Universidade de Évora - Escola de Ciências e Tecnologia

Mestrado em Modelação Estatística e Análise de Dados

Área de especialização | Modelação Estatística e Análise de Dados

Dissertação

Análise estatística da produção de vitelão Mertolengo

Ana Paula Ferrari Januario

Orientador(es) | Patrícia A. Filipe

Gonçalo João Jacinto

Évora 2021

Universidade de Évora - Escola de Ciências e Tecnologia

Mestrado em Modelação Estatística e Análise de Dados

Área de especialização | Modelação Estatística e Análise de Dados

Dissertação

Análise estatística da produção de vitelão Mertolengo

Ana Paula Ferrari Januario

Orientador(es) | Patrícia A. Filipe

Gonçalo João Jacinto

Évora 2021

A dissertação foi objeto de apreciação e discussão pública pelo seguinte júri nomeado pelo Diretor da Escola de Ciências e Tecnologia:

Presidente	|	Anabela Afonso (Universidade de Évora)
Vogais	|	Dulce Gamito Pereira (Universidade de Évora) Patrícia A. Filipe (Universidade de Évora) (Orientador)

Évora 2021

À minha família.

Agradecimentos

A produção deste trabalho foi longa, uma vez que meu contato com a base de dados aconteceu no primeiro semestre do mestrado. Porém, alcançar esta etapa não teria sido possível sem a colaboração, auxílio, carinho e dedicação por parte de várias pessoas ao longo de todo o percurso. Por esta mesma razão, quero aproveitar esta oportunidade para agradecer a todos aqueles que, direta ou indiretamente, contribuíram para que eu chegasse até aqui.

Em primeiro lugar quero agradecer à minha família por me permitir sonhar e voar para o outro lado do Atlântico atrás dos meus objetivos.

À minha mãe, pelo carinho e pela compreensão, mesmo à distância, e apesar de toda a saudade e vontade de ter minha presença física, deu-me força para que eu pudesse chegar até aqui e concluir meu sonho do mestrado. Ao meu pai por me dar forças nos momentos de fraqueza e por ter me ensinado, desde criança, que a melhor coisa que podemos fazer para nós mesmos é investir em nossa educação. Ao meu irmão pelo auxílio, pela "presença online" e pelas sugestões, sempre tentando me ajudar. Obrigado por tudo, não consigo nem imaginar como seria minha vida sem vocês.

À minha avó Hilda por todo o carinho e pela compreensão ao longo desses anos de ausência. Peço desculpas, pois não pude estar presente fisicamente no seu aniversário de 90 anos, mas foi por uma boa causa. Os frutos foram colhidos e a senhora está sempre presente na minha mente e no meu coração. À minha tinha Lúcia, por me manter informada sobre o estado da família em tempos de pandemia e por cuidar de todos com muito carinho.

Também quero aproveitar para agradecer aos meus "irmãos do lado de cá do Atlântico": Rodrigo, Finório e Fernando. Obrigado pela presença, pelo carinho e pelas brigas, pois isso também faz a gente crescer como ser humano. Obrigado por me ensinarem a pensar como estatísticos e economistas e me mostrarem o mundo com outro ponto de vista. E principalmente, obrigado por estarem lá no momento que eu "caí". Os golpes da vida vêm quando a gente menos espera, mas a força que vocês me deram foi essencial para que eu não desistisse de tudo!

Quero agradecer a Associação de Criadores de Bovinos da Raça Mertolenga (ACBM) e ao projeto GoBovMais por cederem a base de dados que deu origem a este estudo. E em especial ao Engenheiro José Pais, por toda a disponibilidade e compartilhamento de informações ao longo deste período de quase dois anos de trabalho.

Agradeço a todos os professores do curso de mestrado em modelação estatística e análise de dados da Universidade de Évora, que têm essa missão tão bela que é compartilhar conhecimento, e me

ensinaram muito no decorrer deste mestrado. Em especial à professora Dra. Patrícia Filipe por ter aceitado o desafio de orientar uma pessoa que é de fora da matemática, obrigada pela paciência e por ter sido amiga e compreensiva ao longo deste processo. Ao professor Dr. Gonçalo Jacinto, por ter compartilhado comigo tanto conhecimento, pela presença e pelas broncas. Pois são nas tensões que eu pude consolidar o conhecimento. À professora Anabela Afonso, por ter sido muito receptiva, desde quando eu ainda estava no Brasil. Obrigada pela disposição em tirar nossas dúvidas e também obrigada pela amizade nos momentos de crise.

Agradeço a Verónica e a Rosana, por me ajudarem a trilhar o caminho do autoconhecimento, a ajuda de vocês foi muito importante para manter o foco nos meus objetivos.

Quero agradecer a Déborah, Suehellen, Solange e todos os demais familiares e amigos que não foram mencionados, mas ficaram torcendo por mim e me apoiaram à distância. Obrigado por acreditarem em minhas capacidades.

