o posto da matriz $\mathbf{X}$, n a dimensão do vetor $\mathbf{Y}$, q o número de elementos aleatórios a serem preditos e $tr(C_{22})$ o traço correspondente à parte $\mathbf{Z}'\mathbf{Z} + \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I}$ da matriz:

$$V = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \frac{\sigma_e^2}{\sigma_u^2}A^{-}1 \end{pmatrix}^{-1}. \tag{2.12}$$

Essas estimativas de componentes atualizadas são novamente substituídas nas equações de Henderson, e assim por diante até a convergência (LIMA; BUENO FILHO, 2006). Na convergência de $\hat{\sigma}_u^2$ e $\hat{\sigma}_e^2$ a análise dos parâmetros $\hat{\beta}, \hat{\mathbf{u}}$ se torna possível.

Uma abordagem estatística que leva a equações como a 2.8 é aquela que incorpora conhecimento prévio na análise de dados. Embora a derivação do Sistema de Equações de Henderson seja essencialmente de natureza clássica, ela produz os mesmos resultados da análise de Bayes usual com *prioris* pouco informativas (LIMA; BUENO FILHO, 2006; SEARLE; CASSELLA; MCCULLOUCH, 1992).

O procedimento Máxima Verossimilhança Restrita envolve a simplificação da função verossimilhança ou função densidade de probabilidade conjunta das observações em termos de seus resíduos (RESENDE, 2000; SEARLE; CASSELLA; MCCULLOUCH, 1992). Como o modelo 2.7 é normal, encontrar os valores dos parâmetros para os quais a função é máxima resulta na distribuição marginal simétrica de interesse. O procedimento resulta em estimadores análogos (e estimativas muito similares) à média marginal que se obteria da inferência bayesiana, dadas *prioris* pouco informativas (SÖRENSEN; GIANOLA, 2002). Desse modo, por conveniência para as formas usais de análise (caso fossem conhecidas todas as observações, considerando zeros para os censurados e considerando observações perdidas nas censuras) foi utilizado a estimação REML ao invés do amostrador de Gibbs.

3 MATERIAL E MÉTODOS

Nos programas de melhoramento vegetal, é comum a análise de um grande número de tratamentos. Por essa razão, para este estudo foi simulado um experimento em blocos incompletos parcialmente balanceados (PBIB) organizados em látice quadrado triplo ($v = k^2$, $r = 3$, $k = 11$, $b = 33$, $\lambda_1 = 1$ e $\lambda_2 = 0$), delineamento este, frequentemente utilizado no melhoramento genético vegetal. A variável resposta foi simulada para que se pudesse associar à produção de batata-doce, e constitui um vetor de tamanho $n = 363$. A rotina (**R**) utilizada para simulação do experimento e obtenção da variável resposta simulada, encontra-se no Apêndice A.

3.0.1 Modelo da simulação do experimento

Para obtenção dos valores paramétricos foi considerado o seguinte modelo para o experimento:

$$\mathbf{W} = \mathbf{1}\mu + \mathbf{X_R}\rho + \mathbf{X_B}\beta + \mathbf{X_G}\gamma + \mathbf{I}\varepsilon$$

em que: $\mathbf{W}$ é o valor originalmente simulado; $\mathbf{1}$ é o vetor de uns aplicado à constante experimental comum μ; $\mathbf{X_R}$ é a matriz do delineamento das repetições aplicadas ao vetor de efeitos de repetições ρ; $\mathbf{X_B}$ é a matriz do delineamento dos blocos aplicadas ao vetor de efeitos de blocos β; $\mathbf{X_G}$ é a matriz do delineamento dos tratamentos aplicada ao vetor de efeitos genéticos dos tratamentos γ; $\mathbf{I}$ é a matriz de delineamento identidade de ordem n aplicada ao vetor de erros ε com distribuição normal padrão.

Todos os efeitos foram simulados como função do número de ordem do fator, para facilitar a verificação de acertos. Para construir o vetor de efeitos de três repetições foi utilizado o seguinte código **R** (como descrito no Apêndice A):

```
> r <- 3
> (1:r-mean(1:r))/sd(1:r))
[1] -1  0  1
```

O mesmo foi feito com os 33 blocos (crescimento linear dos efeitos, padronizado).

```
> b <- 33
> (1:b-mean(1:b))/sd(1:b))
```

Para construir o vetor de efeitos dos genótipos de tratamentos, foram calculados quantis normais da seguinte forma:

```
> t <- 121
> qnorm((1:121-0.5)/121))
```

Desta forma, garante-se que os efeitos originalmente simulados têm sempre média zero e variância um, embora apenas os efeitos dos genótipos de tratamento sejam claramente normais. Ou seja: $N(0, \sigma_G^2 = 1)$ para os tratamentos e para os demais efeitos, distribuições uniformes discretas.

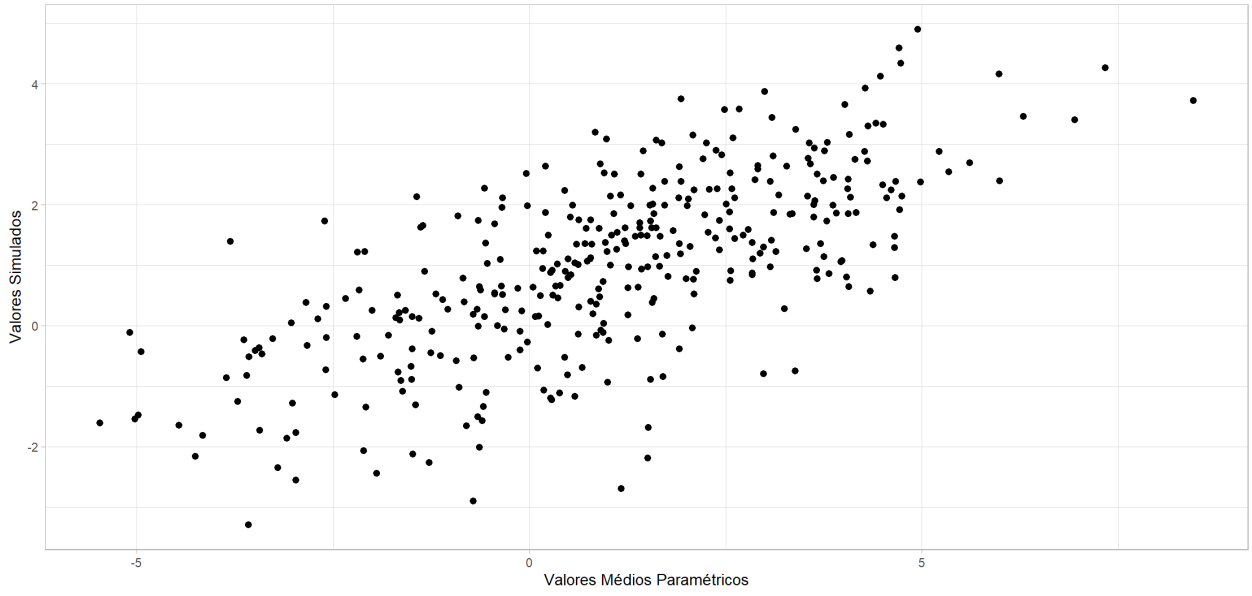
Considerando que o vetor simulado da variável resposta é dado por:

$$\mathbf{w} = E[\mathbf{W}|\rho, \beta, \gamma] + \varepsilon$$

a correlação entre a esperança condicional (valor médio simulado) e o valor das observações foi de $69,49\%$. Note que este valor é muito próximo do limite teórico para a herdabilidade de 50% $(\sqrt{0,5} = 0,7071)$. Em estudos posteriores pode-se investigar o uso de correlações menores aumentando o efeito do erro experimental.

A Figura 3.1 a seguir demonstra uma correlação positiva entre os valores simulados e os valores médios paramétricos.

Figura 3.1 – Gráfico de dispersão entre os valores simulados e sua esperança condicional aos valores de efeitos genéticos e de controle local. A correlação é de $69,49\%$.



Fonte: Do autor (2022).

3.1 Modelo da análise e suas pressuposições

A teoria de modelos mistos, baseia-se em modelos lineares que possuem tanto efeitos fixos como efeitos aleatórios. Desse modo, considerando **w** o vetor de observações da variável resposta produção com dimensão $n = 363$, com ausência de censura. Temos:

$$\mathbf{w} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon$$

em que:

- **X**: é a matriz (nxp) de incidência dos efeitos fixos (que no caso, serão blocos hierarquizados em repetições);
- β: é o vetor de dimensões $(px1)$ das médias de blocos (tomados como efeitos fixos);
- **Z**: é a matriz (nxq) de incidência dos efeitos aleatórios de tratamentos (genótipos);
- **u**: é o vetor de dimensões $(qx1)$ dos parâmetros de efeitos aleatórios dos tratamentos (supostamente, efeitos de genótipos);
- ε: é o vetor de resíduos $(nx1)$.

A partir das definições, os vetores definidos acima seguem as seguintes distribuições:

$$\begin{aligned} \mathbf{w} &\sim N(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}, \mathbf{I}\sigma^2_{\mathbf{e}}) \\ \mathbf{u} &\sim N(\phi,\ \mathbf{I}\sigma^2_{\mathbf{u}}) \\ \varepsilon &\sim N(\phi,\ \mathbf{I}\sigma^2_{\mathbf{e}}) \end{aligned} \tag{3.1}$$

sendo que ϕ representa um vetor nulo de dimensão igual à do vetor considerado.

Para emular medidas estritamente positivas advindas de pesagens, os dados de produção simulados foram transformados em uma variável log-normal a partir da padronização da variável normal *w* inicialmente simulada, conforme segue:

$$\mathbf{y} = e^{\frac{\mathbf{w}-\bar{\mathbf{w}}}{s_w}} \tag{3.2}$$

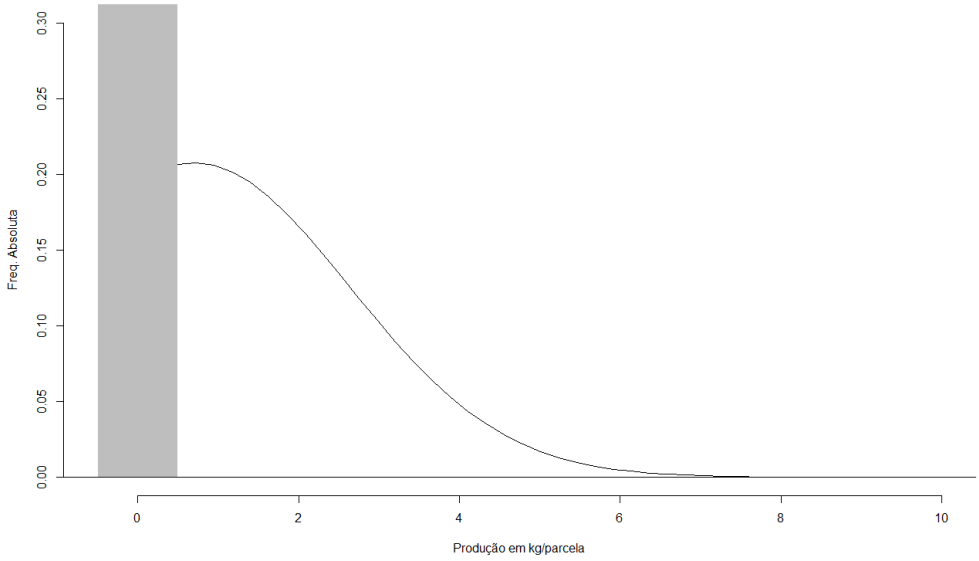
em que $\bar{\mathbf{w}}$ é a média e $\mathbf{s_w}$ o desvio padrão da variável normal originalmente simulada. Essa distribuição log-normal com parâmetros 0 e 1 no núcleo normal tem por média $\mu = e^{0.5} \approx 1,65$ e variância $(e^1 - 1) \times e^1 \approx 4,67$.

3.1.1 Metodologia empregada para se simular a censura à esquerda

No banco de dados de batata-doce apresentado em Silva (2019), cerca de 50% dos dados são censurados, dessa maneira, o mesmo foi feito para banco de dados simulado. Foi estabelecido um valor crítico de detecção abaixo do qual os genótipos seriam desprezados e seu pesos nem mesmo registrados. No cenário 1, aproximadamente 30% dos dados foram censurados, já no cenário 2 a censura foi de aproximadamente 50%, de modo a verificar as consequências desse aumento nas estimativas dos parâmetros genéticos de interesse.

A figura 3.2 a seguir ilustra um experimento com censura (50% de censura) à esquerda em que os dados censurados são substituídos por zeros.

Figura 3.2 – Ilustração de uma censura à esquerda com os dados censurados substituídos por zeros.



Fonte: Do autor (2022).

3.2 Estimação de componentes da variância e valores genéticos

Por conveniência, ao invés de se implementar algoritmos de amostragem Gibbs e inferência bayesiana para a estimação das componentes de variância, foi adotada a predição pelo método da máxima verossimilhança restrita (REML) que dá eventualmente os mesmos resultados que *prioris* não informativas (LIMA; BUENO FILHO, 2006) e está prontamente disponível em diversos pa-

cotes do software **R** v.4.1.0 (BATES et al., 2015; R CORE TEAM, 2022). Utilizamos a biblioteca *lme*4 e a função *lmer*() para os ajustes.

A sintaxe **R** para a análise dos modelos empregando a biblioteca *lme4* é a que se segue:

```
modelo <- lmer(y ~ Bloco + (1|Tratamento) )
```

em que: **y** é o vetor de observações; "Bloco" é o vetor de blocos considerados como de efeito fixo e "(1|Tratamento)" é a sintaxe para associar ao vetor de tratamentos efeitos aleatórios de distribuição normal com uma só média e variância. A rotina (**R**) utilizada para análise dos dados (caso fossem conhecidas todas as observações, considerando zeros para os censurados e considerando observações perdidas nas censuras) pelo método (REML), encontra-se no Apêndice B.

3.3 Previsão condicional em modelagem hierárquica bayesiana

Utilizando de uma generalização do teorema de Bayes juntamente com informações *a priori* sobre os parâmetros e as informações sobre θ fornecidas pelo conjunto de dados (verossimilhança), uma inferência *a posteriori* sobre os parâmetros se torna possível. A seguir estão descritos os componentes necessários para uma análise bayesiana do modelo linear misto (SÖRENSEN, 1996). Note que, para fins de comparação serão utilizadas *prioris* pouco informativas.

O modelo amostral segue duas expressões diferentes:

a) caso não haja censuras e se conheça todos os dados;
b) caso haja censura.

No primeiro caso, o vetor de observações é integralmente observado e a distribuição condicional dos dados **y** é uma normal multivariada:

$$\mathbf{y} \mid \beta, \mathbf{u}, \sigma^2_{\mathbf{u}}, \sigma^2_{\mathbf{e}} \sim N(\mathbf{X}\beta + \mathbf{ZIZ}'\sigma^2_{\mathbf{u}}, \mathbf{I}\sigma^2_{\mathbf{e}}), \tag{3.3}$$

com a eq. (3.3) desempenhando o papel do modelo de amostragem em um contexto bayesiano, sendo o vetor de parâmetros de interesse dado por:

$$\theta = (\beta, \mathbf{u}, \sigma^2_{\mathbf{u}}, \sigma^2_{\mathbf{e}}),$$

sendo: β o efeito de blocos (fixo); u o efeito de tratamento (aleatório); $\sigma^2_{\mathbf{u}}$ a componente de variância genética; $\sigma^2_{\mathbf{e}}$ a componente de variância residual. Note que esta análise apenas será possível para o chamado padrão ouro, sem censura.

No segundo caso, ocorrendo a censura de dados, o vetor de observações $\mathbf{y}$ pode ser posto em forma $\mathbf{y} = (\mathbf{y}^*, \mathbf{y_0})$, onde $\mathbf{y}^* = (y_i^*)_{i=1,2.3...M}$ é um vetor M-dimensional de valores não censurados, e $\mathbf{y_0} = (y_{0i})_{i=M+1,M+2,...,N}$ um vetor $(N-M)$-dimensional de pontos sabidamente censurados (SÖRENSEN; GIANOLA; KORSGAARD, 1998). Foram censurados os valores que estavam abaixo do limite de detecção previamente estabelecido τ e receberam como indicação de censura o valor zero, conforme descrito a seguir:

$$\mathbf{y} = \begin{cases} y_0 = 0, \text{ para } y < \tau \\ \eta = \ln(\mathbf{y}^*), \text{ para } \geq \tau \end{cases} \tag{3.4}$$

Portanto, os valores em $\mathbf{y_0}$ eram abaixo de τ e recebem valor zero, os demais valores em $\mathbf{y}^*$ assumem valor $\eta = \ln(\mathbf{y}^*)$.

3.3.1 Conjunto de parâmetros

Com a censura à esquerda o vetor dos parâmetros de interesse é dado por:

$$\theta = (\beta, \mathbf{u}, \sigma^2_{\mathbf{u}}, \sigma^2_{\mathbf{e}}, \eta)$$

em que:

- β: é o efeito de bloco;
- $\mathbf{u}$: é o efeito de tratamento (genótipo);
- σ^2_u: variância genética;
- σ^2_e: variância residual;
- η: é o valor de $\eta^{-*} = log(\mathbf{y})$ caso ele seja observado, e não entra diretamente na densidade conjunta da verossimilhança mas apenas compõe um parâmetro de mistura no caso de dado censurado η^*.

3.3.2 Verossimilhança

A verossimilhança dos dados completos, incluindo as censuras, é dada por:

$$p(\mathbf{y}|\beta,\mathbf{u},\sigma_u^2,\sigma_e^2) = \prod_{y_i=0} \pi_i \; d_i \prod_{y_i>0} (1-\pi_i) \; d_i$$

em que:

$$d_i = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left\{ -\frac{1}{2\sigma_e^2} (\eta_i - \mathbf{x_i}\beta - \mathbf{z_i}\mathbf{u})'(\eta_i - \mathbf{x_i}\beta - \mathbf{z_i}\mathbf{u}) \right\}, \tag{3.5}$$

é a densidade de η_i na distribuição normal cuja média de bloco é $\mathbf{x_i}\beta$ e o desvio de tratamento é $\mathbf{z_i}\mathbf{u}$. Note que $\mathbf{x_i}$ e $\mathbf{z_i}$ são as linhas das matrizes de delineamento correspondentes à *i*-ésima observação e que, para fins de construir a aproximação pela proporção de dados censurados:

$$\pi_i = \int_{-\infty}^{\tau} f(\eta) d\eta$$

em que (π_i) é a probabilidade do dado ser censurado. Note que o vetor completo para os dados não censurados é $\eta = \log(\mathbf{y}^*)$, ou seja $(\eta - \mathbf{X}\beta - \mathbf{Z}\mathbf{u})$ é o vetor de erros para este caso, suposto normal, que é o mesmo que o vetor de resíduos para a análise dos demais casos em que não se contempla censuras.

3.3.3 Especificação das *prioris*

Para os efeitos fixos β, foram definidas distribuições *a priori* não informativas independentes normais degeneradas, com variância infinita, ou simplesmente uniformes, como se segue:

$$p(\beta) \propto \mathbf{C}$$

em que $\mathbf{C}$ é uma constante.

Do mesmo modo, as componentes da variância terão distribuição correspondentemente pouco informativas. Para obtenção de uma análise conjugada, foram assumidas *prioris* Qui-Quadradas Inversas Escaladas para as componentes da variância, com valores baixos dos hiperparâmetros.

Portanto, foram consideradas para as componentes da variância, $\sigma_u^2 \sim \chi^{-2}(\nu_u = 2; s_u^2 = 5)$ e $\sigma_e^2 \sim \chi^{-2}(\nu = 2; s^2 = 5)$.

Para tais distribuições *a priori* das componentes da variância, a distribuição *a priori* da correlação intraclasse $\frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}$ torna-se aproximadamente uniforme entre 0 e 1.

Para estas análises a matriz de parentesco genético, a qual possui elementos que descrevem os coeficientes das covariâncias entre os efeitos aleatórios genéticos, foi tratada como a matriz identidade ($\mathbf{I}$). Para análise de um banco de dados real pode ser levado em consideração a covariância dos genótipos.

3.3.4 Distribuição *a posteriori* conjunta

Para a obtenção da *posteriori* conjunta de todos os parâmetros, dado que η foi amostrado, fez-se o produto de todas as distribuições *a priori* dadas anteriormente e a verossimilhança do vetor η de observações, ou seja,

$$p(\theta|\eta) \propto p(\eta|\theta)p(\mathbf{u}|\sigma_u^2)p(\sigma_u^2)p(\sigma_e^2).$$

Portanto, a distribuição conjunta *a posteriori*, supondo-se independência entre os parâmetros, é dada pela seguinte expressão:

$$\begin{aligned} p(\beta, \mathbf{u}, \sigma_u^2, \sigma_e^2|\eta) \propto & \prod_{y_i=0} \pi_i \, d_i \prod_{y_i>0} (1-\pi_i) \, d_i \left(\frac{1}{\sigma_u^2}\right)^{\frac{q}{2}} \exp\left\{-\frac{1}{2}\frac{\mathbf{u}'A^1\mathbf{u}}{\sigma_u^2}\right\} (\sigma_u^2)^{-\left(\frac{\nu_u}{2}+1\right)} \\ & \exp\left\{-\frac{\nu_u s_u^2}{2\sigma_u^2}\right\} (\sigma_e^2)^{-\left(\frac{\nu}{2}+1\right)} \exp\left\{-\frac{\nu s^2}{2\sigma_e^2}\right\}. \end{aligned} \tag{3.6}$$

Para implementação do algoritmo do Amostrador de Gibbs, foram obtidas as distribuições *a posteriori* condicionais completas para cada parâmetro, a partir da distribuição conjunta 3.6.

3.3.5 Amostra condicional do vetor η

Seja o vetor original de observações dado por:

$$\eta = \begin{cases} \ln(\mathbf{y}), \text{ para } \mathbf{y}^* \\ \eta^*, \text{ para } \mathbf{y_0} \end{cases} \tag{3.7}$$

ou seja, caso o valor seja censurado, η^* é o conjunto de pseudo-observações η_i^*, que serão amostradas da distribuição normal truncada em τ_i do seguinte modo:

$$\eta^* | \beta, \mathbf{u}, \sigma_u^2, \sigma_e^2 \ \sim \ N_{[-\infty,\tau]}(\mathbf{X}\beta + \mathbf{Zu}, \mathbf{I}\sigma_e^2).$$

Portanto, supondo que se conheça $(\beta, \mathbf{u}, \sigma_u^2, \sigma_e^2)$, η^* pode ser amostrado, conhecidos o genótipo e o bloco da parcela e seus respectivos efeitos.

Calculada a probabilidade π_i^* de se encontrar o valor τ na distribuição normal padrão correspondente a η^*, pode-se amostrar este último de uma distribuição normal com média $(\mathbf{X}\beta + \mathbf{Zu})$ e variância σ_e^2, truncada em τ. Nas implementações, foi amostrado um valor ω_i de uma distribuição uniforme $[0,1]$ e a partir deste, foi obtido o quantil normal para a probabilidade $\omega_i \pi_i^*$.

Desta forma, η^* foi amostrado de da distribuição multivariada normal truncada entre $-\infty$ até τ, com média $(\mathbf{X}\beta + \mathbf{Zu})$ e variância $\mathbf{I}\sigma_e^2$.

Com o vetor completo η, segue-se a análise bayesiana usual do modelo de blocos, em que apenas usamos o amostrador de Gibbs por simplicidade de implementação. No que apresentamos a seguir, para as demais distribuições condicionais completas, usamos η no lugar em que estaria $\mathbf{y}$ no desenvolvimento de Lima e Bueno Filho (2006).

A rotina (**R**) utilizada para amostragem condicional dos dados censurados junto a implementação da Amostragem Gibbs das distribuições condicionais completas *a posteriori*, encontra-se no Apêndice C.

3.3.6 Distribuições condicionais completas para os demais parâmetros

A distribuição condicional completa *a posteriori* para os efeitos fixos, segundo a equação 3.6, é dada por:

$$
\begin{aligned}
p(\beta|\mathbf{u},\sigma_u^2,\sigma_e^2,\eta) &\propto \exp\left\{-\frac{1}{2\sigma_e^2}(\eta-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})'(\eta-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left[(\eta-\mathbf{Z}\mathbf{u})'-(\mathbf{X}\beta)'\right]\left[(\eta-\mathbf{Z}\mathbf{u})-(\mathbf{X}\beta)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left[(\eta-\mathbf{Z}\mathbf{u})'(\eta-\mathbf{Z}\mathbf{u})-(\eta-\mathbf{Z}\mathbf{u})'(\mathbf{X}\beta)\right.\right. \\
&\quad \left.\left.-(\mathbf{X}\beta)'(\eta-\mathbf{Z}\mathbf{u})+(\mathbf{X}\beta)'(\mathbf{X}\beta)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left[-(\eta-\mathbf{Z}\mathbf{u})'\mathbf{X}\beta-\beta'\mathbf{X}'(\eta-\mathbf{Z}\mathbf{u})+\beta'\mathbf{X}'(\mathbf{X}\beta)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left[\beta-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\eta-\mathbf{Z}\mathbf{u})\right]'(\mathbf{X}'\mathbf{X})\left[\beta-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\eta-\mathbf{Z}\mathbf{u})\right]\right\} \\
&\quad \exp\left\{-\frac{1}{2\sigma_e^2}\left[-(\eta-\mathbf{Z}\beta)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\eta-\mathbf{Z}\mathbf{u})\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left[\beta-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\eta-\mathbf{Z}\mathbf{u})\right]'(\mathbf{X}'\mathbf{X})\left[\beta-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\eta-\mathbf{Z}\mathbf{u})\right]\right\}.
\end{aligned}
$$

Portanto:

$$
\beta|(\mathbf{u},\sigma_u^2,\sigma_e^2,\eta) \sim N\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\eta-\mathbf{Z}\mathbf{u}),\sigma_e^2(\mathbf{X}'\mathbf{X})^{-1}\right).
$$

De forma análoga, *a posteriori* condicional completa para os efeitos aleatórios (tratamentos) pode ser obtida, ou seja, dada *a posteriori* conjunta 3.6, tem-se:

$$
p(\mathbf{u}|\beta,\sigma_u^2,\sigma_e^2,\eta) \propto \exp\left\{-\frac{1}{2\sigma_e^2}\left(\frac{\mathbf{u}'\mathbf{A}^{-1}\sigma_e^2\mathbf{u}}{\sigma_u^2}+(\eta-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})'(\eta-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})\right)\right\},
$$

sendo $\delta=\dfrac{\sigma_e^2}{\sigma_u^2}$, tem-se:

$$
\begin{aligned}
p(\mathbf{u}|\beta,\sigma_u^2,\sigma_e^2,\eta) &\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left(\mathbf{u}'\delta\mathbf{A}^{-1}\mathbf{u}+(\eta-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})'(\eta-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left(\mathbf{u}'\delta\mathbf{A}^{-1}\mathbf{u}+\left[(\eta-\mathbf{X}\beta)'-(\mathbf{Z}\mathbf{u})'\right]\left[(\eta-\mathbf{X}\beta)-(\mathbf{Z}\mathbf{u})\right]\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left(\mathbf{u}'\delta\mathbf{A}^{-1}\mathbf{u}+(\eta-\mathbf{X}\beta)'(\eta-\mathbf{X}\beta)-(\eta-\mathbf{X}\beta)'(\mathbf{Z}\mathbf{u})\right.\right. \\
&\quad \left.\left.-(\mathbf{Z}\mathbf{u})'(\eta-\mathbf{X}\beta)+(\mathbf{Z}\mathbf{u})'(\mathbf{Z}\mathbf{u})\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left(\mathbf{u}'\delta\mathbf{A}^{-1}\mathbf{u}-(\eta-\mathbf{X}\beta)'(\mathbf{Z}\mathbf{u})-(\mathbf{Z}\mathbf{u})'(\eta-\mathbf{X}\beta)+(\mathbf{Z}\mathbf{u})'(\mathbf{Z}\mathbf{u})\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left[\mathbf{u}-(\mathbf{Z}'\mathbf{Z}+\delta\mathbf{A}^{-1})^{-1}\mathbf{Z}'(\eta-\mathbf{X}\beta)\right]'(\mathbf{Z}\mathbf{Z}'+\delta\mathbf{A}^{-1})\right. \\
&\quad \left.\left[\mathbf{u}-(\mathbf{Z}\mathbf{Z}'+\delta\mathbf{A}^{-1})^{-1}\mathbf{Z}'(\eta-\mathbf{X}\beta)\right]\right\}.
\end{aligned}
$$

Portanto:

$$
\mathbf{u}|(\beta,\sigma_u^2,\sigma_e^2,\eta) \sim N\left((\mathbf{Z}\mathbf{Z}'+\delta\mathbf{A}^{-1})^{-1}\mathbf{Z}'(\eta-\mathbf{X}\beta),(\mathbf{Z}\mathbf{Z}'+\delta\mathbf{A}^{-1})^{-1}\sigma_e^2\right).
$$

A distribuição condicional completa para a componente da variância σ_e^2, em 3.6, é dada por:

$$
\begin{aligned}
p(\sigma_e^2|\beta,\sigma_u^2,\eta) &\propto \underbrace{(\sigma_e^2)^{-\left(\frac{n}{2}\right)}\exp\left\{-\frac{1}{2\sigma_e^2}\left[(\eta-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})'(\eta-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})\right]\right\}}_{\text{porção da Normal Multivariada para } \eta} \\
&\quad \underbrace{(\sigma_e^2)^{-\left(\frac{\nu}{2}+1\right)}\exp\left\{-\frac{\nu s^2}{2\sigma_e^2}\right\}}_{\text{porção da } \chi^{-2}_{escalada} \text{ para } \sigma_e^2} \\
&\propto (\sigma_e^2)^{-\left(\frac{n+\nu}{2}+1\right)}\exp\left\{-\frac{1}{2\sigma_e^2}\left[(\eta-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})'(\eta-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})+\nu s^2\right]\frac{(n+\nu)}{(n+\nu)}\right\},
\end{aligned}
$$

ou seja, uma Qui-Quadrado Inversa Escalada para a componente da variância, com $n+\nu$ graus de liberdade e parâmetro de escala dado por:

$$\sigma_e^2|(\beta,\mathbf{u},\sigma_u^2,\eta) \sim \chi^{-2}\left(n+v,\frac{(\eta-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})'(\eta-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})+vs^2}{n+v}\right).$$

Da mesma maneira, para a componente da variância σ_u^2, pode-se obter sua distribuição condicional completa por meio de 3.6, como se segue:

$$p(\sigma_u^2|\beta,\mathbf{u},\sigma_e^2,\eta) \propto (\sigma_u^2)^{-\left(\frac{q+v_u}{2}+1\right)}\exp\left\{-\frac{1}{2\sigma_u^2}\left(\frac{\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}+v_u s_u^2}{q+v_u}\right)(q+v_u)\right\},$$

que é também uma Qui-Quadrado Inversa Escalada com $q+\nu_u$ graus de liberdade e parâmetro de escala dado por:

$$\sigma_u^2|(\beta,\mathbf{u},\sigma_e^2,\eta) \sim \chi^{-2}\left(q+v_u,\frac{\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}+v_u s_u^2}{q+v_u}\right).$$

3.4 Análises

Para as análises REML foram verificadas a acurácia seletiva, calculada como a proporção dos sets (7) maiores valores paramétricos de genótipos incluída entre as sete melhores estimativas. Foram calculadas também correlações para indicar esta acurácia. O mesmo foi feito com os genótipos de valores estimados baixos e respectivos ordenamentos e efeitos paramétricos. Foi também avaliada a precisão e o viés na estimação de parâmetros genéticos como as componentes da variância (calculadas usando distribuições qui-quadradas com os graus de liberdade estimados segundo a aproximação de Satterthwaite (1946), implementada na biblioteca *lmerTest* do **R**) e as herdabilidades na seleção entre parcelas $\left(h^2=\frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}\right)$ e entre médias de tratamentos experimentais $\left(h_m^2=\frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2/r}\right)$. Para obter as distribuições e estimativas de herdabilidades foram tomadas amostras das distribuições das componentes da variância, consideradas independentes.

No caso da análise proposta ("PC"), além de se obter distribuições *a posteriori* para todos os parâmetros acima mencionados, foram também calculadas as correlações entre os valores originalmente simulados para as observações que foram censuradas e suas respectivas médias *a posteriori* obtidos com a amostragem Gibbs. Para as análises anteriores este valor é zero e para haver algum indício de melhora é preciso verificar correlações positivas. Por fim, em cada uma das formas de

análise, foi calculada a correlação entre as predições dos efeitos genéticos dos tratamentos e seus respectivos valores paramétricos.

3.5 Implementação

Todas as análises estatísticas desenvolvidas nesta seção foram realizadas utilizando-se o software **R** v.4.2.0 (R CORE TEAM, 2022).

4 RESULTADOS E DISCUSSÃO

Na presente seção serão apresentados os resultados da simulação em dois cenários, quais sejam:

a) taxa de censura moderada: $\sim 30\%$ com $\sigma_e^2 = 3$ e $\sigma_u^2 = 1$;

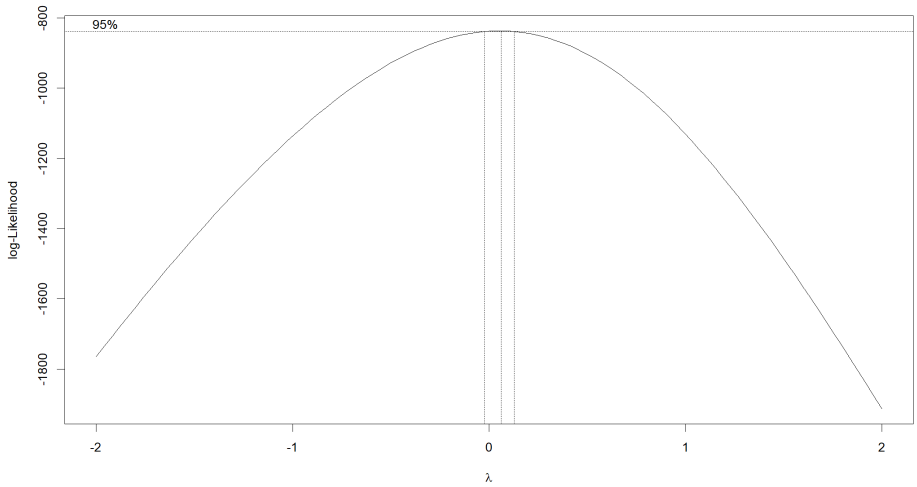
b) taxa de censura alta: $\sim 50\%$ com $\sigma_e^2 = 3$ e $\sigma_u^2 = 1$.

São a seguir apresentados os resultados de quatro formas de análise: Dados completos "DC" (caso fossem conhecidas todas as observações); Censura Zero "C0" (considerando zero para as censuras); Censura à Esquerda "CE" (considerando observações perdidas nas censuras); Previsão Condicional "PC" (substituindo os valores censurados por valores estimados).

4.1 Cenário: supondo que se conheça os dados completos ("DC"):

Para esta análise, suposta como padrão ouro, a transformação não linear de Box e Cox (1964) foi utilizada para melhorar a aproximação dos dados experimentais à distribuição normal (Figura 4.1). Portanto, para a variável produção **y** sem a censura de dados, foi utilizada a transformação logarítmica (o que equivale à suposição de que **y** tem distribuição log-normal).

Figura 4.1 – Transformação de Box e Cox aplicada a **y**, com $\lambda = 0$ (Cenário: DC).



Fonte: Do autor (2022).

Com as componentes da variância obtidas é possível obter estimativas da herdabilidade para a seleção entre parcelas $\left(h^2 = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}\right)$ e para a seleção entre médias de tratamentos $\left(h_m^2 = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2/r}\right)$.

Os resumos destes parâmetros genéticos são apresentados na 4.1. Observa-se que o experimento estimou de forma adequada as componentes da variância genética e do erro experimental, bem como as herdabilidades.

Tabela 4.1 – Estimativas das componentes de variância e herdabilidades do caso "DC", com seus respectivos IC (95%) e valores paramétricos usando os dados transformados.

Parâmetro	**Valor Paramétrico**	**Estimativa**	**IC**	
			LI	**LS**
σ_u^2	1,00	1,07	0,84	1,39
σ_e^2	3,00	2,98	2,54	3,54
h^2	0,25	0,26	0,21	0,32
h_m^2	0,50	0,52	0,44	0,59

Fonte: Do autor (2022).