Isaac Newton mencionou em uma carta certa vez: "Se vi mais longe foi por estar de pé sobre ombros de gigantes", todos vocês foram meus gigantes que me permitiram subir nos ombros para que eu enxergasse mais longe, e por isso meu MUITO OBRIGADO!

Lista de Acrónimos

ACBM Associação de Criadores de Bovinos da raça Mertolenga

AIC Critério de informação AKAIKE

BIC Critério de Informação Bayesiano

CTR Centro de Testagem e Recria

DOP Denominação de origem protegida

EQM Erro quadrático médio

FAO Food and Agriculture Organization of the United Nations

GMD Ganho Médio Diário

GLM Modelos lineares generalizados

HCCM Matriz de covariância heterocedástica corrigida

INE Instituto Nacional de Estatística

MMQ Método dos mínimos quadrados

MMV Método da máxima verossimilhança

RLM Regressão linear múltipla

ROC Curva característica de operação do receptor

SAU Superfície Agrícola Utilizada

SQR Soma dos quadrados dos desvios explicáveis da regressão

SQT Soma dos quadrados do desvio total

TLC Teorema do limite central

VG Valor genético

Abstract

Statistical analysis of Mertolengo cattle production

The work was intended to support the Association of mertolenga cattle breed in its breeding process and decision making, namely in modeling the cost per day of production of the male mertolenga cattle, and in identifying the variables that favor the sale of the animal as a product with a protected designation of origin (PDO) seal. The database contained information on 716 male animals, of which 54 % went to the slaughter that guarantees the PDO seal. We also had data on the cost structure production of the animals from when it enters into the CTR to slaughter, in addition to the individual characteristics of each animal, in particular, of its estimated breeding value.

To obtain the cost-per-day production model, multiple linear regression models and other generalized linear models were used. For the classification of the animal as a PDO slaughter destination, a logistic regression model was used.

When we comparing the generalized linear models tested, the multiple linear regression model was confirmed as the best technique to explain the cost per day of production. For this model, it was found that information such as weight at entry as well as different estimated breeding value positively influence the cost of production.

With regard to logistic regression, weight at entry, age at entry and genetic values referring to maternal capacity and calving interval are factors that enhance the animal being sold under the PDO seal.

Keywords: Mertolenga breed, costs, multiple linear regression, generalized linear model, logistic regression

Sumário

Com o trabalho desenvolvido nesta dissertação, pretendeu-se apoiar a Associação de produtores de bovinos da raça mertolenga no seu processo de recria e nas tomadas de decisão, nomeadamente na modelação do custo por dia de produção de bovinos machos da raça mertolenga, e na identificação das variáveis que favorecem a venda do animal como um produto com selo de denominação de origem protegida (DOP). A base de dados continha a informação de 716 animais machos, dos quais 54% foram para o abate que garante o selo DOP, dados referentes à estrutura de custo de produção dos animais desde a entrada no CTR até o abate, além das características individuais de cada animal, em particular, dos seus valores genéticos.

Para obter o modelo do custo por dia de produção, utilizou-se modelos de regressão linear múltipla e outros modelos lineares generalizados. Para a classificação do animal por destino de abate DOP, utilizou-se um modelo de regressão logística. Quando se comparou os diferentes modelos lineares generalizados testados, confirmou-se o modelo de regressão linear múltipla como o mais adequado para explicar o custo por dia de produção. Para este modelo, verificou-se que informações como o peso à entrada bem como diferentes valores genéticos influenciam de forma positiva o custo de produção.

No que diz respeito a regressão logística, o peso à entrada, a idade à entrada e os valores genéticos referentes à capacidade maternal e intervalo entre partos são fatores potenciadores do animal ser vendido com o selo DOP.

Palavras chave: Raça Mertolenga, custos, regressão linear múltipla, modelo linear generalizado, regressão logística

1

Introdução

Os bovinos são grandes animais ruminantes e formam a espécie mais comumente difundida no mundo, criados principalmente para a produção de leite, carne, couro e para fornecimento de energia de tração (Food and Agriculture Organization of the United Nations- FAO, 2020). Para a FAO a bovinocultura fornece bens e serviços às pessoas e desempenha um papel social e financeiro importante. Atualmente o Alentejo é detentor da maior proporção de Superfície Agrícola Utilizada de Portugal (INE, 2017). Além disso, segundo o Instituto Nacional de Estatística (INE, 2017), o Alentejo também é o maior produtor de carne bovina, possuindo a maior manada do país. Por haver alta densidade deste tipo de produção animal, percebe-se a importância de estudar a produtividade dos bovinos da região. Nesse contexto surgiu o trabalho desenvolvido com a Associação de Criadores de Bovinos da raça Mertolenga (ACBM).

A ACBM tem sua sede na cidade de Évora, e representa os seus associados perante o Estado e outros organismos nacionais ou estrangeiros. Além disso, de forma geral, defende os interesses dos seus associados no que se relaciona com a preservação, melhoramento genético, criação e comercialização dos bovinos de raça mertolenga. ACBM acompanha os animais que entram na associação e recolhe as informações referente a esses animais e à produção. A base de dados que foi concebida com essas informações foi cedida pela associação para este estudo, no âmbito do projeto Go BovMais (www.bovmais.pt).