O limite máximo esperado para correlação genética dos efeitos de tratamentos com seus efeitos paramétricos é de $\sqrt{h_m^2}.100 = \sqrt{0,50}.100 = 70,71\%$. O valor observado, ou seja, a correlação entre os BLUP dos efeitos de tratamentos da análise "DC" e seus respectivos valores simulados foi de $\hat{\rho} = 69,29\%$, sendo significativamente não nula pelo teste "t de Student", o que pode ser visualizado na Tabela 4.2 e na Figura 4.2. A porcentagem de recuperação da informação genética para esta análise foi de:

$$\frac{\hat{\rho}}{\sqrt{h_m^2}.100} = \frac{69,29\%}{70,71\%} = 97,99\%.$$

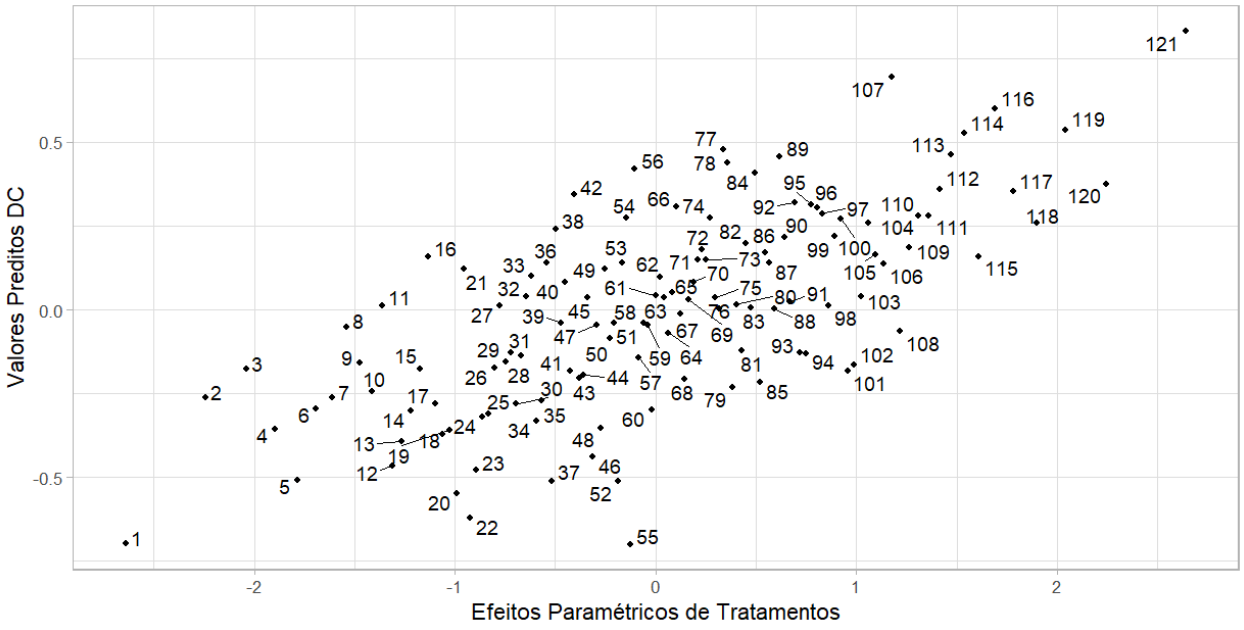
Com o propósito de comparação com as demais formas de análises "D0", "CE" e "PC", é apresentado na Tabela 4.2 a seguir, as correlações de Pearson e Sperman considerando que todos os dados foram observados.

Tabela 4.2 – Correlações de Pearson e Sperman dos efeitos de tratamentos com seus respectivos efeitos paramétricos para o caso "DC". Intervalo de confiança (95%) usando a aproximação *t* de Student para correlações Pearson e Sperman.

Método	**Correlação (%)**	**valor-*p***	**IC**	
			LI	**LS**
Pearson	69,29	2,2e-16	58,79	77,55
Spearman	68,17	2,2e-16	60,60	75,75

Fonte: Do autor (2022).

Figura 4.2 – Gráfico de dispersão das predições dos efeitos de tratamentos para o caso "DC" (Cenário 1) e seus respectivos valores paramétricos ($\hat{\rho} = 69,29\%$).



Fonte: Do autor (2022).

Encontram-se na Tabela 4.3, os resumos das estimativas (BLUP e respectivos erros-padrão) dos 7 piores e 7 melhores efeitos de tratamentos da análise "DC". Os efeitos estão ordenados em ordem crescente do BLUP, espera-se valores baixos para os primeiro grupo (idealmente, de 1 a 7) e valores altos para o segundo grupo (idealmente, de 115 a 121). Nota-se que os maiores problemas estão entre os valores menores, que na escala original tendem a ser mais comprimidos.

Tabela 4.3 – BLUP para os efeitos de tratamentos considerando todas as observações conhecidas, e seus respectivos erros-padrões, IC (95%) e valores paramétricos.

Tratamento	Valor Paramétrico	BLUP	SD_{BLUP}	LI_{BLUP}	LS_{BLUP}
55	-0,12	-1,70	0,72	-3,10	-0,29
1	-2,64	-1,69	0,72	-3,10	-0,29
22	-0,92	-1,51	0,72	-2,91	-0,10
20	-0,99	-1,33	0,72	-2,74	0,07
52	-0,19	-1,25	0,72	-2,65	0,16
37	-0,52	-1,24	0,72	-2,65	0,16
5	-1,78	-1,23	0,72	-2,64	0,17
⋮	⋮	⋮	⋮	⋮	⋮
113	1,47	1,12	0,72	-0,29	2,53
77	0,34	1,16	0,72	-0,25	2,56
114	1,54	1,28	0,72	-0,13	2,68
119	2,04	1,30	0,72	-0,10	2,71
116	**1,69**	**1,46**	**0,72**	**0,05**	**2,86**
107	**1,18**	**1,69**	**0,72**	**0,28**	**3,09**
121	**2,64**	**2,02**	**0,72**	**0,61**	**3,42**

Fonte: Do autor (2022).

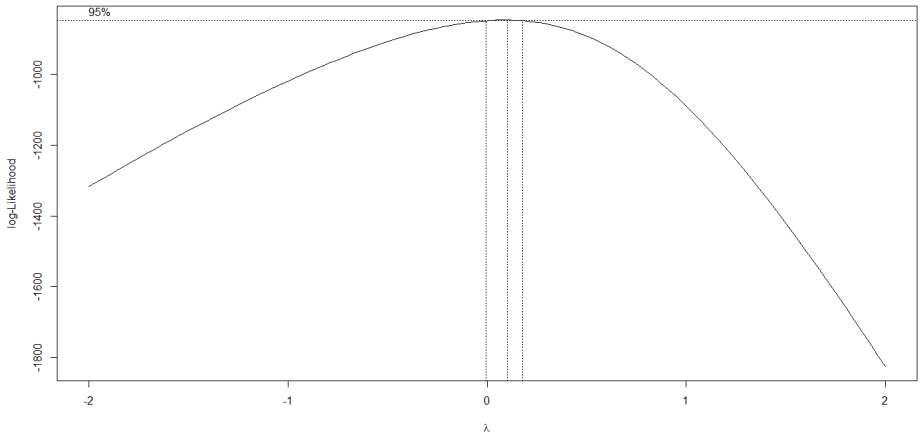
Vale ressaltar que, mesmo com todos os dados observados, para os melhores tratamentos somente os valores em negrito na Tabela 4.3 apresentaram resultados acima de zero (IC não contém o zero). Portanto, somente os três tratamentos 116, 107 e 121 (3/7) seriam bem selecionados, ou seja, entrariam no grupo dos genótipos de elite.

4.2 Cenário 1: substituindo censuras por 0 ("C0")

Para a variável produção **y** com a censura de dados substituída por zeros, foi utilizada a transformação não linear de Box e Cox (1964) para melhorar a aproximação dos dados experimentais à distribuição normal. Como esta metodologia indicava a proximidade do $\lambda = 0$, conforme

pode ser verificado na figura 4.3 foi escolhida a transformação logarítmica com $\mathbf{y}+\mathbf{0},\mathbf{3}$, constante escolhida para eliminar os zeros do conjunto de dados e permitir o cálculo do logaritmo natural.

Figura 4.3 – Transformação de Box e Cox aplicada a $\mathbf{y}+\mathbf{0},\mathbf{3}$, com $\lambda = 0$ (Cenário 1: C0).



Fonte: Do autor (2022).

Os resumos dos parâmetros genéticos para esta análise são apresentados na Tabela 4.4. Observa-se estimativas ligeiramente superestimadas das componentes da variância, embora as herdabilidades sejam bem estimadas. Este resultado é consistente com o que se espera, sendo esta uma abordagem utilizada com muita frequência, embora os valores reais não sejam tão extremos quanto o zero.

Tabela 4.4 – Estimativas das componentes de variância e herdabilidades do caso "C0" (Cenário 1) , com seus respectivos IC (95%) e valores paramétricos usando os dados transformados.

Parâmetro	**Valor Paramétrico**	**Estimativa**	**IC**	
			LI	**LS**
σ_u^2	1,00	1,84	1,45	2,40
σ_e^2	3,00	5,67	4,84	6,73
h^2	0,25	0,24	0,19	0,30
h_m^2	0,50	0,49	0,42	0,56

Fonte: Do autor (2022).

As predições dos efeitos de tratamentos substituindo todos os valores faltantes por zeros, apresentou correlação significativamente não nula $\hat{\rho} = 69,06\%$ com os efeitos de tratamentos paramétricos. Correlação esta, semelhante a da análise "DC", o que pode ser verificado pela Tabela 4.5 e visualizado pela Figura 4.4. A forma de análise "C0" também apresentou boa porcentagem de recuperação da informação genética, sendo de:

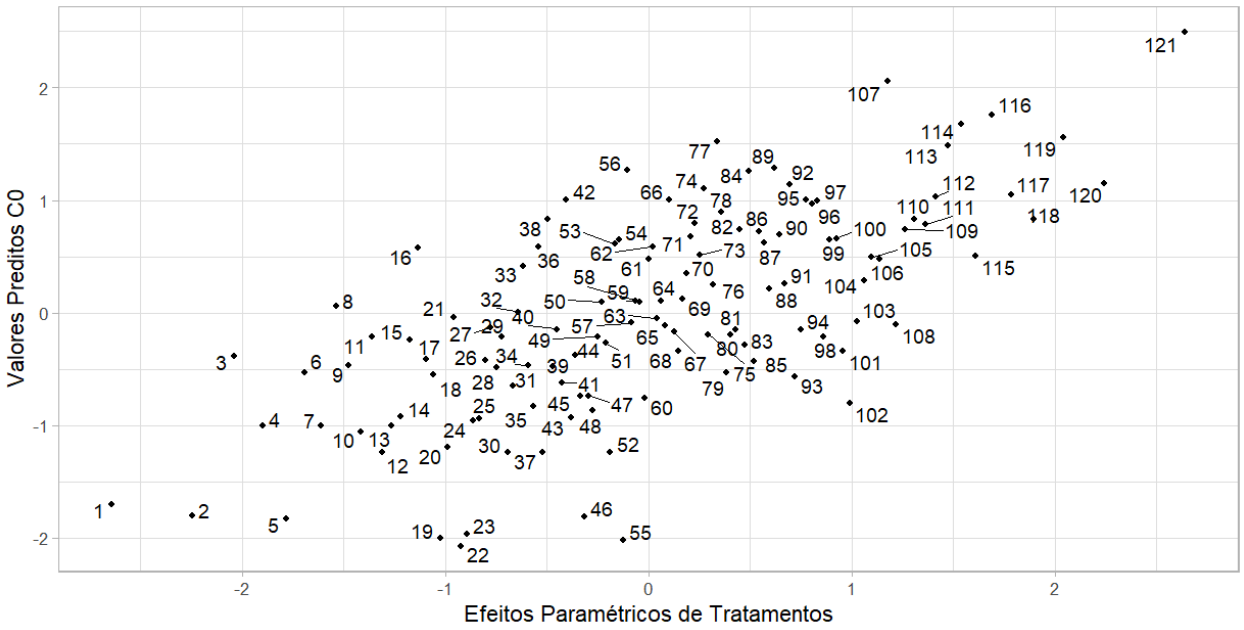
$$\frac{\hat{\rho}}{\sqrt{h_m^2}} = \frac{69,06\%}{70,71\%} = 97,66\%.$$

Tabela 4.5 – Correlações de Pearson e Sperman dos efeitos de tratamentos com seus respectivos efeitos paramétricos para o caso "C0" (Cenário 1). Intervalo de confiança (95%) usando a aproximação *t* de Student para correlações Pearson e Sperman.

Método	**Correlação (%)**	**valor-*p***	**IC**	
			LI	**LS**
Pearson	69,06	2,2e-16	58,41	77,37
Spearman	68,44	2,2e-16	60,90	75,99

Fonte: Do autor (2022).

Figura 4.4 – Gráfico de dispersão das predições dos efeitos de tratamentos para o caso "C0" e seus respectivos valores paramétricos ($\hat{\rho} = 69,06\%$).



Fonte: Do autor (2022).

Os resumos das estimativas (BLUP e respectivos erros-padrão) dos 7 piores e 7 melhores efeitos de tratamentos da análise "C0" é apresentado na Tabela 4.6. Assim como a análise análise "DC", os tratamentos estão ordenados em ordem crescente dos BLUP. Valores baixos para os primeiro grupo (idealmente, de 1 a 7) e valores altos para o segundo grupo (idealmente, de 115 a 121) são esperados. Assim como na Tabela 4.3 da análise "DC", para a análise "C0" na Tabela 4.6, nota-se que os maiores problemas estão entre os menores valores. Além dos tratamento 2 e 5, o mais próximo do grupo (1 a 7) é o tratamento 19.

Tabela 4.6 – BLUP para os efeitos de tratamentos com a perda dos dados censurados, e seus respectivos erros-padrões, IC (95%) e valores paramétricos.

Tratamento	Valor Paramétrico	BLUP	SD_{BLUP}	LI_{BLUP}	LS_{BLUP}
22	-0,92	-2,07	0,97	-3,96	-0,18
55	-0,12	-2,02	0,97	-3,91	-0,12
19	-1,02	-2,00	0,97	-3,90	-0,11
23	-0,89	-1,97	0,97	-3,86	-0,07
5	-1,78	-1,83	0,97	-3,72	0,06
46	-0,32	-1,81	0,97	-3,70	0,08
2	-2,24	-1,81	0,97	-3,70	0,09
⋮	⋮	⋮	⋮	⋮	⋮
113	1,47	1,49	0,97	-0,40	3,38
77	0,34	1,52	0,97	-0,37	3,41
119	2,04	1,56	0,97	-0,33	3,45
114	1,54	1,68	0,97	-0,22	3,57
116	1,69	1,76	0,97	-0,13	3,65
107	**1,18**	**2,06**	**0,97**	**0,16**	**3,95**
121	**2,64**	**2,49**	**0,97**	**0,60**	**4,38**

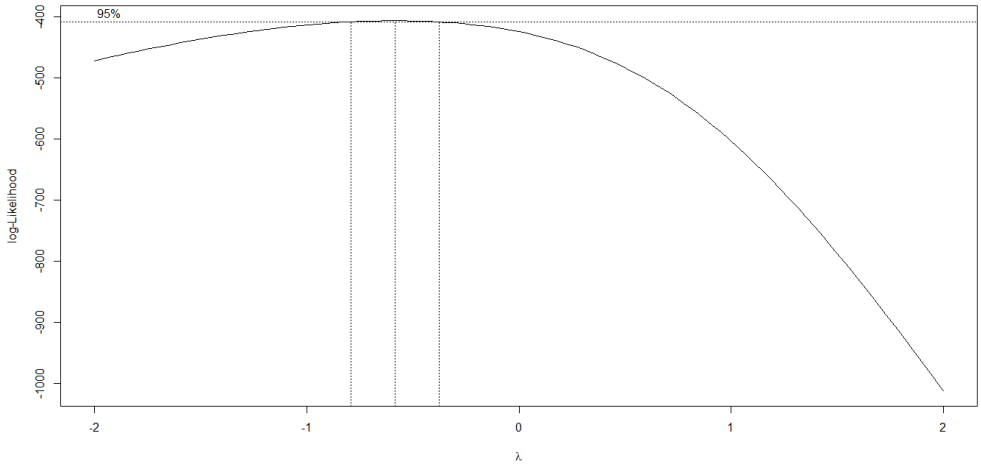
Fonte: Do autor (2022).

A análise "C0" para seleção de genótipos de elite apresentou-se ligeiramente inferior que a análise "DC", somente os tratamentos 107 e 121 (2/7) estão acima de zero, ou seja, sem alternância de sinal em seus respectivos IC.

4.3 Cenário 1: eliminando censuras à esquerda ("CE")

Considerando a perda de dados, as observações censuradas foram retiradas da variável produção **y**. O modelo precisou de uma transformação não linear Box e Cox (1964), seguida de uma correção na escala (Figura 4.5). Portanto, para cada observação da variável produção foi obtida sua transformação $Pt = \frac{(\mathbf{y}+\mathbf{0,2})^{\lambda}-1}{\lambda}$.

Figura 4.5 – Transformação de Box e Cox aplicada a $\mathbf{y}+\mathbf{0,2}$, com $\lambda \neq 0$ (Cenário 1: CE).



Fonte: Do autor (2022).

A Tabela 4.7 contém os resumos dos parâmetros genéticos para a análise onde os valores abaixo do limite de detecção são desconsiderados. Uma estimativa adequada da variância do erro experimental é observada, assim como uma certa subestimação da componente da variância genética e das herdabilidades.

Tabela 4.7 – Estimativas das componentes de variância e herdabilidades do caso "CE" (Cenário 1) , com seus respectivos IC (95%) e valores paramétricos usando os dados transformados.

Parâmetro	**Valor Paramétrico**	**Estimativa**	**IC**	
			LI	**LS**
σ_u^2	1,00	0,53	0,42	0,69
σ_e^2	3,00	3,08	2,56	3,77
h^2	0,25	0,15	0,11	0,19
h_m^2	0,50	0,27	0,21	0,33

Fonte: Do autor (2022).

Devido a remoção dos dados censurados, 7 dos tratamentos $(1,2,5,19,22,23,55)$ foram desconsiderados da análise por não aparecerem com dados em qualquer unidade experimental. Isto representa $5,8\%$ dos tratamentos que ficariam sem estimativas. Embora estes sejam tratamentos ruins, deve-se verificar se a retirada desses tratamentos beneficia de alguma forma a análise.

Com a remoção dos dados censurados, as estimativas (BLUP e respectivos erros-padrão) dos efeitos de tratamento apresentou correlação significativamente não nula $\hat{\rho} = 52,78\%$ com seus respectivos valores simulados, o que pode ser visualizado pela Tabela 4.8.

Tabela 4.8 – Correlações de Pearson e Sperman dos efeitos de tratamentos com seus respectivos efeitos paramétricos para o caso "CE". Intervalo de confiança (95%) usando a aproximação *t* de Student para correlações Pearson e Sperman.

Método	**Correlação (%)**	**valor-*p***	**IC**	
			LI	**LS**
Pearson	52,78	1,9e-09	38,01	64,92
Spearman	48,92	5,9e-08	38,14	59,69

Fonte: Do autor (2022).

Quando comparado com os dados completos e os dados censurados substituídos por zeros, a simples retirada dos dados censurados ocasionou diminuição da correlação dos efeitos de tratamentos com seus respectivos valores paramétricos. Dessa maneira, ignorar esses dados em uma situação com censura moderada, quanto ao efeito de seleção, mostrou-se o pior cenário entre

as demais alternativas até aqui analisadas. Consequentemente, a porcentagem de recuperação da informação genética para esta análise foi a menor comparada com "DC" e "C0", sendo de:

$$\frac{\hat{\rho}}{\sqrt{h_m^2}} = \frac{52,78\%}{70,71\%} = 74,64\%.$$

É exibido na Tabela 4.9, os resumos das estimativas (BLUP e respectivos erros-padrão) dos 7 piores (excluindo os que não foram estimados, que podem ser considerados ainda piores) e 7 melhores efeitos de tratamentos da análise "CE".

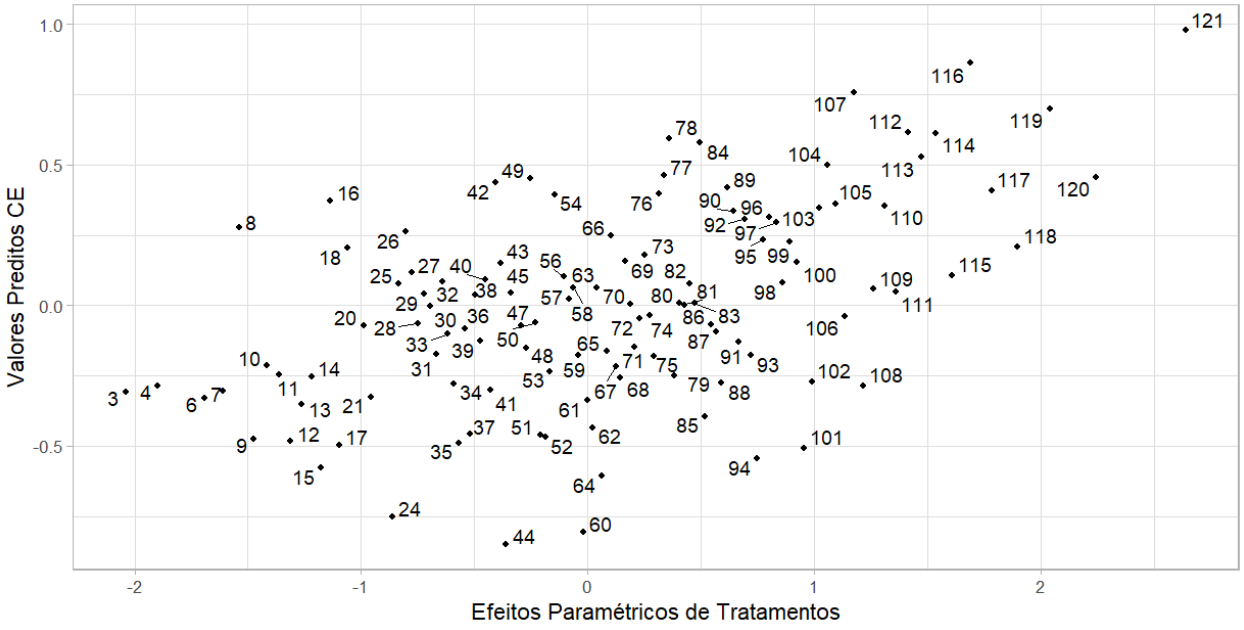
Tabela 4.9 – BLUP para os efeitos de tratamentos com a perda dos dados censurados, e seus respectivos erros-padrões, IC (95%) e valores paramétricos.

Tratamento	Valor Paramétrico	BLUP	SD_{BLUP}	LI_{BLUP}	LS_{BLUP}
44	-0,36	-0,85	0,59	-2,01	0,31
60	-0,02	-0,81	0,63	-2,04	0,42
24	-0,86	-0,75	0,63	-1,98	0,48
64	0,06	-0,61	0,59	-1,76	0,55
15	-1,18	-0,58	0,63	-1,81	0,65
94	0,75	-0,54	0,59	-1,70	0,61
101	0,96	-0,51	0,59	-1,66	0,65
⋮	⋮	⋮	⋮	⋮	⋮
78	0,36	0,59	0,63	-0,64	1,82
114	1,54	0,61	0,59	-0,55	1,77
112	1,42	0,62	0,59	-0,54	1,77
119	2,04	0,70	0,59	-0,46	1,86
107	1,18	0,76	0,59	-0,40	1,91
116	1,69	0,86	0,59	-0,30	2,02
121	2,64	0,98	0,59	-0,18	2,14

Fonte: Do autor (2022).

Quando comparada com a Tabela 4.3 da análise "DC", uma alteração na ordem dos 7 melhores tratamentos é observada, sendo ainda mais considerável na ordem dos 7 piores. Como referência é preciso lembrar que os valores esperados para os postos de tratamento seriam de 8 a 14, e o mais próximo disso é o 15, por outro lado, entre os melhores, apenas o 78 é bem distante do grupo. As alterações para os demais tratamentos é visualizada na Figura 4.6.

Figura 4.6 – Gráfico de dispersão das predições dos efeitos de tratamentos "CE" e seus respectivos valores paramétricos ($\hat{\rho} = 52,78\%$).



Fonte: Do autor (2022).

Para seleção dos genótipos de elite, a análise "CE" apresentou o pior cenário, já que nenhum dos 7 melhores tratamentos na Tabela 4.9 apresentaram resultados maiores que zero.

4.4 Cenário 1: previsão condicional dos dados censurados ("PC")

Para o caso dos dados censurados em si, a análise dos valores de produção η obtidos através da amostragem de Gibbs, apresenta-se os resumos da distribuição marginal *a posteriori* de algumas destas observações estimadas na Tabela 4.10.

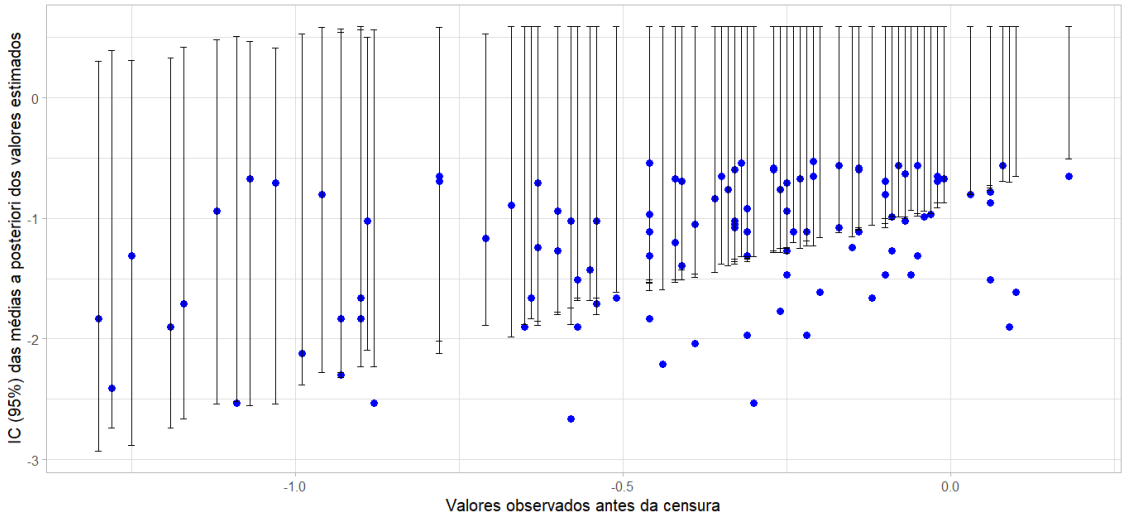
Tabela 4.10 – Resumos da distribuição *a posteriori* obtida pela amostragem Gibbs em comparação aos valores paramétricos gerados na simulação.

Censuras	Valor gerado	Média	SD	LI	LS
1	-1,90	-0,65	0,69	-1,88	0,59
2	-0,71	-0,63	0,69	-1,85	0,59
3	-1,66	-0,64	0,68	-1,83	0,59
4	-1,24	-0,63	0,70	-1,89	0,59
5	-1,71	-0,54	0,63	-1,66	0,59
6	-0,94	-0,60	0,66	-1,80	0,59
7	-1,02	-0,58	0,65	-1,74	0,59
⋮	⋮	⋮	⋮	⋮	⋮
101	-0,69	-0,10	0,49	-1,04	0,59
102	-1,24	-0,15	0,53	-1,15	0,59
103	-1,31	-0,05	0,48	-0,96	0,59
104	-0,56	-0,05	0,48	-0,98	0,59
105	-1,02	-0,07	0,49	-1,00	0,59
106	-0,99	-0,04	0,48	-0,94	0,59
107	-0,97	-0,03	0,48	-0,96	0,59

Fonte: Do autor (2022).

Verifica-se na Figura 4.7, que grande parte dos intervalos de credibilidade (95%) da média marginal *a posteriori* dos valores estimados contém os valores conhecidos da variável produção gerados na simulação, antes do processo de censura. Este é um primeiro indício de que o processo de inferência escolhido é interessante para a análise prática.

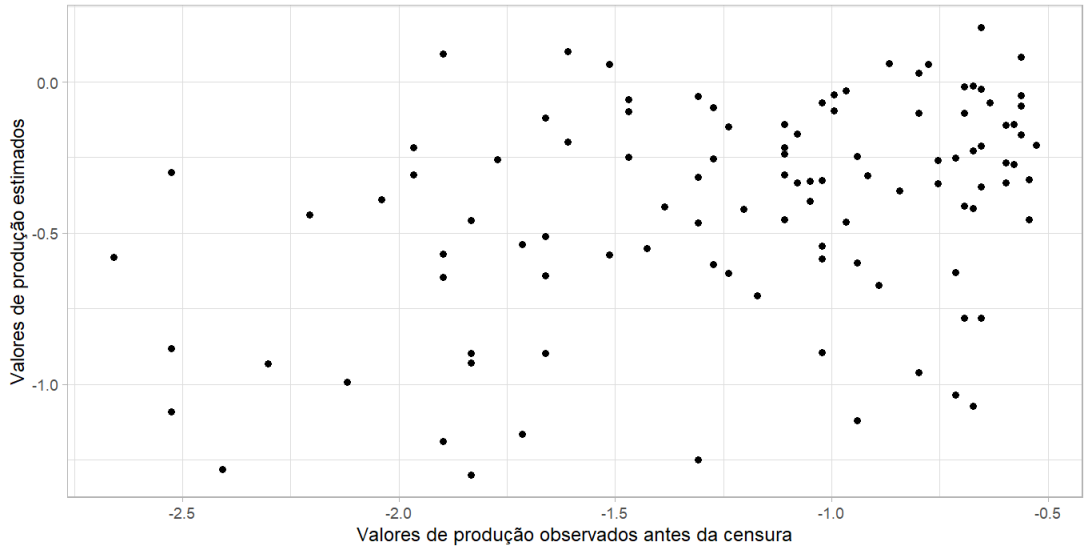
Figura 4.7 – Intervalos de credibilidade (95%) dos 107 valores de produção estimados em relação aos valores conhecidos da variável produção gerados na simulação, antes do processo de censura.



Fonte: Do autor (2022).

A correlação entre as médias *a posteriori* e os valores efetivamente simulados, antes do processo de censura, foi de $r = 39,07\%$ (Figura 4.8) com IC: $(21,70;\ 54,05)$, que é significativamente positiva pelo teste "t de Student" com probabilidade de significância $p = 3,2e-05$. Isto é um outro indício da adequação do método.

Figura 4.8 – Gráfico de dispersão das médias *a posteriori* dos 107 valores de produção estimados em relação aos seus respectivos valores simulados antes da censura ($r = 39,07\%$).



Fonte: Do autor (2022).

Com as cadeias de Markov geradas para os parâmetros das componentes da variância foi possível gerar estimativas *a posteriori* destas componentes e de formas da herdabilidade para a seleção entre parcelas $\left(h^2 = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}\right)$ e entre médias de tratamentos $\left(h_m^2 = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2/r}\right)$. Os resumos a *posteriori* destes parâmetros genéticos são apresentados na Tabela 4.11.

Quando comparado com os valores paramétricos da simulação original, observa-se um efeito de encolhimento das estimativa da variância genética e da variância do erro experimental, possivelmente devido ao efeito de escala. Para as herdabilidades, no entanto, em ambos os casos observa-se superestimação. Isto pode ser devido à escala e, ou, à consideração de informação adicional de que os "zeros" são valores ruins, associados ao efeito médio. Este tipo de resultado é mais difícil de interpretar e é interessante observar que consequências acarreta nas predições dos efeitos de tratamento.

Tabela 4.11 – Resumos das estimativas (distribuição *a posteriori*) das componentes de variância e herdabilidades em comparação aos seus respectivos valores paramétricos.

Parâmetro	**Valor Paramétrico**	**Estimativa**	**SD**	**LI**	**LS**
σ_u^2	1,00	0,30	0,05	0,20	0,40
σ_e^2	3,00	0,32	0,04	0,26	0,39
h^2	0,25	0,48	0,05	0,38	0,58
h_m^2	0,50	0,65	0,05	0,56	0,74

Fonte: Do autor (2022).

Os resumos das distribuições marginais *a posteriori* para os 7 piores e 7 melhores efeitos de tratamentos respectivamente, encontram-se na Tabela 4.12 a seguir. Nota-se que, para os efeitos abaixo da média, os sinais estão adequados, mas nenhum dos valores (1 a 7) foi relacionado entre os piores. Por outro lado, entre os melhores, apenas o 78 é bem distante do grupo. A maioria dos valores foram preditos corretamente, mas houve certa subestimação nos efeitos (encolhimento dos seus módulos). Podemos considerar, no entanto,que o processo de predição foi satisfatório pois os efeitos foram preditos na direção "correta". No entanto, avaliou-se também as correlações para concluir sobre o processo de seleção.

Tabela 4.12 – Resumos da distribuição *a posteriori* obtida pela amostragem Gibbs dos efeitos de tratamentos em comparação aos valores paramétricos gerados na simulação.

Tratamento	Valor Paramétrico	Média	SD	LI	LS
24	-0,86	-0,58	0,34	-1,23	0,10
12	-1,31	-0,58	0,39	-1,33	0,17
60	-0,02	-0,56	0,33	-1,20	0,10
22	-0,92	-0,48	0,42	-1,30	0,34
37	-0,52	-0,48	0,37	-1,22	0,24
19	-1,02	-0,47	0,42	-1,31	0,32
55	-0,12	-0,47	0,42	-1,30	0,35
⋮	⋮	⋮	⋮	⋮	⋮
78	0,36	0,60	0,32	-0,03	1,22
113	**1,47**	**0,62**	**0,30**	**0,04**	**1,21**
114	**1,54**	**0,68**	**0,30**	**0,08**	**1,26**
119	**2,04**	**0,69**	**0,30**	**0,11**	**1,27**
116	**1,69**	**0,81**	**0,30**	**0,25**	**1,41**
107	**1,18**	**0,89**	**0,30**	**0,32**	**1,48**
121	**2,64**	**1,12**	**0,30**	**0,52**	**1,70**

Fonte: Do autor (2022).

Para efeito de seleção dos genótipos de elite, a análise "PC" mostrou-se superior aos demais métodos de análise (até mesmo que o método "DC"), pois 6 tratamentos entre os 7 melhores apresentaram resultados acima de zero. Estes tratamentos encontram-se em negrito na Tabela 4.12.

As médias marginais *a posteriori* dos efeitos de tratamentos apresentaram correlação de $\hat{\rho} = 66,79\%$ com seus respectivos efeitos paramétricos, sendo significativamente positiva pelo teste "t de Student", o que é visualizado pela Tabela 4.13 e pela Figura 4.9. A porcentagem de recuperação da informação genética obtida com a previsão condicional dos dados foi de:

$$\frac{\hat{\rho}}{\sqrt{h_m^2}.100} = \frac{66,79\%}{70,71\%} = 94,46\%,$$

resultado este bem próximo do obtido na análise considerando que todos dados foram observados, mas ainda inferior ao da análise substituindo os censurados por zeros.

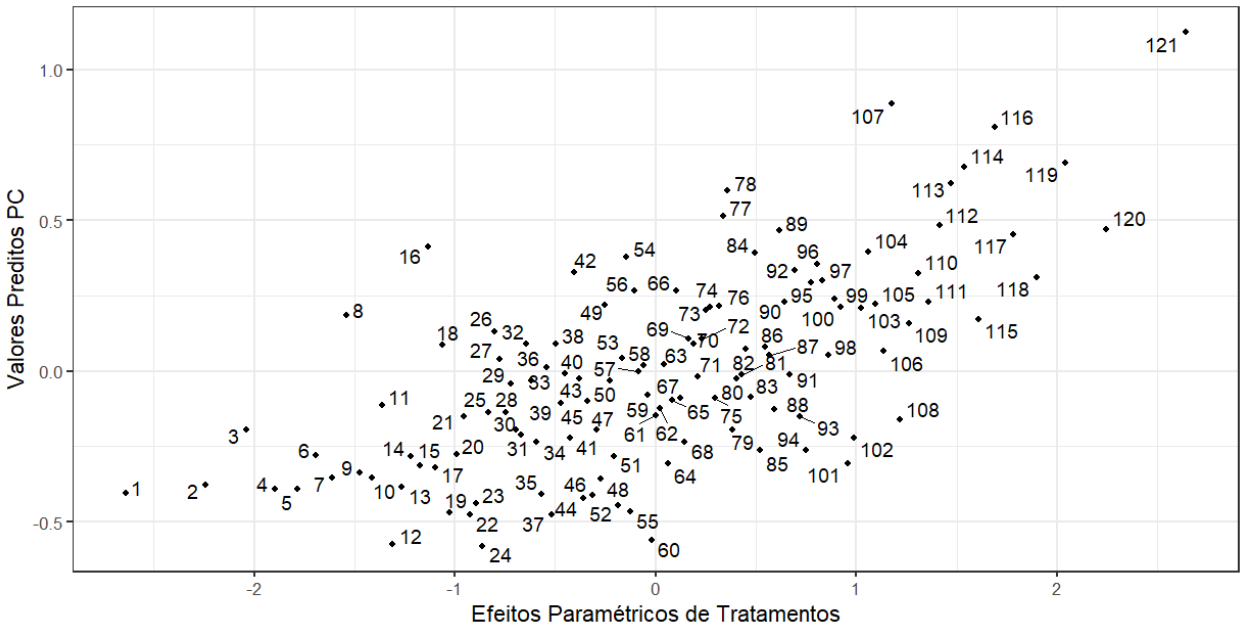
Verifica-se na Tabela 4.13, que os valores obtidos das correlações de Pearson e Sperman na análise "PC" (Predição Condicional) apresentaram-se próximos dos resultados obtidos nas análises "DC" e "C0", e superior aos resultados obtidos na análise "CE".

Tabela 4.13 – Correlações de Pearson e Sperman dos efeitos de tratamentos com seus respectivos efeitos paramétricos para o caso "PC" (Cenário 1). Intervalo de confiança (95%) usando a aproximação *t* de Student para correlações Pearson e Sperman.

Método	**Correlação (%)**	**valor-*p***	**IC**	
			LI	**LS**
Pearson	66,79	2,2e-16	55,56	75,62
Spearman	63,85	2,2e-16	55,88	71,81

Fonte: Do autor (2022).

Figura 4.9 – Gráfico de dispersão das predições dos efeitos de tratamentos "PC" (Cenário 1) e seus respectivos valores paramétricos ($\hat{\rho} = 66,79\%$).

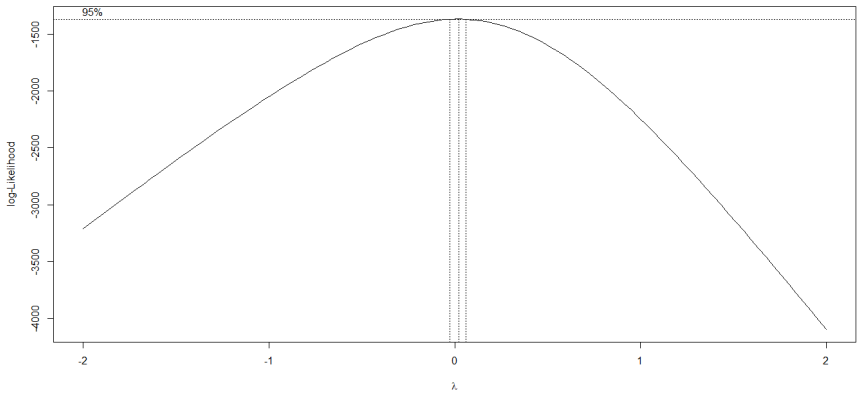


Fonte: Do autor (2022).