1.1 Raça mertolenga e processo de recria

De acordo com Crepaldi (2005, p.213) a atividade da bovinocultura pode ser classificada em fases de desenvolvimento do animal como: cria (foco na produção do bezerro que é vendido após o desmame, e normalmente ocorre uma vez ao ano), recria (com o objetivo de produzir e vender o vitelo ou novilho, a partir do bezerro desmamado) e acabamento ou engorda (a partir do novilho ou vitelo magro, produz-se o boi gordo para venda).

A raça mertolenga é caracterizada por ser uma raça pequena, enérgica e bem-adaptada à região do Alentejo. Estes animais são desmamados com cerca de 6 a 8 meses e a ACBM realiza as fases de recria e acabamento de jovens machos dessa raça, provenientes de criadores seus associados, cujas explorações agrícolas se situam maioritariamente no Alentejo. As ações acontecem no Centro de Testagem e Recria da Herdade dos Currais e Simalhas (CTR).

Segundo o Engenheiro José Pais, diretor executivo da ACBM, o trabalho da associação é manter as características que a raça mertolenga tem hoje e adquiriu ao longo da evolução ocorrida com alguns cruzamentos indeterminados há dezenas de anos. Uma vez que esta raça produz vacas muito bem adaptadas ao meio ambiente em que é explorada, apresentando boa capacidade maternal e facilidade de parto. O trunfo dos animais de raça mertolenga é o de possuir características genéticas que podem ser usadas como linha materna em cruzamentos dirigidos, para a produção de fêmeas reprodutoras mais adaptadas. Portanto, o trabalho desenvolvido pela associação tem sido no sentido de melhorar a conformação e um pouco a capacidade de crescimento após o desmame sem querer, necessariamente, aumentar o peso adulto.

Com relação à Carne Mertolenga que possui o selo de denominação de origem protegida (DOP), esta possui características organolépticas próprias e alguma infiltração de gordura intra-muscular, sendo o marmoreado de dispersão médio. A Direção Geral de Agricultura e Desenvolvimento Rural (2020) especifica que esta carne possui cor rosa escura e pode apresentar gordura branca ou amarela, conforme se trate de vitela ou animal adulto.

A recria dos machos mertolengos é feita em modos semi-intensivos/intensivos e o sistema de alimentação, na maioria dos casos, é constituído por concentrado e palha ou feno. No CTR, o sistema de alimentação recorre à distribuição de uma mistura feita com unifeed, à base de silagem de milho, silagem de consociação gramíneas/leguminosas, feno, feno-silagem e farinado para complementar o perfil nutricional que se pretende obter.

No CTR também acontece o teste de machos mertolengos para reprodutores. Os machos testados, são escolhidos à entrada tendo em atenção os valores genéticos dos mesmos. Dão-se especial atenção aos valores genéticos para o intervalo entre partos e capacidade maternal (Carolino et al., 2016).

Pais et al. (2019) informa que nos últimos anos a ACBM tem produzido dois tipos de animais: o vitelão Mertolengo DOP, com idade ao abate entre 10 e 15 meses e peso mínimo da carcaça de 120 kg (sem limite superior) e o vitelão convencional (termo utilizado pela ACBM para distinguir do produto DOP), com idade ao abate entre 8 e 12 meses (até um dia antes de completar 12 meses) e peso de carcaça entre 120 e 250 kg.

O DOP é um selo que consiste na utilização do nome duma região ou localidade, para designar que um produto é dela originário, e à partida, garante uma valorização ao produto que o detém.

Essa ação de produção da ACBM está relacionada com a fraca valorização dos jovens animais provenientes da manada de raça Mertolenga com linhagem pura. E assim, segundo Pais et al. (2019), pretende-se possibilitar aos criadores a realização da recria e acabamento dos seus vitelos com um valor econômico superior ao que normalmente é atingido com as vendas no desmame. Além de ajudar na recria

Figura 1.1: Bovino da raça mertolenga. Fotografia disponível na site da ACBM.

e acabamento quando a exploração não tem condições técnicas/económicas para isso.

1.2 Motivação

Segundo o Instituto Nacional de Estatística (INE, 2017) o Alentejo é o maior produtor de carne bovina de Portugal, possuindo também a maior manada do país. Por haver essa alta densidade, percebe-se a importância de estudar a produtividade dos bovinos na região do Alentejo.

Dentro da produção animal e entre as espécies mais produzidas no mundo, o bovino é o animal que apresenta o maior porte, que necessita de mais alimento do ponto de vista unitário e que ocupa mais espaço na terra, cerca de um terço da terra agrária do planeta (FAO, 2014).

A produção bovina mobiliza, assim, um número significativo de recursos e por essa razão a sustentabilidade de produção é tão questionada. Para a FAO as previsões atuais indicam que o consumo de carne em todo o mundo dobrará nos próximos 20 anos. E ainda que esta seja uma boa notícia em termos da segurança alimentar de milhões de pessoas, para satisfazer tal demanda a fronteira agrícola e pecuária, cada vez mais, é empurrada para áreas de maior vulnerabilidade ambiental, e isto preocupa os consumidores com consciência ambiental.