4.5 Cenário 2: substituindo censuras por 0 ("C0")

Para a variável produção **y** com a censura de dados substituída por zeros, foi utilizada a transformação logarítmica para melhorar a aproximação dos dados experimentais à distribuição normal. Como esta metodologia indicava a proximidade do $\lambda = 0$, conforme pode ser verificado na Figura 4.10 foi escolhida a transformação logarítmica com $\mathbf{y} + \mathbf{0,001}$, constante escolhida para eliminar os zeros do conjunto de dados e permitir o cálculo do logaritmo natural.

Figura 4.10 – Transformação de Box e Cox aplicada a $\mathbf{y} + \mathbf{0,001}$, com $\lambda = 0$ (Cenário 2: C0).



Fonte: Do autor (2022).

As estimativas dos parâmetros genéticos para está análise encontram-se na Tabela 4.14. Estimativas superestimadas das componentes da variância (superestimação superior a do cenário 1) são observadas.

Tabela 4.14 – Estimativas das componentes de variância e herdabilidades do caso "C0" (Cenário 2) , com seus respectivos IC (95%) e valores paramétricos usando os dados transformados.

Parâmetro	**Valor Paramétrico**	**Estimativa**	**IC**	
			LI	**LS**
σ_u^2	1,00	3,56	2,81	4,66
σ_e^2	3,00	17,97	15,41	21,23
h^2	0,25	0,17	0,13	0,21
h_m^2	0,50	0,37	0,30	0,44

Fonte: Do autor (2022).

Verifica-se também, que na análise "C0" o aumento da censura resultou na subestimação das estimativas das herdabilidades.

As predições dos efeitos de tratamentos para uma censura de aproximadamente 50% dos dados, apresentou correlação significativamente não nula $\hat{\rho} = 58,66\%$ com seus respectivos valores paramétricos, o que pode ser visualizado na Tabela 4.15 a seguir. A porcentagem de recuperação da informação genética para esta análise no cenário 2 foi de:

$$\frac{\hat{\rho}}{\sqrt{h_m^2}.100} = \frac{58,66\%}{70,71\%} = 82,96\%.$$

Tabela 4.15 – Correlações de Pearson e Sperman dos efeitos de tratamentos com seus respectivos efeitos paramétricos para o caso "C0" (Cenário 2). Intervalo de confiança (95%) usando a aproximação *t* de Student para correlações Pearson e Sperman.

Método	**Correlação (%)**	**valor-*p***	**IC**	
			LI	**LS**
Pearson	58,66	1,5e-12	45,59	69,26
Spearman	60,52	2,2e-16	52,28	68,76

Fonte: Do autor (2022).

É apresentado na Tabela 4.16 em ordem crescente do BLUP, os resumos das estimativas (BLUP e respectivos erros-padrão) dos 7 piores e 7 melhores efeitos de tratamentos. Espera-se valores baixos para o primeiro grupo (idealmente, de 1 a 7) e valores altos para o segundo grupo (idealmente, de 115 a 121). Apesar de uma boa correlação (Pearson) $\hat{\rho} = 58,66\%$, quando comparada com a Tabela 4.3 da análise "DC", uma alteração considerável na ordem dos melhores e dos piores tratamentos é observada. Alterações essas, desvantajosas no processo de seleção.

Tabela 4.16 – BLUP para os efeitos de tratamentos com a perda dos dados censurados, e seus respectivos erros-padrões, IC (95%) e valores paramétricos. (continua)

Tratamento	Valor Paramétrico	BLUP	SD_{BLUP}	LI_{BLUP}	LS_{BLUP}
22	-0,92	-2,34	1,49	-5,27	0,59
19	-1,02	-2,20	1,49	-5,13	0,72

Tabela 4.16 – BLUP para os efeitos de tratamentos com a perda dos dados censurados, e seus respectivos erros-padrões, IC (95%) e valores paramétricos. (conclusão)

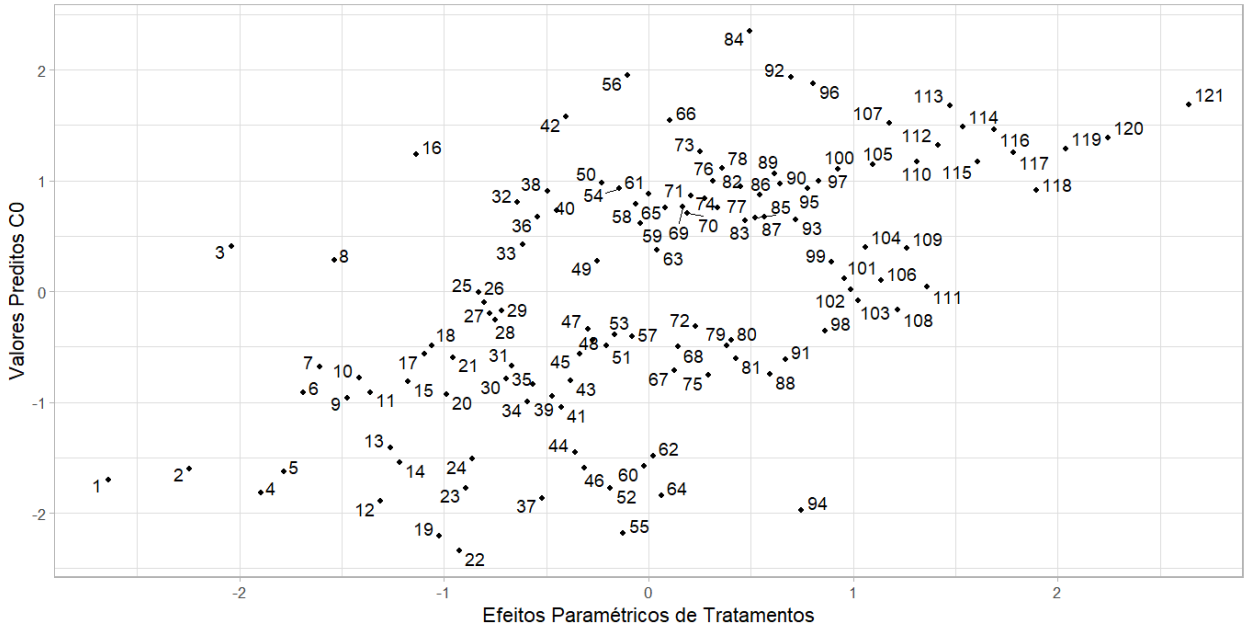
Tratamento	Valor Paramétrico	BLUP	SD_{BLUP}	LI_{BLUP}	LS_{BLUP}
55	-0,12	-2,18	1,49	-5,11	0,75
94	0,75	-1,97	1,49	-4,90	0,95
12	-1,31	-1,89	1,49	-4,82	1,03
37	-0,52	-1,86	1,49	-4,79	1,07
64	0,06	-1,84	1,49	-4,77	1,09
⋮	⋮	⋮	⋮	⋮	⋮
42	-0,40	1,58	1,49	-1,35	4,51
113	1,47	1,68	1,49	-1,25	4,61
121	2,64	1,69	1,49	-1,24	4,62
96	0,80	1,88	1,49	-1,05	4,80
92	0,69	1,94	1,49	-0,99	4,86
56	-0,10	1,95	1,49	-0,98	4,88
84	0,50	2,35	1,49	-0,57	5,28

Fonte: Do autor (2022).

Os resultados esperados para os postos de tratamentos seria de 1 a 7, e o mais próximo disso é o 12, por outro lado entre os melhores, os tratamentos 42, 56, 84 estão bem distantes do grupo (115 a 121). As alterações para os demais tratamentos é visualizada na Figura 4.13. Para os efeitos acima da média, os sinais não estão adequados para nenhum dos 7 melhores tratamentos.

Quando comparado com os dados completos, diferente do cenário 1, o aumento da censura substituindo os valores censurados por zeros acarretou na subestimação da correlação dos BLUP dos efeitos de tratamentos com seus valores paramétricos, o que pode ser visualizado na Figura 4.11.

Figura 4.11 – Gráfico de dispersão das predições dos efeitos de tratamentos "C0" e seus respectivos valores paramétricos ($\hat{\rho} = 58,66\%$).



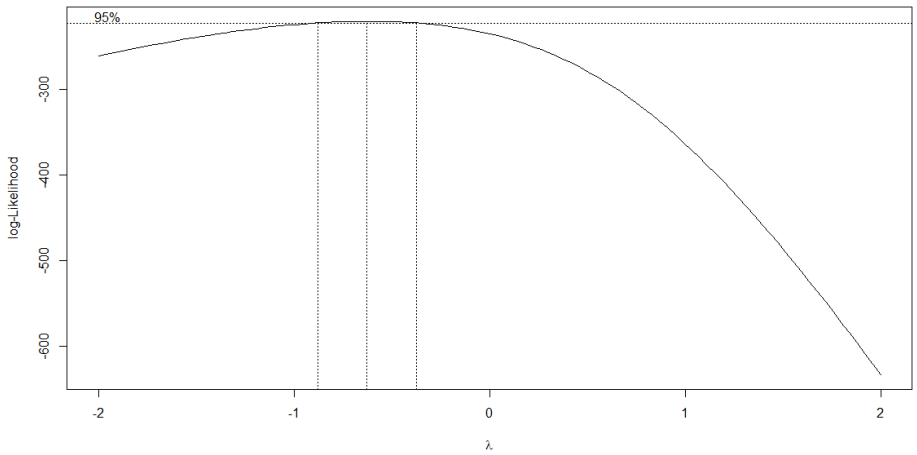
Fonte: Do autor (2022).

4.6 Cenário 2: eliminando censuras à esquerda ("CE")

Com a remoção da censura de aproximadamente 50% dos dados, a estimativa da componente de variância genética obtida para um primeiro banco de dados simulado foi igual a zero. Para seguimento da análise, foi necessário para este caso realizar um novo sorteio para obtenção de um novo conjunto de dados simulado. Em um pequeno estudo de simulação, gerando 100 vetores para variável produção **y**, com os mesmos parâmetros da análise apresentada, em 33% dos casos no método "CE" não se obteve convergência para a componente da variância genética. Certamente este é um problema comum neste tipo de análise quando se perde muitas unidades experimentais e isto pode levar a problemas na especificação da estrutura de covariâncias e nas estimativas resultantes.

Para os dados censurados removidos (perdidos) em que ocorreu a convergência das componentes de variância, precisou-se de uma transformação não linear Box e Cox (1964), seguida de uma correção na escala (Figura 4.12). Para cada observação da variável produção foi obtida sua transformação $Pt = \frac{(\mathbf{y}+\mathbf{0,1})^{\lambda}-1}{\lambda}$.

Figura 4.12 – Transformação de Box e Cox aplicada a $\mathbf{y}+\mathbf{0},\mathbf{1}$, com $\lambda \neq 0$ (Cenário 2: CE).



Fonte: Do autor (2022).

Com as componentes da variância obtidas é possível obter estimativas da herdabilidade para a seleção entre parcelas $\left(h^2 = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}\right)$ e entre médias de tratamentos $\left(h_m^2 = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2/r}\right)$. Os resumos destes parâmetros genéticos para análise "CE" são apresentados na Tabela 4.18. Nota-se uma ligeira subestimação das componentes da variância e das herdabilidades.

Tabela 4.18 – Estimativas das componentes de variância e herdabilidades do caso "CE" (Cenário 2) , com seus respectivos IC (95%) e valores paramétricos usando os dados transformados.

Parâmetro	**Valor Paramétrico**	**Estimativa**	**IC**	
			LI	**LS**
σ_u^2	1,00	0,74	0,59	0,97
σ_e^2	3,00	2,55	2,04	3,28
h^2	0,25	0,23	0,17	0,29
h_m^2	0,50	0,31	0,24	0,38

Fonte: Do autor (2022).

É interessante notar que, com o aumento da porcentagem de dados censurados, 18 dos tratamentos $(1,2,4,5,12,13,14,19,22,23,24,37,46,52,55,60,62,64$ e $94)$ foram desconsiderados da análise por não aparecerem com dados em qualquer unidade experimental. Isto representa $14,8\%$

dos tratamentos que ficariam sem estimativas. Embora estes sejam tratamentos ruins, devemos verificar se a retirada beneficia de alguma forma a análise.

Com a censura alta, quando comparada com as análises "DC" e "C0", a simples remoção dos dados censurados apresentou correlação significativamente não nula $\hat{\rho} = 44,68\%$ (Tabela 4.19) dos BLUP dos efeitos de tratamentos com seus respectivos valores paramétricos. O gráfico que ilustra esta dependência pode ser visualizado pela Figura 4.13. Consequentemente a recuperação da informação genética para esta análise foi de:

$$\frac{\hat{\rho}}{\sqrt{h_m^2}.100} = \frac{44,68\%}{70,71\%} = 63.19\%.$$

resultado significativamente inferior que as demais análises.

Tabela 4.19 – Correlações de Pearson e Sperman dos efeitos de tratamentos com seus respectivos efeitos paramétricos para o caso "CE" (Cenário 2). Intervalo de confiança (95%) usando a aproximação *t* de Student para correlações Pearson e Sperman.

Método	**Correlação (%)**	**valor-*p***	**IC**	
			LI	**LS**
Pearson	44,68	2,5e-06	27,63	59,00
Spearman	40,69	2,6e-05	27,29	54,08

Fonte: Do autor (2022).

Os resumos das estimativas no cenário 2 (BLUP e respectivos erros-padrão) dos 7 piores (esperava-se do 19 ao 25) e 7 melhores (esperava-se do 115 ao 121) efeitos de tratamentos da análise "CE" são exibidos na Tabela 4.20.

Tabela 4.20 – BLUP para os efeitos de tratamentos com a perda dos dados censurados, e seus respectivos erros-padrões, IC (95%) e valores paramétricos. (continua)

Tratamento	Valor Paramétrico	BLUP	SD_{BLUP}	LI_{BLUP}	LS_{BLUP}
85	0,52	-1,02	0,68	-2,37	0,32
61	0,00	-0,98	0,68	-2,33	0,36
101	0,96	-0,92	0,68	-2,26	0,42
93	0,72	-0,87	0,68	-2,21	0,47

Tabela 4.20 – BLUP para os efeitos de tratamentos com a perda dos dados censurados, e seus respectivos erros-padrões, IC (95%) e valores paramétricos. (conclusão)

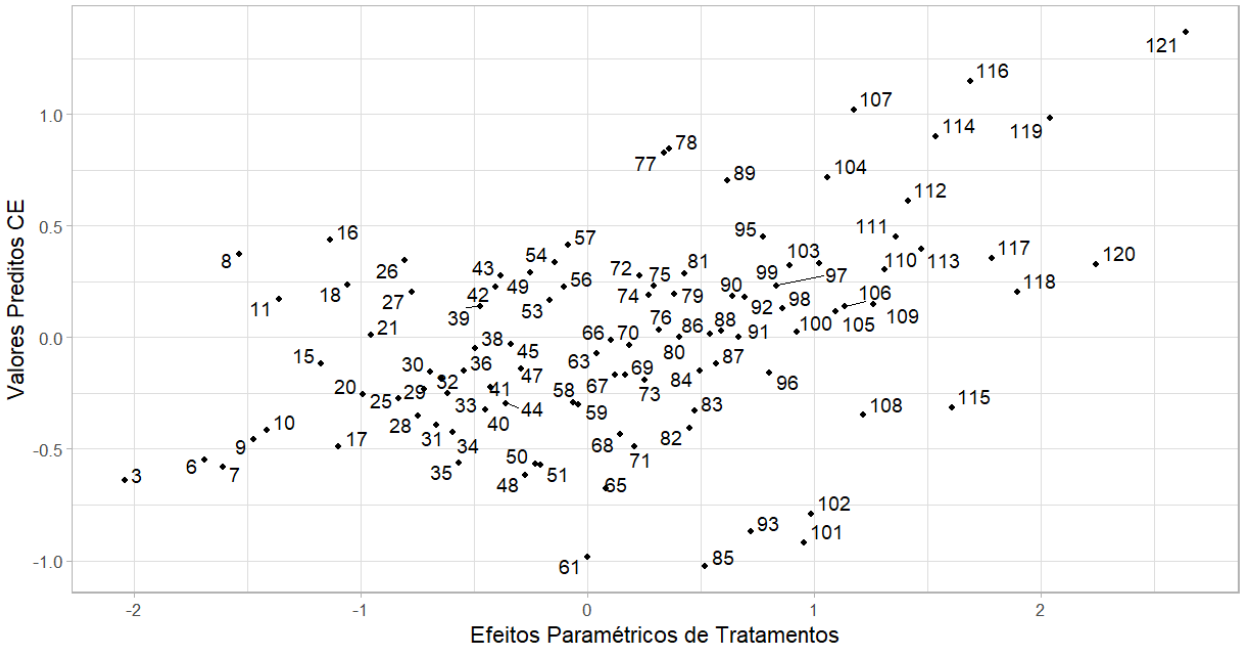
Tratamento	Valor Paramétrico	BLUP	SD_{BLUP}	LI_{BLUP}	LS_{BLUP}
102	0,99	-0,79	0,68	-2,13	0,55
65	0,08	-0,68	0,68	-2,02	0,66
3	-2,04	-0,64	0,68	-1,98	0,70
⋮	⋮	⋮	⋮	⋮	⋮
77	0,34	0,83	0,68	-0,52	2,17
78	0,36	0,84	0,68	-0,50	2,19
114	1,54	0,90	0,63	-0,33	2,14
119	2,04	0,98	0,63	-0,25	2,22
107	1,18	1,02	0,63	-0,22	2,25
116	1,69	1,15	0,63	-0,09	2,38
121	**2,64**	**1,37**	**0,63**	**0,13**	**2,60**

Fonte: Do autor (2022).

Nota-se que os maiores problemas estão entre os valores menores. Como referência é preciso lembrar que os valores esperados para os postos de tratamentos seria de 19 a 25, e o mais próximo disso é o 61. Para os efeitos acima da média, os sinais estão adequados somente para o tratamento 121, por outro lado, entre os melhores, apenas os tratamentos 77 e 78 estão bem distantes do grupo. As alterações para os demais tratamentos é visualizada na Figura 4.13.

Apesar de uma menor correlação na análise "CE" do que na análise "C0", tem-se pelo menos o tratamento 121 (1/7) presente no grupo dos genótipos de elite. Portanto, não considerar os dados censurados em um cenário com taxa de censura alta, mostrou-se superior para fins de seleção dos genótipos de elite do que a análise substituindo os dados censurados por zeros.

Figura 4.13 – Gráfico de dispersão das predições dos efeitos de tratamentos "CE" e seus respectivos valores paramétricos ($\hat{\rho} = 44,68\%$).



Fonte: Do autor (2022).

4.7 Cenário 2: previsão condicional dos dados censurados ("PC")

Na Tabela 4.22, apresenta-se os resumos da distribuição marginal *a posteriori* de alguns dos 179 valores de produção η obtidos através da amostragem de Gibbs.

Tabela 4.22 – Resumos da distribuição *a posteriori* obtida pela amostragem Gibbs em comparação aos valores paramétricos gerados na simulação. (continua)

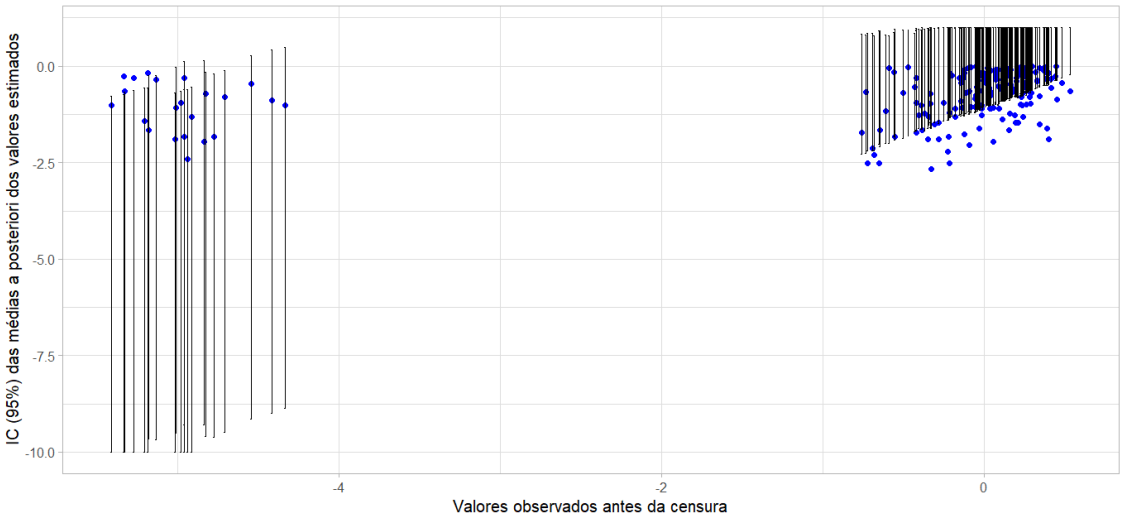
Censuras	Valor gerado	Média	SD	LI	LS
1	-1,90	-0,35	0,72	-1,61	0,99
2	-0,71	-0,33	0,71	-1,59	1,00
3	-1,66	-0,38	0,72	-1,64	1,00
4	-1,24	-0,37	0,71	-1,67	0,96
5	-1,71	-0,42	0,70	-1,69	0,91
6	-0,94	-0,42	0,70	-1,68	0,95

Tabela 4.22 – Resumos da distribuição *a posteriori* obtida pela amostragem Gibbs em comparação aos valores paramétricos gerados na simulação. (conclusão)

Censuras	Valor gerado	Média	SD	LI	LS
⋮	⋮	⋮	⋮	⋮	⋮
7	-1,02	-0,39	0,69	-1,63	0,95
173	-1,02	0,23	0,53	-0,78	1,00
174	-0,07	0,21	0,54	-0,81	1,00
175	-0,99	0,26	0,51	-0,70	1,00
176	-0,97	0,29	0,52	-0,70	1,00
177	-0,17	0,40	0,46	-0,49	1,00
178	-0,20	0,40	0,46	-0,51	1,00
179	-0,33	0,39	0,46	-0,50	1,00

Fonte: Do autor (2022).

Figura 4.14 – Intervalos de credibilidade (95%) dos 179 valores de produção estimados em relação aos valores conhecidos da variável produção gerados na simulação, antes do processo de censura.



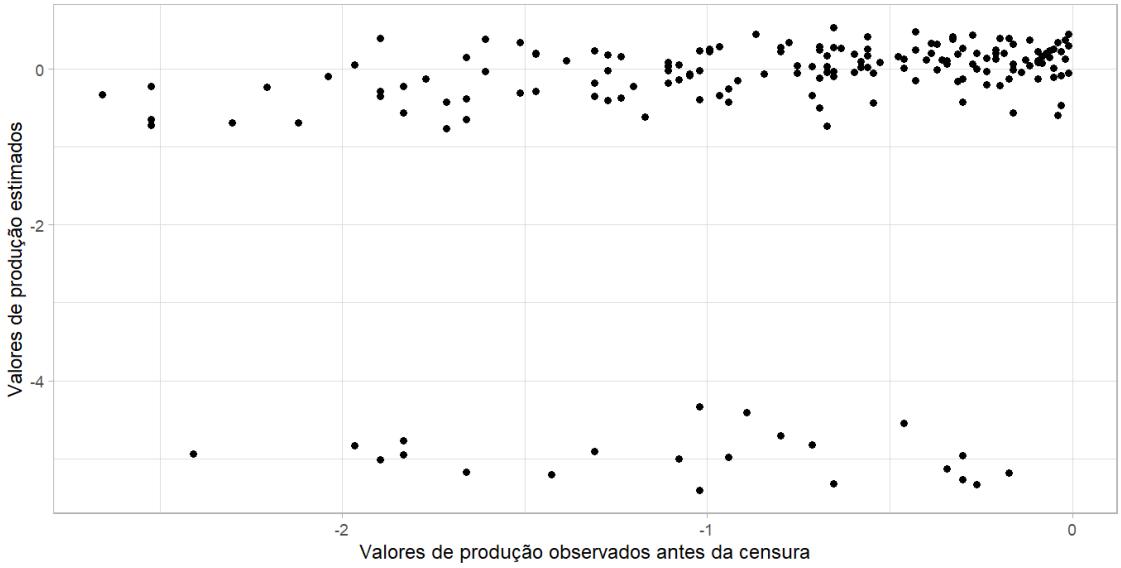
Fonte: Do autor (2022).

Como na censura moderada de dados, verifica-se na Figura 4.14 que grande parte dos intervalos de credibilidade (95%) da média marginal *aposteriori* dos valores estimados contém os

valores paramétricos conhecidos da variável produção na simulação, antes do processo de censura. Novamente, o processo de inferência escolhido mostrou-se interessante para a análise prática.

A correlação entre as médias *a posteriori* e os valores efetivamente simulados, antes do processo de censura, foi de $r = 20,99\%$ (Figura 4.15) e IC: $(0,0652;\ 0,3459)$, que é significativamente positiva pelo teste "t de Student" com probabilidade de significância $p = 0,0048$. Isto é um outro indício da adequação do método.

Figura 4.15 – Gráfico de dispersão das médias *a posteriori* dos 179 valores de produção estimados em relação aos seus respectivos valores simulados antes da censura ($r = 20,99\%$).



Fonte: Do autor (2022).

Com as cadeias de Markov geradas para os parâmetros das componentes da variância foi possível gerar estimativas a posteriori destas componentes e de formas da herdabilidade para a seleção entre parcelas $\left(h^2 = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}\right)$ e entre médias de tratamentos $\left(h_m^2 = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2/r}\right)$. Os resumos *a posteriori* destes parâmetros genéticos são apresentados na Tabela 4.24.

Devido as escalas utilizadas, observa-se no cenário 2 um encolhimento ligeiramente maior das estimativas da variância genética e variância do erro experimental. Para as herdabilidades, no entanto, em ambos os casos observa-se uma superestimação ligeiramente maior que a do cenário 1. Isto pode ser devido à escala e, ou, à consideração de informação adicional de que os "zeros" são valores ruins, associados ao efeito médio. Este tipo de resultado é mais difícil de interpretar e é interessante observar que consequências acarreta nas predições dos efeitos de tratamento.

Tabela 4.24 – Resumos das estimativas (distribuição *a posteriori*) das componentes de variância e herdabilidades em comparação aos seus respectivos valores paramétricos.

Parâmetro	**Valor Paramétrico**	**Estimativa**	**SD**	**LI**	**LS**
σ_u^2	1,00	0,29	0,05	0,19	0,29
σ_e^2	3,00	0,29	0,04	0,22	0,37
h^2	0,25	0,50	0,05	0,40	0,60
h_m^2	0,50	0,66	0,05	0,58	0,75

Fonte: Do autor (2022).

Os resumos das distribuições marginais *a posteriori* para os 7 piores e 7 melhores efeitos de tratamentos encontram-se na Tabela 4.25 a seguir. Nota-se que, para os efeitos abaixo da média, os sinais estão adequados, sendo dois (6 e 7) dos valores de 1 a 7 relacionados entre os piores. Por outro lado, entre os melhores, apenas o 78 é bem distante do grupo. A maioria dos valores foi estimada corretamente, mas houve certa subestimação nos efeitos (encolhimento dos seus módulos). Podemos considerar, no entanto, que o processo de estimação foi satisfatório pois os efeitos foram estimados na direção "correta". No entanto, avaliou-se também as correlações para concluir sobre o processo de seleção.

Tabela 4.25 – Resumos da distribuição *a posteriori* obtida pela amostragem Gibbs dos efeitos de tratamentos em comparação aos valores paramétricos gerados na simulação. (continua)

Tratamento	Valor Paramétrico	Média	SD	LI	LS
35	-0.57	-0.40	0.37	-1.13	0.31
60	-0.02	-0.39	0.42	-1.23	0.44
48	-0.27	-0.38	0.38	-1.10	0.37
24	-0.86	-0.38	0.43	-1.28	0.43
12	-1.31	-0.36	0.43	-1.23	0.47
6	-1.69	-0.36	0.37	-1.06	0.37
7	-1.61	-0.36	0.38	-1.10	0.38
⋮	⋮	⋮	⋮	⋮	⋮

Tabela 4.25 – Resumos da distribuição *a posteriori* obtida pela amostragem Gibbs dos efeitos de tratamentos em comparação aos valores paramétricos gerados na simulação. (conclusão)

Tratamento	Valor Paramétrico	Média	SD	LI	LS
113	1,47	0,53	0,29	-0,02	1,10
78	0,36	0,58	0,31	-0,03	1,19
114	**1,54**	**0,62**	**0,29**	**0,06**	**1,18**
119	**2,04**	**0,64**	**0,29**	**0,08**	**1,21**
116	**1,69**	**0,74**	**0,29**	**0,18**	**1,30**
107	**1,18**	**0,84**	**0,29**	**0,27**	**1,42**
121	**2,64**	**1,07**	**0,29**	**0,51**	**1,66**

Fonte: Do autor (2022).

Para o cenário 2, para efeito de seleção dos genótipos de elite, a análise "PC" mostrou-se novamente superior que os métodos de análise "DC", "C0" e "CE", pois 5 dos 7 tratamentos que se encontram próximos ou entre os melhores, pertencem ao grupo dos genótipos de elite. Observa-se esses tratamentos em negrito na Tabela 4.25.

As médias marginais a *posteriori* dos efeitos de tratamentos apresentaram correlação $\hat{\rho} = 62,72\%$ com seus respectivos efeitos paramétricos, sendo significativamente não nulo pelo pelo teste "t de Studant" com probabilidade de significância vista na Tabela 4.27 e na Figura 4.16. A porcentagem de recuperação da informação genética foi de:

$$\frac{\hat{\rho}}{\sqrt{h_m^2}.100} = \frac{62,72\%}{70,71\%} = 88,71\%.$$

Resultado este próximo do obtido na análise considerando que todos dados foram observados, e superior às demais análises.

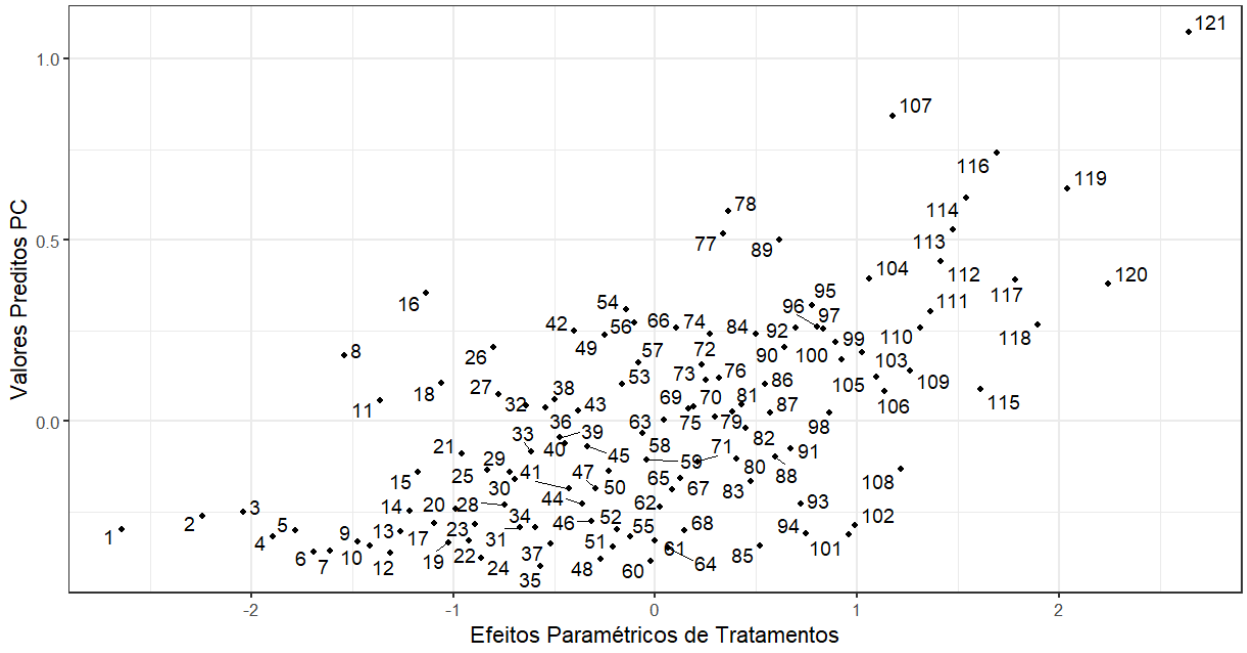
Verifica-se na Tabela 4.27, que os valores obtidos das correlações de Pearson e Sperman na análise "PC" (Previsão Condicional) apresentaram-se próximos dos resultados obtidos nas análises "DC", e superiores que os resultados obtidos nas análise "CE".

Tabela 4.27 – Correlações de Pearson e Sperman dos efeitos de tratamentos com seus respectivos efeitos paramétricos para o caso "PC" (Cenário 2). Intervalo de confiança (95%) usando a aproximação *t* de Student para correlações Pearson e Sperman.

Método	**Correlação (%)**	**valor-*p***	**IC**	
			LI	**LS**
Pearson	62,72	1,4e-14	50,53	72,46
Spearman	58,44	2,2e-16	50,04	66,84

Fonte: Do autor (2022).

Figura 4.16 – Gráfico de dispersão das predições dos efeitos de tratamentos ("PC") e seus respectivos valores paramétricos (correlação estimada: $\hat{\rho} = 62,72\%$).



Fonte: Do autor (2022).

4.8 Síntese do Cenário 1

Na Tabela 4.28 a seguir, encontram-se as estimativas de herdabilidade para a seleção entre parcelas h^2 e para a seleção entre médias de h^2_m, obtidas tanto pela amostragem de Gibbs quanto pela REML e seus respectivos valores paramétricos.

A análise "PC" apresentou estimativas superiores para as herdabilidades que as estimativas obtidas pelas demais análises. Se tem uma herdabilidade acima de seu valor paramétrico, que é dado por $h^2 = 0,25$ para herdabilidade entre parcelas e $h^2_m = 0,50$ para herdabilidade entre as

médias de tratamentos. Já a análise "CE" apresentou estimativas para as herdabilidades inferiores que seus valores paramétricos.

Tabela 4.28 – Quadro-resumo das estimativas das herdabilidades, com seus respectivos IC com (95%) de probabilidade e seus valores paramétricos.

Herdabilidade	**Tipo de análise**	**Estimativa**	**IC**		**Valor Paramétrico**
			LI	**LS**	
$h^2 = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$	**DC**	0,26	0,21	0,32	**0,25**
	C0	0,24	0,19	0,30	
	CE	0,15	0,11	0,19	
	PC	0,48	0,38	0,58	
$h_m^2 = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_e^2}{r}}$	**DC**	0,52	0,44	0,59	**0,50**
	C0	0,49	0,42	0,56	
	CE	0,27	0,21	0,33	
	PC	0,65	0,56	0,74	

Fonte: Do autor (2022).

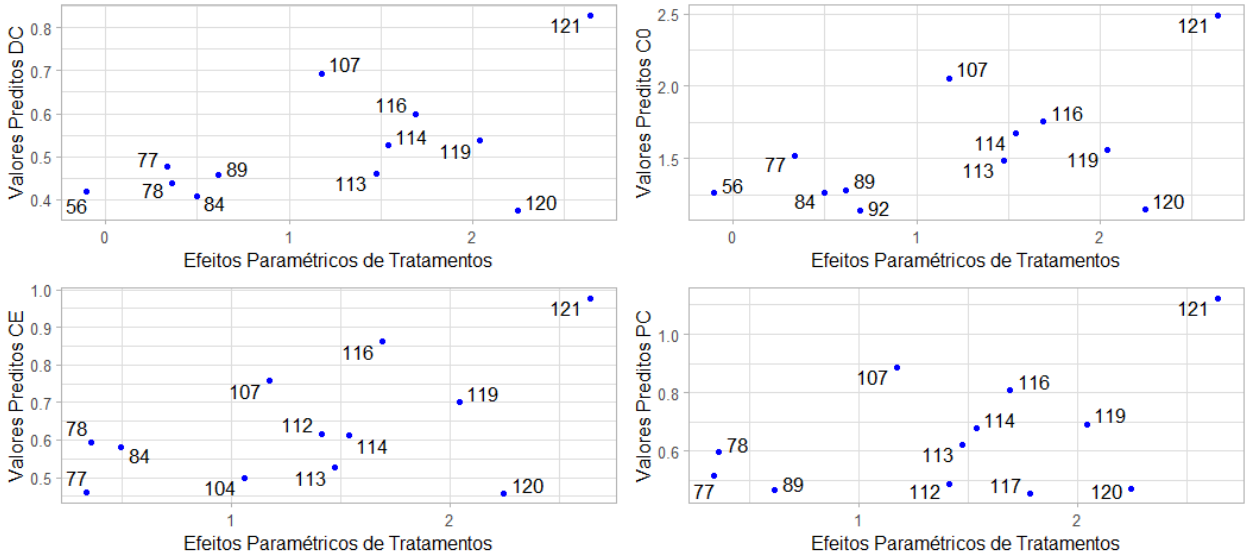
É apresentada na Tabela 4.29, a reunião das porcentagens de recuperação da informação genética e das correlações "Pearson" (entre as predições dos efeitos de tratamentos com seus respectivos valores simulados) obtidas pelos diferentes métodos de análise. Já na Figura 4.17, é apresentada a reunião dos gráficos de dispersão (e coeficientes de correlação) entre as 12 maiores predições de efeitos de tratamentos e seus respectivos valores paramétricos para cada tipo de análise.

Tabela 4.29 – Quadro-resumo das porcentagens de recuperação da informação genética e correlações (Pearson) dos efeitos paramétricos de tratamentos, com suas respectivas predições pelos diferentes métodos de análise. Intervalo de confiança (95%) usando a aproximação *t* de Student.

Tipo de análise	**Informação Genética (%)**	**Correlação (%)**	**IC**	
			LI	**LS**
DC	97,99	69,29	58,79	77,55
C0	97,66	69,06	58,41	77,37
CE	74,64	52,78	38,01	64,92
PC	94,46	66,79	55,56	75,62

Fonte: Do autor (2022).