Para a FAO (2014), a pecuária pode desempenhar um papel importante tanto na adaptação às alterações climáticas como para atenuar os seus efeitos sobre o bem-estar da humanidade e sugere que as raças adaptadas apresentam potencial para melhorar substancialmente a produtividade a nível global e contribuir para a redução da pobreza. Quando refere-se a produção de raças adaptadas, pode-se pensar como adaptação vinda do melhoramento genético, ou através de raças nativas da região como é o caso da raça mertolenga.

A FAO (2019) prevê ainda que o consumo de carne mundial deve dobrar até 2050, e segundo o INE (2020), o consumo de carne bovina em Portugal vêm aumentando de forma gradual desde 2015.

A motivação que deu origem a este estudo é o contexto mundial da bovinocultura, e também, por haver possibilidade de impacto desse tipo de produção no mercado consumidor português nos próximos anos, visto os hábitos de consumo de carne bovina por parte da população. Por isso, esse trabalho pode vir a oferecer bases aos criadores e à associação para o processo crítico de tomada de decisão através do estudo detalhado do custo por dia de produção.

As técnicas de modelação utilizadas foram avançando à medida que era observado o comportamento das variáveis de interesse. Para esta dissertação foram usados tipos de modelos diferentes: modelos de regressão linear múltipla e modelos lineares generalizados.

As técnicas foram aplicadas na tentativa de responder as perguntas que apareciam no processo de modelação e que deram base à conclusão deste trabalho, como por exemplo: Quais são as variáveis que impactam as chances de um animal ir ao abate que garante o selo DOP? Essas variáveis têm impacto sobre o custo de produção por dia? Será que um modelo com distribuição de probabilidade diferente pode explicar melhor o custo de produção por dia? Será que a função de ligação apresenta efeito sobre o custo? E à medida que as perguntas foram respondidas, estruturou-se esta dissertação.

1.3 Objetivos

Procurou-se com este trabalho, realizar uma análise estatística dos dados técnicos e económicos da exploração de bovinos da raça mertolenga, usando para o efeito a base de dados da associação de criadores da raça mertolenga que se considerou como caso de estudo. Por via desta base analisada pode-se entender e caracterizar os animais e a dinâmica de produção dos seus associados que se situam maioritariamente no Alentejo. Neste sentido, pretendeu-se na análise construir modelos que expliquem a dinâmica do custo de produção em função de certas variáveis de interesse à associação e aos criadores, fornecendo informações mais detalhadas sobre quais as variáveis genéticas com os dados conhecidos à entrada do animal na recria, constituem um aumento dos custos diários de produção e com isso a diminuição do lucro. Numa segunda fase, procurou-se construir um modelo logístico, cuja variável resposta corresponde ao destino de abate (DOP ou convencional).

1.4 Estrutura do trabalho

Este trabalho está organizado da seguinte forma:

O processo da recolha e o tratamento aplicado aos dados serão detalhados no capítulo 2.

No capítulo 3, é apresentada uma análise descritiva e inferencial das principais variáveis presentes na base de dados.

No capítulo 4, é apresentado o modelo de regressão linear múltipla para a variável custo por dia de produção e os resultados da avaliação dos pressupostos exigido por essa técnica.

No capítulo 5, é apresentado um modelo linear generalizado do custo por dia de produção com 3 tipos diferentes de distribuição de probabilidade e diferentes funções de ligação. E também são apresentados os resultados de análise dos resíduos, bem como a validação dos seus pressupostos.

No capítulo 6, é apresentado um modelo de regressão logística que investigou as características que o animal deve possuir para garantir o selo DOP (mercadoria com denominação de origem protegida), que a princípio, dá ao produtor maior lucratividade na produção.

No capítulo 7, são apresentadas as conclusões do melhor modelo do custo por dia de produção, e também, as conclusões sobre as variáveis que impactam as chances do animal ir apara o abate tipo DOP.

No final também há 3 anexos, nos quais são apresentados alguns scripts e gráficos utilizados para obter os resultados apresentados nos capítulos 3 (anexo A), 4 (anexo B) e 5 (anexo C).

2

Metodologia estatística

Com o intuito de apresentar um modelo aos produtores com as variáveis mais significativas para custo do processo produtivo, utilizou-se técnicas de regressão que serão explicadas a seguir. Os modelos são os principais instrumentos utilizados na estatística, e, segundo Braumann (2019) a união da teoria e aplicação é uma ferramenta poderosa na modelagem matemática e contribui para uma melhor compreensão da realidade e suas motivações. Ainda Braumann (2005) diz que os modelos são uma versão simplificada de algum problema ou situação da vida real e destinam-se a ilustrar certos aspectos de um fenómeno.

O modelo de regressão linear descreve a relação da média da variável resposta e um conjunto de variáveis explicativas de forma linear, assumindo que a inferência é efetuada com uma distribuição normal para a variável resposta. Os GLM extendem os modelos de regressão linear para incorporar distribuições não normais na variável resposta e inclusive possíveis funções não lineares da média (Agresti, 2015). A escolha entre um modelo ou outro dá-se pela observação da variável resposta e das avaliações dos resíduos dos modelos obtidos.

A seleção do melhor modelo é a tarefa de escolher um modelo estatístico a partir de um conjunto de modelos plausíveis. Segundo Konish & Kitagawa (2008), um boa técnica para seleção de modelos equilibrará qualidade do ajuste e complexidade. Modelos mais complexos poderão melhor adaptar a sua

forma para ajustar-se aos dados, entretanto um modelo com muitos parâmetros pode não explicar nada de útil.

Este capítulo está organizado em 3 secções. Na secção 2.1 é apresentada a metodologia de recolha de dados da ACBM. Na secção 2.2 é apresentado o modelo de regressão linear múltipla, os seus pressupostos e os métodos de estimação dos parâmetros, método dos mínimos quadrados e método da máxima verossimilhança. Ainda nesta secção são apresentadas técnicas para quando os pressupostos da regressão linear múltipla não se verificam. Na secção 2.3 são apresentados os modelos lineares generalizados (GLM).

2.1 Metodologia da recolha de dados da ACBM

Nesta secção é apresentado como é feito o levantamento das informações que compõem a base usada no estudo. Como já informado anteriormente a ACBM acompanha os animais que entram na associação e recolhe de forma regular, informações relativas ao animal, como por exemplo, idade, peso, criador de origem e etc. E também organiza as informações da genealogia acumulada que vão para o livro genealógico (Capacidade maternal, ganho de peso médio diário no teste de performance, intervalo entre partos e etc).

Para melhorar uma determinada característica produtiva do rebanho, o criador deverá identificar os animais que, geneticamente, são superiores e selecioná-los para que sejam os antecessores das futuras crias. Ao selecionar os animais geneticamente superiores é possível aumentar o termo médio da produção das gerações seguintes. E assim, ao longo das gerações, o valor produtivo dos animais irá aumentando e com ele os benefícios económicos associados.

A avaliação genética baseia-se nos registos de partos, abates e pesagens realizados pela ACBM e pelos criadores, assim como nas genealogias acumuladas no Livro Genealógico. A avaliação genética da raça bovina mertolenga foi elaborada na Unidade de Recursos Genéticos, Reprodução e Melhoramento Animal do INRB I.P., a partir de toda a informação de campo recolhida. Todos os caracteres selecionados foram submetidos a análises univariadas, através do BLUP - Modelo Animal, utilizando-se para o efeito o programa informático MTDFREML. Esta metodologia permite estimar os valores genéticos de cada animal para os tipos de caracteres, tendo em conta a sua performance, no caso de ser conhecida, e as performances de todos os seus parentes (ascendentes, descendentes e colaterais) (ACBM, 2018). Estas informações genéticas dos animais foram fornecidas pela Ruralbit, empresa parceira da ACBM, e extraídas da base de dados GENPRO.

Estas informações genéticas dos animais encontram-se disponíveis na base de dados GENPRO e foram cedidas no âmbito do projecto Go BovMais - Melhoria da produtividade da fileira dos bovinos de carne. Utilizou-se técnicas de análise de dados e inferencia estatística para caracterizar a base de dados. A análise decorreu com recurso ao software RStudio versão 4.0.2.

2.2 Regressão linear múltipla

Atualmente, a análise de regressão é uma das mais importantes técnicas estatísticas, sendo utilizada em aplicações de diversas áreas. Segundo Rodrigues (2012) na regressão linear múltipla assume-se que existe uma relação linear entre uma variável dependente (Y) e uma certa quantidade de variáveis independentes preditoras $(X_j, j = 1, 2, \ldots, k)$. E que através dessa relação é possível construir uma equação para se estimar os valores esperados da variável Y, dado os valores das variáveis independentes $X_j, j = 1, 2, \ldots, k$. Sharma (1996) completa ao referir que a regressão linear múltipla possui três principais objetivos: descrição, controlo e previsão.

Para Pek et al., (2018) a regressão linear múltipla (RLM) é um quadro analítico geral em que o teste t e a ANOVA são casos especiais. Através dos valores observados obtem-se a equação de regressão linear para estimação dos valores esperados e apresenta-se em notação como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

De modo que:
Y é a variável dependente que o modelo tenta explicar,
β_0 é a constante que representa a intercepção da reta com o eixo,
β_1, β_2, ..., β_k representam a inclinação (coeficiente angular) em relação as variáveis independentes $X_j, j = 1, 2, \ldots, k$.
$X_j, j = 1, 2, \ldots, k$ representam as variáveis independentes.
ϵ representa os fatores resíduais e os possíveis erros de medição. Para que este tipo de modelo seja corretamente utilizados, esses fatores resíduais devem satisfazer alguns pressupostos.