Figura 4.17 – Gráfico de dispersão e coeficientes de correlação entre as 12 maiores predições de efeitos de tratamentos e respectivos valores paramétricos entre eles. Análises: DC ($\hat{\rho} = 0,54$), C0 ($\hat{\rho} = 0,54$), CE ($\hat{\rho} = 0,51$) e PC ($\hat{\rho} = 0,44$).



Fonte: Do autor (2022).

Quando comparada com a análise ideal (sem a censura de dados "DC"), a censura dos dados "CE" apresentou pior cenário entre as formas de análise, tanto para estimação dos parâmetros genéticos, tal como para efeito de seleção. Para um grau moderado de perda de dados, substituir os valores censurado por zeros mostrou-se uma alternativa interessante.

Apesar do método "DC" apresentar correlação (para efeitos de tratamentos com seus valores paramétricos) maior do que o método "PC", para seleção dos genótipos de elite, a análise "PC" apresentou maior acurácia, sendo que em 6 dos 7 melhores tratamentos seus IC não continham o valor zero, ou seja, apresentaram estimativas que diferem significativamente de zero.

4.9 Síntese do Cenário 2

Na Tabela 4.30 a seguir, encontram-se as estimativas de herdabilidade para a seleção entre parcelas h^2 e para a seleção entre médias de h^2_m, obtidas tanto pela amostragem de Gibbs quanto pela REML e seus respectivos valores paramétricos.

Tabela 4.30 – Quadro-resumo das estimativas da herdabilidade e herdabilidade experimental, com seus respectivos IC com (95%) de probabilidade e seus valores paramétricos.

Herdabilidade	**Tipo de análise**	**Estimativa**	**IC**		**Valor Paramétrico**
			LI	**LS**	
$h^2 = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$	**DC**	0,26	0,21	0,32	**0,25**
	C0	0,17	0,13	0,21	
	CE	0,23	0,17	0,29	
	PC	0,50	0,40	0,60	
$h_m^2 = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_e^2}{r}}$	**DC**	0,52	0,44	0,59	**0,50**
	C0	0,37	0,30	0,44	
	CE	0,31	0,24	0,38	
	PC	0,66	0,58	0,75	

Fonte: Do autor (2022).

No cenário 2, assim como na censura moderada de dados, a análise "PC" apresentou estimativas maiores para as herdabilidades do que as estimativas obtidas pelas análises "DC", "C0" e "CE". Em relação as herdabilidade paramétrica, uma subestimação ocorreu nas estimativas das herdabilidades nas análises "C0" e "CE".

Com a taxa de censura alta de dados, quando comparada com o cenário ideal de análise (sem censura), a análise "PC" apresentou correlações dos efeitos de tratamentos com os efeitos paramétricos maiores e mais próximos da análise sem censura do que os demais métodos de análise. Desse modo, para uma censura maior de dados, o método de previsão condicional "PC" mostrou-se mais acurada para fins de seleção. Apesar do método "C0" apresentar correlação (para efeitos de tratamentos com seus valores paramétricos) maior do que o método "CE", para fins de seleção dos genótipos de elite, a análise "CE" mostrou-se melhor que a análise "C0".

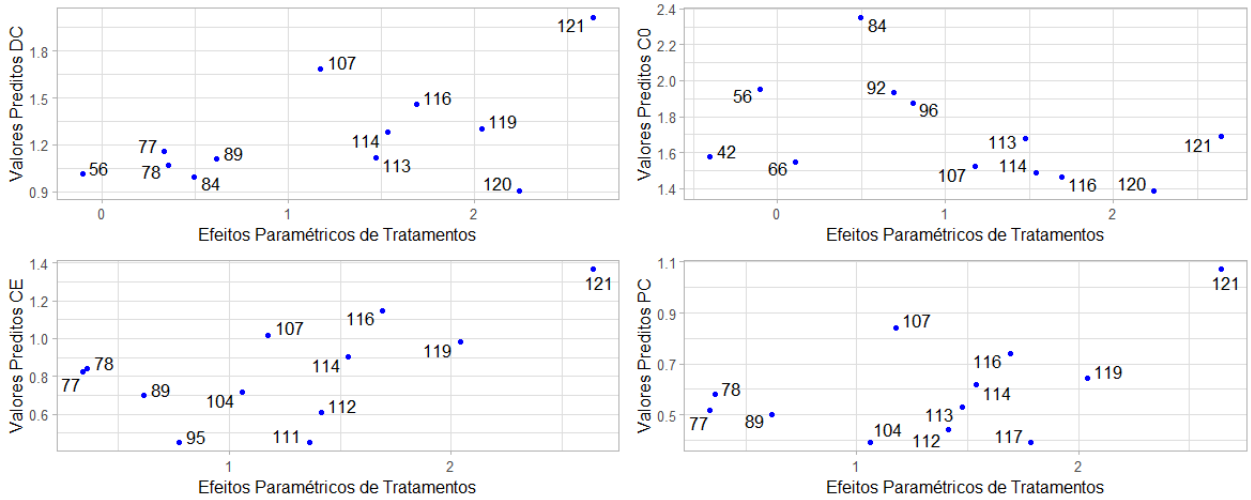
É apresentado na Tabela 4.31, a reunião das porcentagens de recuperação da informação genética e das correlações "Pearson" (entre as predições dos efeitos de tratamentos com seus respectivos valores simulados) obtidas pelos diferentes métodos de análise. Já na Figura 4.18, é apresentado a reunião dos gráficos de dispersão (e coeficientes de correlação) entre as 12 maiores predições de efeitos de tratamentos e seus respectivos valores paramétricos para cada tipo de análise.

Tabela 4.31 – Quadro-resumo das porcentagens de recuperação da informação genética e correlações (Pearson) dos efeitos paramétricos de tratamentos, com suas respectivas predições pelos diferentes métodos de análise. Intervalo de confiança (95%) usando a aproximação *t* de Student.

Tipo de análise	**Informação Genética (%)**	**Correlação (%)**	**IC**	
			LI	**LS**
DC	97,99	69,29	58,79	77,55
C0	82,96	58,66	45,59	69,26
CE	63,19	44,68	27,63	59,00
PC	88,71	62,72	50,53	72,46

Fonte: Do autor (2022).

Figura 4.18 – Gráfico de dispersão e coeficientes de correlação entre as 12 maiores predições de efeitos de tratamentos e respectivos valores paramétricos entre eles. Análises: DC ($\hat{\rho} = 0,54$), C0 ($\hat{\rho} = -0,40$), CE ($\hat{\rho} = 0,59$) e PC ($\hat{\rho} = 0,52$).



Fonte: Do autor (2022).

4.10 Discussão Geral

No cenário 2 com taxa de censura alta ($\sim 50\%$) a análise estimando os dados censurados mostrou-se promissora para correlação dos efeitos de tratamentos, apresentando maior correlação do que as demais análises, ficando próximo do valor ideal obtido quando se tem todos os dados observados.

Em comparação com o cenário 1, vale ressaltar que o aumento da proporção de dados censurados para análise "C0", resultou na diminuição das correlações dos efeitos de tratamentos

com os valores paramétricos. Por outro lado, para a taxa de cesura baixa ($\sim 30\%$), substituir os valores censurados por zeros mostrou-se um procedimento interessante.

O método de análise ignorando os dados abaixo da linha crítica "CE" foi adotado no experimento de Silva (2019). De acordo com os resultados obtidos no presente trabalho esse método de análise deve ser reconsiderado, já que em ambos os cenários as estimativas dos parâmetros genéticos de interesse e predições dos efeitos de seleção apresentaram-se inferiores aos resultados obtidos na análise proposta.

O resultado mais surpreendente foi a superestimação das herdabilidades no método "PC" em ambos os cenários. Como tentativa de explicação podemos levantar a hipótese de que a análise é de fato bivariada, ou seja, uma combinação entre o caractere efetivamente medido e a censura por razão prática, que na verdade é uma nota subjetiva aplicada pelo melhorista prático. Como o primeiro é uma variável contínua e o segundo uma nota binária extra, esta informação pode melhorar a qualidade de um dado ruim que tenha sido efetivamente observado, mas ignorado ou tratado como zero pelas metodologias convencionalmente empregadas. É claro que a herdabilidade bivariada é um resultado conhecido da genética quantitativa e isto poderia ser verificado na simulação, mas não nos ocorreu imaginar que a análise poderia ser melhor que a de dados completos e assim planejar este estudo em especial. Para maior confiança nesta afirmação maiores estudos com dados reais e simulações são recomendáveis.

De forma geral, o experimentador prefere estabelecer um limiar de detecção por restrições práticas como minorar a carga de trabalho e facilitar a decisão imediata sobre o descarte de material ruim. A hipótese após este estudo de simulação é que esta decisão pode ser justificada, embora possa levar a problemas na estimação das variâncias genéticas e ambiental. Consequentemente isto pode levar a predições ruins de ganho por seleção e progresso genético, que na verdade não são o objetivo de estudos iniciais, mas podem, por outro lado, trazer ganhos inesperados de seleção, este sim o objetivo principal.

5 CONCLUSÕES

No cenário com moderada taxa de censura (30%), "PC" e "C0" foram equivalentes em termos de correlações, mas para a seleção dos melhores genótipos a forma proposta foi superior, proporcionando melhor seleção para o grupo dos genótipos de elite. A forma usual "CE" no entanto, apresenta resultados bastante ruins tanto por subestimar a componente genética da variância como resultar em correlação mais baixa e seleção ruim. Neste cenário, no entanto, a análise proposta superestima ligeiramente as herdabilidades, embora esta interpretação não seja direta.

Para o cenário com alta taxa de censura (50%), "PC" foi a mais acurada em termos de correlações e para fins de seleção. Note-se que, neste caso, "CE" frequentemente nem converge, falhando em reconhecer variância genética. No exemplo convergente apresentou correlações menores que "C0" mas na seleção dos genótipos de elite foi melhor. Todas as formas de análise (incluindo "C0" e "PC") foram ruins para estimar herdabilidades.

Concluímos que, em estágios iniciais do programa de melhoramento e com ocorrência de altas taxas de censura, é mais recomendada a análise "PC" pois as decisões sobre seleção são melhores e isso é mais importante do que interpretação de estimativas de herdabilidade.

REFERÊNCIAS

AMIRI, M.; JENSEN, R. Missing data imputation using fuzzy-rough methods. **Neurocomputing**, Elsevier, v. 205, p. 152–164, 2016.

BATES, D. et al. Fitting Linear Mixed-Effects Models Using lme4. **Journal of Statistical Software**, v. 67, n. 1, p. 1–48, 2015. DOI: 10.18637/jss.v067.i01. Disponível em: https://www.jstatsoft.org/index.php/jss/article/view/v067i01.

BOLSTAD, W. M.; CURRAN, J. M. **Introduction to Bayesian statistics**. [S.l.]: John Wiley & Sons, 2016.

BOX, G. E.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 26, n. 2, p. 211–243, 1964.

BOX, G. E.; TIAO, G. C. **Bayesian inference in statistical analysis**. [S.l.]: John Wiley & Sons, 1973. v. 40.

CHRISTOFARO, C.; LEÃO, M. Tratamento de dados censurados em estudos ambientais. **Química Nova**, SciELO Brasil, v. 37, p. 104–110, 2014.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. [S.l.]: Editora Blucher, 2006.

DEMIDENKO, E. **Mixed models: theory and applications with R**. [S.l.]: John Wiley & Sons, 2013.

DEMPSTER, A.; RUBIN, D. **Overview. Incomplete Data in Sample Surveys, Vol. II: Theory and Annotated Bibliography**. [S.l.]: New York: Academic Press, 1983.

GELFAND, A. E.; SMITH, A. F. Sampling-based approaches to calculating marginal densities. **Journal of the American statistical association**, Taylor & Francis, v. 85, n. 410, p. 398–409, 1990.

GELMAN, A. Prior distribution. **Encyclopedia of environmetrics**, Citeseer, v. 3, n. 4, p. 1634–1637, 2002.

GELMAN, A.; CARLIN, J. B. et al. **Bayesian data analysis**. [S.l.]: Chapman e Hall/CRC, 1995.

GELMAN, A.; CARLIN, J. B. et al. **Bayesian data analysis**. [S.l.]: Chapman e Hall/CRC, 2013.

GEMAN, S.; GEMAN, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, n. 6, p. 721–741, 1984.

GUO, S. **Survival analysis**. [S.l.]: Oxford University Press, 2010.

HENDERSON, C. R. Applications of linear models in animal breeding. 462 pp. **University of Guelph, Guelph, Ontario**, 1984.

HENDERSON, C. R. Sire evaluation and genetic trends. **Journal of Animal Science**, Oxford University Press, v. 1973, Symposium, p. 10–41, 1973.

HENDERSON, C. R. et al. The estimation of environmental and genetic trends from records subject to culling. **Biometrics**, JSTOR, v. 15, n. 2, p. 192–218, 1959.

HINKELMANN, K.; KEMPTHORNE, O. **Design and analysis of experiments, volume 1: Introduction to experimental design**. [S.l.]: John Wiley & Sons, 2007. v. 1.

KINAS, P. G.; ANDRADE, H. A. **Introdução à Análise Bayesiana (com R) 2a Edição**. [S.l.]: Consultor Editorial, 2021.

KLEIN, J. P.; MOESCHBERGER, M. L. **Survival analysis: techniques for censored and truncated data**. [S.l.]: Springer, 2003. v. 2.

KLEINBAUM, D. G.; KLEIN, M. et al. **Survival analysis: a self-learning text**. [S.l.]: Springer, 2012. v. 3.

LEUNG, K.-M.; ELASHOFF, R. M.; AFIFI, A. A. Censoring issues in survival analysis. **Annual review of public health**, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 18, n. 1, p. 83–104, 1997.

LIMA, C. N. d. L.; BUENO FILHO, J. S. d. S. Ajuste de uma superfıécie de resposta no delineamento em blocos incompletos: análise de verossimilhança restrita e bayesiana em uma simulação de adubação em citros. **Rev. Mat. Estat**, n. 4, p. 133–157, 2006.

LITTLE, R. J.; RUBIN, D. B. **Statistical analysis with missing data**. [S.l.]: John Wiley & Sons, 2019. v. 793.

LIU, X. **Survival analysis: models and applications**. [S.l.]: John Wiley & Sons, 2012.

MCGRORY, E.; HOLIAN, E.; MORRISON, L. Assessment of groundwater processes using censored data analysis incorporating non-detect chemical, physical, and biological data. **Journal of Contaminant Hydrology**, Elsevier, v. 235, p. 103706, 2020.

ONOFRI, A.; PIEPHO, H.-P.; KOZAK, M. Analysing censored data in agricultural research: A review with examples and software tips. **Annals of applied biology**, Wiley Online Library, v. 174, n. 1, p. 3–13, 2019.

PEREIRA, G. M. d. C. Medidas alternativas para comparação de modelos e aplicação de métodos de aprendizado de máquina e de redução de dimensionalidade para seleção genômica com dados censurados. Universidade Federal de Viçosa, 2020.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2022. Disponível em: https://www.R-project.org/.

RAGHUNATHAN, T. **Missing data analysis in practice**. [S.l.]: CRC press, 2015.

REID, N.; COX, D. **Analysis of survival data**. [S.l.]: Chapman e Hall/CRC, 2018.

RESENDE, M. D. V. de. Análise estatística de modelos mistos via REML/BLUP na experimentação em melhoramento de plantas perenes. Colombo: Embrapa Florestas, 2000., 2000.

RESENDE, M. D. V. de et al. Estimação de componentes de variância e predição de valores genéticos pelo método da máxima verossimilhança restrita (Reml) e melhor predição linear não viciada (Blup) em Pinus. Boletim de Pesquisa Florestal, Colombo, n. 32/33, p. 23-42, jan./dez. 1996., 1996.

ROBINSON, G. K. That BLUP is a good thing: the estimation of random effects. **Statistical science**, JSTOR, 1991.

SATTERTHWAITE, F. E. An approximate distribution of estimates of variance components. **Biometrics bulletin**, JSTOR, v. 2, n. 6, p. 110–114, 1946.

SEARLE, S.; CASSELLA, G.; MCCULLOUCH, C. Variance Components In: New York: Jonh Wiley-Sons. **DOI: http://dx. doi. org/10.1002/9780470316856**, 1992.

SILVA. J. C. d. O. Seleção de clones de batata-doce para diferentes aptidões agronômicas. 2019. 86 p. Dissertação (Mestrado em Agronomia/Fitotecnia). Universidade Federal de Lavras,Lavras, 2019.

SMITH, P. J. **Analysis of failure and survival data**. [S.l.]: Chapman e Hall/CRC, 2017.

SÖRENSEN, D. Gibbs sampling in quantitative genetics. **Intern Rapport. Statens Husdyrbrugsforsoeg (Denmark). no. 82.**, 1996.

SÖRENSEN, D.; GIANOLA, D. Likelihood, Bayesian and MCMC Methods in Quantitative Genetics. Springer, p. 313, 2002.

SÖRENSEN, D. A.; GIANOLA, D.; KORSGAARD, I. R. Bayesian mixed-effects model analysis of a censored normal distribution with animal breeding applications. **Acta Agriculturae Scandinavica A—Animal Sciences**, Taylor & Francis, v. 48, n. 4, p. 222–229, 1998.

TURKMAN, M. A. A.; PAULINO, C. D.; MÜLLER, P. **Computational bayesian statistics: an introduction**. [S.l.]: Cambridge University Press, 2019. v. 11.

VAN BUUREN, S. **Flexible imputation of missing data**. [S.l.]: CRC press, 2018.