Segundo Afonso & Nunes (2019), durante o processo de modelação, para verificar se o modelo obtido é significativo, testa-se o modelo de regressão múltipla com um teste F da significância global do modelo, que avalia a vários coeficientes ao mesmo tempo. As hipóteses desse teste são:
H_0: $\beta_1 = \beta_2 = \ldots = \beta_k$ =0
H_1: Pelo menos um $\beta_j \neq 0$, j = 1, 2, 3, ..., k
Se o p-valor para o teste F apresentar significância pode-se rejeitar a hipótese nula e concluir que os coeficientes do modelo são significativos e ajudam a explicar a variável resposta.

2.2.1 Pressupostos para a regressão linear

Para avaliar os pressupostos, avaliam-se os resíduos do modelo. O resíduo é definido para uma observação como a diferença entre o seu valor real e o ajustado ($\epsilon_i = y_i$ - $\hat{y}_i$) (Afonso & Nunes, 2019).

Segundo Pek et al. (2018) para possibilitar a inferência do modelo de regressão desenvolvido é necessário que o mesmo cumpra 5 pressupostos. O primeiro é que a relação entre a variável dependente e as variáveis independentes deve ser linear. Para esse tipo de modelo as variáveis independentes são controladas e não estão sujeitas a variações. O segundo pressuposto prevê que os resíduos do modelo obtido deve seguir a normalidade [i.e., $\epsilon_i \sim$ N(0,σ^2)], para a verificação desse pressuposto foi realizado o teste de Kolmogorov-Smirnoff com correção Lilliefors (que é o indicado para amostras grandes n≥ 30). E este fato ($\epsilon_i \sim$ N) assegura que a distribuição amostral das estimativas segue uma distribuição t quando o σ é estimado. Ou quando a dimensão da amostra (N) é grande o suficiente a distribuição amostral será aproximadamente normal, de acordo com o teorema do limite central (TLC).

O terceiro pressuposto diz que é necessário verificar se os erros apresentam variância finita e a mesma variabilidade em torno dos níveis das variáveis independentes (homocedasticidade), para verificar este pressuposto foi utilizado o teste de Breusch-Pagan. Este teste é indicado para grandes amostras (como é o caso da base de dados em estudo) e assume a suposição de normalidade dos erros.

A multicolinearidade refere-se as variáveis preditoras estarem relacionadas entre si, e com isso, o modelo apresenta informações, de certa forma, redundantes. Para a verificação desse pressuposto, realizou-se o teste VIF (Variance Inflation Factor) que mede o quanto da variância de cada coeficiente de regressão do modelo estatístico se encontra inflado, devido à multicolinearidade nos estimadores calculados pelo método dos mínimos quadrados.

E por último, a amostra deve ser independente e identicamente distribuída (i.i.d.), para além disso, os resíduos também devem ser independentes. Para avaliar este pressuposto utilizou-se do teste de Durbin-Watson que é utilizado para detectar a presença de autocorrelação (dependência) nos resíduos de uma análise de regressão.

Resumindo os pressupostos e as observações registadas, os erros ϵ devem seguir uma distribuição normal de média zero e variância homogénea. Além de se verificar ausência de multicolinearidade e de autocorrelação (Bussabi & Morettin, 2012).

2.2.2 Método dos mínimos quadrados (MMQ)

Neste trabalho os parâmetros da regressão linear múltipla foram obtidos pelo método dos mínimos quadrados (MMQ), que segundo Agresti (2015) é a abordagem padrão para esse tipo de técnica de modelação. O MMQ minimiza a soma dos quadrados dos erros e os seus coeficientes possibilitam estimar a mudança prevista em Y por unidade de mudança num X_j específico, mantendo constante o efeito das outras variáveis $X_i (i \neq j)$ (Berenson et al., 2012). Com a suposição adicional de função de ligação identidade (Agresti, 2015).

Para Agresti (2015), minimizar uma soma de quadrados é matemáticamente e computacionalmente a forma mais simples de se estimar parâmetros, e também é a melhor classe de estimadores não enviesados. Segundo Pek et al. (2018) a obtenção de estimativas para os parâmetros na equação obtida, não requer nenhuma premissa de distribuição. Entretanto, Agresti (2015) demonstra que usar o MMQ corresponde a máxima verossimilhança quando é adicionado o pressuposto da normalidade ao modelo, uma vez que para maximizar o log da verossimilhança precisa de minimizar o $\sum(y_i - \mu_i)^2$.

2.2.3 O método da máxima verossimilhança (MMV)

O método da máxima verosimilhança consiste em maximizar o valor da função de verosimilhança para uma determinada amostra, ou seja, maximiza a probabilidade com base nos dados observados, obtendo o valor de θ mais provável (Meyer, 1983). Para Verbek (2017) esse método garante propriedades assintóticas para consistência, eficiência e normalidade.

Definição: Seja $\{X_1, X_2, \ldots, X_n\}$ uma amostra aleatória independente e identicamente distribuída (i.i.d.), de tamanho n da variável aleatória X e de valores amostrais $x_1, \ldots, x_n$, com função densidade de probabilidade (f.d.p.) g (x; $\theta_1, \theta_2, \ldots, \theta_k$) = g (x | $\boldsymbol{\theta}$), com $\boldsymbol{\theta}(\theta_1, \theta_2, \ldots, \theta_k) \in \Omega$, que é o espaço paramétrico. Considerando uma amostra em concreto, designa-se por função de verosimilhança a função de $\boldsymbol{\theta}$ correspondente à amostra aleatória observada tal que:

$$L(\boldsymbol{\theta}; x_1, x_2, ..., x_n) = \prod_{i=1}^{n} g(x_{ij}; \boldsymbol{\theta}) = g(x_{1j}; \boldsymbol{\theta}) g(x_{2j}; \boldsymbol{\theta}) \ldots g(x_{nj}; \boldsymbol{\theta})$$

Portanto, a função de verossimilhança no contexto da regressão linear múltipla representa a probabilidade de todos os individuos da amostra se comportarem de acordo com a função f (assumindo a independência) (Verbek, 2017). O estimador resulta da maximização de L ou da função de log- verossimilhança.

$$LL(\boldsymbol{\theta}; x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \log(g(x_i); \boldsymbol{\theta})$$

Para o contexto da RLM com $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_k)$

2.2.4 Técnicas de seleção do modelo

A variabilidade total no conjunto de dados pode ser decomposta em duas componentes: uma explicada pela regressão e outra que não é explicada pela reta de regressão (Afonso & Nunes, 2019). O modelo que melhor se ajusta é aquele que consegue explicar melhor os dados utilizando a menor quantidade de variáveis possíveis (princípio da parcimónia). Existem vários métodos para avaliar se o modelo linear postulado é adequado ou não, dependendo das suposições que fizermos sobre ele. Para esse trabalho a seleção do melhor modelo no método clássico considerou a análise dos resíduos, o coeficiente de determinação (R^2) e o menor valor do critério de informação de AKAIKE (AIC). No modelo obtido pelo método generalizado, também foi considerado o Critério de Informação Bayesiano (BIC).

O coeficiente de determinação (R^2) mede a proporção da variável Y que é explicada pelas variáveis $X_j (j = 1, 2, \ldots, k)$ e é simbolizado por R^2 (Ayres, 2012). Esse coeficiente quando associado à reta ajustada, representa a proporção da variabilidade amostral da variável dependente que é explicada pela equação de regressão. O seu valor é obtido através da soma dos quadrados dos desvios explicáveis pela regressão (SQR) dividido pela soma dos quadrados do desvio total (SQT), representada por:

$$R^2 = \frac{SQR}{SQT} = \sum_{i=1}^{n} \frac{(\hat{Y}_i - \bar{Y})^2}{(Y_i - \bar{Y})^2}, 0 \leq R^2 \leq 1$$

Quanto maior o valor de R^2, maior é o poder de explicação da regressão.

No caso do critério de informação, esta é uma técnica que surgiu com o intuito de comparar n modelos, $g_1(X_1 \mid \boldsymbol{\theta}_1), g_2(X_2 \mid \boldsymbol{\theta}_2), \ldots, g_n(X_n \mid \boldsymbol{\theta}_n)$, de acordo com as magnitudes da função maximizada, ou seja, $L(\hat{\boldsymbol{\theta}}_i)$. O AIC, específicamente, foi proposto em 1974 e é uma medida relativa da qualidade do ajuste dum modelo estatístico estimado pelo método da máxima verossimilhança, frequentemente utilizado para selecionar modelos em diversas áreas.

Este critério pretende escolher o modelo cuja densidade de probabilidade seja mais próxima da verdadeira. E baseia-se no fato de que o viés tende ao número de parâmetros a serem estimados no modelo, penalizando modelos com muitas variáveis. Ou seja, o AIC não é uma prova sobre modelo, no sentido de testar hipóteses, mas uma ferramenta para a seleção, sendo escolhidos aqueles que apresentarem menor valor. Esse critério pode ser definido como:
AIC = -2 (Função suporte maximizada) + 2 (número de parâmetros)

$$AIC = -2LL(\hat{\boldsymbol{\theta}}) + 2(k)$$

onde $LL(\hat{\boldsymbol{\theta}})$ é o máximo da função de log- verossimilhança do modelo e k é o número de parâmetros do modelo.

Já o Critério de Informação Bayesiano (BIC) é definido por:

$$BIC = -2LL(\hat{\boldsymbol{\theta}}) + k \log(n)$$

Com n o número de observações. Tanto o AIC quanto o BIC aumentam conforme a soma dos quadrados dos resíduos (SQE) aumenta. É importante ressaltar que ambos os critérios penalizam modelos com muitas variáveis, portanto, valores menores de AIC e BIC são preferíveis. Como estes métodos de avaliação de modelo bonificam estimadores da máxima verossimilhança e o R^2 bonifica estimadores por mínimos quadrados, foi essencial o senso crítico e a análise do contexto de produção para avaliação e seleção dos modelos que serão apresentados.