APÊNDICE A – Simulação do experimento

Rotina (**R**) utilizada para simulação do experimento em blocos incompletos parcialmente balanceados (PBIB) organizados em látice quadrado ($v = k^2, r = 3, k = 11, b = 33, \lambda_1 = 1$ e $\lambda_2 = 0$). Rotina adicional da metodologia utilizada para simular a censura à esquerda.

```
### Pacotes utilizados:
library(ggplot2)
library(dplyr)
library(ggrepel)
library(ggthemes)
library(tidyverse)

### Estrutura do delineaemento:
(pX <- t(matrix(1:121,11,11)))
(pY <- t(pX))
pZ <- pY
for(i in 1:11){for(j in 1:11){k <- (i-1+j)%%11
if(k==0){k <- 11}
pZ[i,j] <- pY[k,j]}}

### Valores paramétricos de variâncias e efeitos simulados:
# Variância do Erro:
ve <- 3

# Variância de Tratamentos:
vg <- 1

# Efeito de bloco - escala normal:
Block <- factor(kronecker(1:(3*11),rep(1,11))) # Bloco de tamanho 11.
b  <- nlevels(Block) # número de blocos.
eB <- (1:33-mean(1:33))/sd(1:33)
```

```
plot(eB)
save(eB,file="eB.rda")

# Efeito de tratamento - escala normal:
Treat <- factor(c(t(pX),t(pY),t(pZ)))
v  <- nlevels(Treat) # número de tratamentos.
eT <- qnorm((1:121-0.5)/121)
plot(eT)
save(eT,file="eT.rda")

### Modelo da simulação do experimento para obtenção
### dos valores paramétricos:
XT <- model.matrix(~-1+Block)
ZB <- model.matrix(~-1+Treat)
mu <- 1
Ypar <- mu+XT%*%eB+ZB%*%eT*sqrt(vg)
erro.par <- rnorm(n,0,sqrt(ve))
yesc <- Ypar + erro.par
cor(yesc,Ypar)
cor <- cbind(yesc, Ypar)
cor <- data.frame(cor)
ggplot(cor,aes(yesc, Ypar)) + geom_point(color = 'black', size = 4) +
theme_bw() +labs(x = "Valores Médios Paramétricos",
y = "Valores Simulados") + theme_light(base_size = 21)

### Dados completos simulados:
Y <- round(exp((yesc - mean(yesc))/sd(yesc)),2)
plot(yesc,Y)

### Metodologia empregada para simular a censura à esquerda:
```

```
# Limite mínimo de pesagem para uma taxa de censura moderada:
pcens <- 0.30
tau <- exp(qnorm(pcens))

# Limite mínimo de pesagem para uma taxa de censura alta:
pcens <- 0.50
tau <- exp(qnorm(pcens))

# Gerando dados de produção com distribuição lognormal
# e censura à esquerda:
(PGZ <- sum(Y < tau)/length(Y)) # proporção geral de zeros.
y <- Y
quais <- which(Y<tau)
y[quais] <- 0
mean(y==0)

# Sorteio dos zeros:
table(Treat[quais])

# Proporção de zeros esperada por observação:
pgz <- pnorm(0,0*quais,1)
plot(Y,y)

### Salvando arquivos de dados simulados
save(Ypar,file="y.par.rda")
save(yesc,file="yescaladosefeitos")
save(y,file="y.rda")
save(Y,file="Ycmp.rda")
```

APÊNDICE B – Análise REML

Rotina (**R**) utilizada para estimação das componentes de variânciae e predição dos valores genéticos pelo método da máxima verossimilhança restrita (REML) para as formas de análise "DC", "C0" e "CE".

```
### Pacotes utilizados:
library(ggrepel)
library(coda)
library(xtable)
library(tidyverse)
library(lme4)
library(MASS)
library(lmerTest)
library(emmeans)
library(gridExtra)

### Leitura de dados:
load("y.par.rda") # y paramétrico.
load("yescaladosefeitos") # y simulado.
load("Ycmp.rda") # valor estandartizado, sem censuras.
load("y.rda") # y observado.
load("eT.rda") # parâmetros.
load("eB.rda") # parâmetros.

### Definição do delinemanto:
dados$Bloc <- factor(dados$Block)
dados$Trat <- factor(dados$Treat)
dados$Prod <- y
attach(dados)

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

```
###                Início da análise
### Supondo que se conheça os dados completos ("DC"):
#####################################################

boxcox(Y ~ Bloc + Trat)
yt <- log(Y)
AO <- max(yesc)-min(yesc)
AT <- max(yt)-min(yt)
yt <- min(yesc)+(yt-min(yt))*AO/AT
plot(Y,yt)
cbind(yesc,yt)
table(Trat)
r <- mean(table(Trat))
modelo <- lmer(log(Y) ~ Bloc + (1|Trat)); modelo
anova(modelo)
saida <- summary(modelo)
nug <- nlevels(Trat)
vg <- unlist(saida[[13]])
LIvg <- vg*(nug/qchisq(0.975,nug))
LSvg <- vg*(nug/qchisq(0.025,nug))
ve <- unlist(saida[[11]])^2
nue <- anova(modelo)[1,4]
LIve <- ve*(nue/qchisq(0.975,nue))
LSve <- ve*(nue/qchisq(0.025,nue))
svg <- vg*rchisq(10000,nug)/nug
sve <- ve*rchisq(10000,nue)/nue
svh2 <- sort(svg/(svg+sve))
LIh2 <- svh2[250]
LSh2 <- svh2[10000-250]
svh2m <- sort(svg/(svg+sve/r))
```

```
LIh2m <- svh2m[250]
LSh2m <- svh2m[10000-250]
h2 <- vg/(vg+ve)
h2m <- vg/(vg+ve/r)
Parâmetros_Geneticos <- t(matrix(c(1,vg,LIvg,LSvg,
3,ve,LIve,LSve,
1/4,h2,LIh2,LSh2,
1/(1+3/3),h2m,LIh2m,LSh2m),4,4))

### Componenetes de variância e herdabilidade do modelo:
xtable(Parametros_Geneticos)

### Efeitos aleatórios:
rand.model <- ranef(modelo,condVar=TRUE)
BLUP <- as.vector(unlist(rand.model[[1]]))
sd_BLUP <- sqrt(as.vector(unlist(attr(rand.model[[1]],"postVar"))))
LI_BLUP <- BLUP-1.96*sd_BLUP
LS_BLUP <- BLUP+1.96*sd_BLUP
ERROS <- residuals(modelo)
plot(density(ERROS))
BLUP_EST <- data.frame(cbind(eT,BLUP,sd_BLUP,LI_BLUP,LS_BLUP))
xtable(BLUP_EST, digits = 4)

# Limite máximo esperado para a correlação genética:
round(sqrt(0.5),5)

# Valor observado rho(g,g_hat):
round(cor(eT,BLUP),5)
# Porcentagem de recuperação da informação genética para esta análise:
round(cor(eT,BLUP)/sqrt(0.5),5)
```

```
cor.test(eT,BLUP) # Pearson.

# ou usar
cor.test(eT,BLUP,
method = "spearman")  # para a correlação de Spearman.

########## FIM da análise "DC" ##########

#####################################
###       Análise com censuras
### Substituindo censuras por 0 ("C0"):

# Cenário 1:
boxcox(y+0.3 ~ Bloc + Trat)
yt <- log(y + 0.3)

# ou

# cenário 2:
boxcox(y+0.001 ~ Bloc + Trat)
yt <- log(y + 0.001)

#########################################
# Daqui em diante segue-se o mesmo código:
###############

AO <- max(yesc)-min(yesc)
AT <- max(yt)-min(yt)
yt <- min(yesc)+(yt-min(yt))*AO/AT
```

```
plot(y,yt)
cbind(y,yt)
table(Trat)
r <- mean(table(Trat))
modelo <- lmer(yt ~ Bloc + (1|Trat))
anova(modelo)
saida <- summary(modelo)
nug <- nlevels(Trat)
vg <- unlist(saida[[13]])
LIvg <- vg*(nug/qchisq(0.975,nug))
LSvg <- vg*(nug/qchisq(0.025,nug))
ve <- unlist(saida[[11]])^2
nue <- anova(modelo)[1,4]
LIve <- ve*(nue/qchisq(0.975,nue))
LSve <- ve*(nue/qchisq(0.025,nue))
svg <- vg*rchisq(10000,nug)/nug
sve <- ve*rchisq(10000,nue)/nue
svh2 <- sort(svg/(svg+sve))
LIh2 <- svh2[250]
LSh2 <- svh2[10000-250]
svh2m <- sort(svg/(svg+sve/r))
LIh2m <- svh2m[250]
LSh2m <- svh2m[10000-250]
h2 <- vg/(vg+ve)
h2m <- vg/(vg+ve/r)
Parametros_Geneticos <- t(matrix(c(1,vg,LIvg,LSvg,
3,ve,LIve,LSve,
1/(1+3),h2,LIh2,LSh2,
1/(1+3/3),h2m,LIh2m,LSh2m),4,4))
```

```
# Componenetes de variância do modelo:
xtable(Parametros_Geneticos)

# Efeitos aleatórios:
rand.model <- ranef(modelo,condVar=TRUE);
BLUP <- as.vector(unlist(rand.model[[1]]))
sd_BLUP <- sqrt(as.vector(unlist(attr(rand.model[[1]],"postVar"))))
LI_BLUP <- BLUP-1.96*sd_BLUP
LS_BLUP <- BLUP+1.96*sd_BLUP
ERROS <- residuals(modelo)
plot(density(ERROS))
BLUP_EST <- data.frame( cbind(eT,BLUP,sd_BLUP,LI_BLUP,LS_BLUP))
xtable(BLUP_EST)

# Valor observado rho(g,g_hat):
round(cor(eT,BLUP),4)
# Porcentagem de recuperação da informação genética para esta análise:
round(cor(eT,BLUP)/sqrt(0.5),4)

cor.test(eT,BLUP) # Pearson.

# ou usar
cor.test(eT,BLUP,
method = "spearman")  # para a correlação de Spearman.

########## FIM da análise "C0" ##########

####################################
###       Análise com censuras
### Eliminando censuras à esquerda ("CE"):
```

```
# Cenário 1:
pcens <- 0.30
tau <- exp(qnorm(pcens))
dados$yesc <- yesc
dados
dados2 <- dados[which(Prod>tau),]

bc <- boxcox(Prod+0.2 ~ Bloc + Trat,data=dados2)
(lambda <- bc$x[which(bc$y==max(bc$y))])
Pt <- (dados2$Prod^lambda-1)/lambda
plot(Prod,Pt)
dados2$Pt <- min(dados2$Prod)+(Pt-min(Pt))*(max(dados2$Prod)-
min(dados2$Prod))/
(max(Pt)-min(Pt))
plot(Prod,dados2$Pt)
boxcox(Pt ~ Bloc + Trat,data=dados2)

# ou

# Cenário 2:
pcens <- 0.50
tau <- exp(qnorm(pcens))
dados$yesc <- yesc
dados
dados2 <- dados[which(Prod>tau),]

detach(dados)
attach(dados2)
```

```
bc <- boxcox(Prod+0.1 ~ Bloc + Trat,data=dados2)
(lambda <- bc$x[which(bc$y==max(bc$y))])
Pt <- (dados2$Prod^lambda-1)/lambda
plot(Prod,Pt)

dados2$Pt <- min(dados2$Prod)+(Pt-min(Pt))*(max(dados2$Prod)-
min(dados2$Prod))/
(max(Pt)-min(Pt))
plot(Prod,dados2$Pt)
boxcox(Pt ~ Bloc + Trat,data=dados2)

########################################
# Daqui em diante segue-se o mesmo código
###############

yt <- Pt
AO <- max(dados2$yesc)-min(dados2$yesc)
AT <- max(yt)-min(yt)
yt <- min(dados2$yesc)+(yt-min(yt))*AO/AT
plot(dados2$y,yt)
cbind(dados2$yesc,yt)
table(dados2$Trat)
(r <- mean(table(dados2$Trat)))
modelo <- lmer(yt ~ Bloc + (1|Trat),data=dados2)
anova(modelo)
saida <- summary(modelo)
nug <- nlevels(Trat)
vg <- unlist(saida[[13]])
LIvg <- vg*(nug/qchisq(0.975,nug))
LSvg <- vg*(nug/qchisq(0.025,nug))
```

```
ve <- unlist(saida[[11]])^2
nue <- anova(modelo)[1,4]
LIve <- ve*(nue/qchisq(0.975,nue))
LSve <- ve*(nue/qchisq(0.025,nue))
svg <- vg*rchisq(10000,nug)/nug
sve <- ve*rchisq(10000,nue)/nue
svh2 <- sort(svg/(svg+sve))
LIh2 <- svh2[250]
LSh2 <- svh2[10000-250]
svh2m <- sort(svg/(svg+sve/r))
LIh2m <- svh2m[250]
LSh2m <- svh2m[10000-250]
h2 <- vg/(vg+ve)
h2m <- vg/(vg+ve/r)
Parametros_Geneticos <- t(matrix(c(1,vg,LIvg,LSvg,
3,ve,LIve,LSve,
1/4,h2,LIh2,LSh2,
1/(1+3/3),h2m,LIh2m,LSh2m),4,4))

# Componenetes de variância do modelo:
xtable(Parametros_Geneticos)

# Efeitos aleatórios:
rand.model <- ranef(modelo,condVar=TRUE);
BLUP <- rand.model[[1]]
as.numeric(rownames(BLUP))
eT[as.numeric(rownames(BLUP))]
cbind(BLUP,eT[as.numeric(rownames(BLUP))])
sd_BLUP <- sqrt(as.vector(unlist(attr(rand.model[[1]],"postVar"))))
LI_BLUP <- BLUP-1.96*sd_BLUP
```

```
LS_BLUP <- BLUP+1.96*sd_BLUP
ERROS <- residuals(modelo)
plot(density(ERROS))
BLUP_EST <- cbind(eT[as.numeric(rownames(BLUP))],
BLUP,sd_BLUP,LI_BLUP,LS_BLUP)
xtable(BLUP_EST)

# Valor observado rho(g,g_hat):
round(cor(eT[as.numeric(rownames(BLUP))],BLUP),4)
# Porcentagem de recuperação da informação  genética para esta análise:
round(cor(eT[as.numeric(rownames(BLUP))],BLUP)/sqrt(0.5),4)

# ou usar
cor.test(c(unlist(BLUP)),eT[as.numeric(rownames(BLUP))],
method = "spearman") # para a correlação de Spearman.

########## FIM da análise "CE" ##########
```

APÊNDICE C – Inferência Bayesiana

Rotina (**R**) utilizada para amostragem condicional dos dados censurados implementando a Amostragem Gibbs a partir das distribuições condicionais completas *a posteriori*. Rotina adicional de diagnóstico das cadeias e resumo da análise.

```
### Carregando pacotes R:
library(MASS)
library(msm)
library(dplyr)
library(ggrepel)

### Leitura de dados
load("y.par.rda") # y paramétrico.
load("yescaladosefeitos") # y simulado.
load("Ycmp.rda") # valor estandartizado, sem censuras.
load("y.rda") # y observado.
load("eT.rda") # parâmetros.
load("eB.rda") # parâmetros.

############################################################
### Carregando (ou ignorando ) a "Numerator Relationship":
A <- diag(rep(1,121))

### Caso houvesse parentesco registrado na matriz A,
### gravada no arquivo "A.rda", trocar a linha anterior por:
### load("A.rda")
Ai <- solve(A)

################################################
# Modelagem com médias de blocos na parte fixa:
X  <- model.matrix(~ -1+Block)
```

```
Z  <- model.matrix(~ -1+Treat)
p  <- b  # número de blocos (com médias supostas fixas)
q  <- v  # número de níveis de efeitos aleatórios

# Prioris
nuu <- 2
s2u <- 5
nue <- 2
s2e <- 5

# Valores iniciais arbitrários:
eta <- log(y+0.1)
eta[y>0] <- log(y[y>0])
eta[quais] <- log(0.1)
plot(Y,eta)
plot(y,eta)

u    <- ginv(t(Z)%*%Z)%*%t(Z)%*% eta
beta <- ginv(t(X)%*%X) %*% t(X)%*% eta
e    <- eta - (X%*%beta + Z%*%u)
vu   <- as.double(var(u))
ve   <- as.double(var(e))

# Decomposiçãoo de Cholesky da matrix X'X:
Fe   <- ginv(t(X)%*%X)
cF   <- chol(Fe)

####################################
# Definições para o laço MCMC:
# Parâmetros de controle da Cadeia:
```

```
BI     <- 1000         # "burn in".
TH     <- 10           # "thinning".
NEC    <- 10000        # effective sample size.
NTotal <- BI + NEC*TH  # número total de passos de amostragem.

### Início do laço:
for(i in 1:NTotal){
  gama <- as.double(ve/vu) # razão das componentes de variância.
  Ag   <- gama * Ai # inversa da matriz de parentesco.

# Atualizando valores de eta
  mu_eta     <- X%*%beta + Z%*%u # média dos valores de eta.
  pgz        <- pnorm(tau,mu_eta[quais],sqrt(ve))
  sorteio    <- runif(length(quais))
  eta[quais] <- qnorm(pgz*sorteio, mean=mu_eta[quais],
  sd=sqrt(ve))

# Prevenção de quebra para números muito discrepantes
  if(sum(eta< -10)>0){eta[eta< -10] <- -10}

# Atualiza o vetor de efeitos fixos:
  meanf <- ginv(t(X)%*%X)%*%t(X)%*%(eta - Z%*%u)
  kf    <- rnorm(p)
  beta  <- sqrt(ve) * t(cF) %*% kf + meanf

# Atualiza o vetor de efeitos aleatórios:
  M     <- solve(t(Z)%*%Z + Ag)
  meanr <- M%*%t(Z)%*%(eta - X%*%beta)
  kr    <- rnorm(q)
  sr    <- sqrt(ve) * chol(M)
```

```
  u     <- t(sr) %*% kr + meanr

# Atualiza a variância residual:
  e  <- eta - X%*%beta-Z%*%u
  c1 <- (n+nue)/2
  c2 <- (t(e)%*%e+nue*s2e)/2
  ve <- rgamma(1,c1,c2)
  ve <- as.double(1/ve)

# Atualiza a variância de tratamentos:
  c1 <- (q+nuu)/2
  c2 <- (t(u) %*% u + nuu*s2u)/2
  vu <- rgamma(1,c1,c2)
  vu <- as.double(1/vu)

# Gravando a cadeia de resultados do Amostrador de Gibbs:
if(((i-BI)%%TH == 0)&(i>BI)){
  print(round(100*i/NTotal,2)) # controle para verificar o andamento
                                                    # da execução interativos.
  amostra <- cbind(ve,vu,t(beta),t(u))
  write(t(amostra),file="cadeia.txt",ncol=length(amostra),append=TRUE)
  write(t(eta[quais]),file="eta_imputados.txt",
  ncol=length(quais),append=TRUE)}
}
### FIM do laço MCMC

################################################
### Diagnóstico das cadeias e resumos da análise:
################################################
library(coda)
```

```
library(xtable)
library(tidyverse)

cadeia <- mcmc(read.table("cadeia.txt"))
raftery.diag(cadeia)
eta <- mcmc(read.table("eta_imputados.txt"))
raftery.diag(eta)

#### Componentes da Variância:
xtable(cbind(c(3,1),round(summary(cadeia)$statistics[1:2,],6),
round(HPDinterval(cadeia)[1:2,],6)))

# Herdabilidades:
h2 <- mcmc(cadeia[,2]/(cadeia[,2]+(cadeia[,1])))
h2m <- mcmc(cadeia[,2]/(cadeia[,2]+
(cadeia[,1]/2))) # herdabilidade experimental
H2 <- rbind(c(1/(1+3),summary(h2)$statistics,HPDinterval(h2)),
c(1/(1+3/3),summary(h2m)$statistics,HPDinterval(h2m)))
round(H2,4)

#### Grau de acerto nos imputados:
quais <- which(y==0)
length(quais)
imputados <- xtable(cbind(round(log(Y[quais]),2),
round(summary(eta)$statistics[1:107,1:2],2),round(HPDinterval(eta),2)))
data.frame(imputados)

# Valor observado rho(g,g_hat):
round(cor(eT,round(summary(cadeia)$statistics[36:156,1],4)),5)
```

```
# Porcentagem de recuperação da informação genética para esta análise:
round(cor(eT,round(summary(cadeia)$statistics[36:156,1],4))/sqrt(0.5),4)

cor.test(eT,round(summary(cadeia)$statistics[36:156,1],4)) # Pearson.

# ou usar
cor.test(eT,round(summary(cadeia)$statistics[36:156,1],4),
         method = "spearman")
# para a correlação de Spearman.
```