2.2.5 Técnicas para quando os pressupostos não se verificam

Quando a não normalidade nos resíduos é observada, duas premissas em modelos lineares podem não ser atendidas. O primeiro problema resultante da não normalidade é a obtenção de resultados inferenciais imprecisos sobre os valores-p e o intervalo de confiança dos parâmetros. E o segundo é que a relação entre os X_j e Y pode não ser linear (Agresti, 2015).

Para Verbek (2017) quando ocorre a falha do pressuposto da normalidade, mas os demais pressupostos se mantem, então os estimadores MMQ não são centrados, mas são consistentes e apresentam eficiencia assintótica. Assim, quando os pressupostos não são verificados no modelo obtido, ainda é possível utilizar algumas abordagens para solução da falha no pressuposto para obter um modelo preciso e que possa garantir inferência.

Segundo Pek (2018) os métodos mais frequentemente observados na literatura para contornar a ausencia de normalidade (e homocedasticidade) são as transformações, seguido do argumento de robustez de amostras grandes devido ao teorema do limite central (TLC), e por último, ainda são abordados os métodos não paramétricos. No entanto, para transformação, Agresti (2015) diz que é dificil encontrar uma transformação que forneça normalidade aproximada e variância constante.

Para os casos de não normalidade dos resíduos de um modelo de regressão linear múltipla também há a hipotese de utilizar os métodos da família Bootstrap proposto por Efron (1979), nos quais os dados não são alterados e a não normalidade dos resíduos é tratada como não informativa, mas apenas um incómodo a ser tratado. Este método baseia-se na suposição da amostra ser respresentativa da população através de reamostragem do conjunto de dados original. Usa-se frequentemente para aproximar o viés ou a variância de um conjunto de dados estatísticos, assim como para construir intervalos de confiança ou realizar contrastes de hipóteses sobre parâmetros de interesse.

Ainda Pek (2018) sugere que para casos de violação da homocedasticidade a abordagem mais utilizada é a matriz de covariância hetereocedástica corrigida (HCCM heteroscedastic corrected covariance matrix), que assim como Bootstrap não altera os dados e assume que a falha de especificação na estrutura de covariância dos erros é algo a ser tratado. Para exemplificar a hetereocedasticidade, pode-se considerar que com dados para N = 2, apresenta uma estrutura heterocedástica de forma $\sum \epsilon = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$, em que $\sigma_1^2 \neq \sigma_2^2$. E o objetivo é corrigir esta matriz de forma a que os resíduos fiquem homocedásticos ($\sigma_1^2 = \sigma_2^2$).

Outra técnica utilizada para casos de violação dos prossupostos é a regressão robusta, que assume que os dados observados estão contaminados com as observações influentes (outliers), e sugere extensivas formas de detecção e tratamento dessas observações extremas, promovendo, portanto, a alteração ou discarte dos dados extremos. Entretanto, muitas informações sobre a natureza dos fenómenos sob estudo e suas características podem estar associados a esses pontos, o que requer cautela para remoção ou modificação desses valores. Os críticos desta técnica alertam que esse processo de tratar os valores extremos, viola o pressuposto de independência requerida pelo método dos mínimos quadrados, o que pode comprometer a inferencia.

Quanto a falha na independência, Wooldridge (2010) aponta que esse é um problema comum, principalmente em aplicações económicas, pois nesses casos, muitas vezes não é possível coletar os dados da variável que verdadeiramente afetam o comportamento económico que se pretende modelar. Explica também que quando se utiliza uma medida imprecisa duma variável económica num modelo de regressão, ele irá conter um erro de medição que afetará os resultados. A falha na independência, também pode ter origem no processo de amostragem, que em muitos casos não é tão aletório assim. Felizmente, para esse tipo de problema Aupy et.al. (2017) diz que pode-se assumir com segurança (porém de maneira errada) que a falha de independência do modelo não afeta o modelo de forma geral, uma vez que conseguiu-se comprovar que o ganho obtido com o modelo em estudo independente foi insignificante, quando comparado a um modelo em que os resíduos apresentavam dependência.

Quando os regressores $X_j, j = 1, 2, \ldots, k$ são dependentes, diz-se que o modelo apresenta multicolinearidade. A multicolinearidade refere-se à correlação entre três ou mais variáveis independentes (Miloca & Conejo, 2012). Para falha na ausência de multicolinearidade, Verbek (2017) diz que torna-se difícil interpretar os coeficientes do modelo, e que nestes casos pode-se fazer uso de variáveis instrumentais (VI) no lugar da variável correlacionada. Entretanto, esse tipo de manobra exige cuidados com o método de estimação utilizado e com os problemas de quebras de outros pressupostos que podem vir a acontecer. Este mesmo autor também diz que, caso os pressupostos não se verifiquem, é indicado o uso de métodos de estimação alternativos como por exemplo, o método dos mínimos quadrados a 2 passos, ou o método da máxima verossimilhança (que utiliza algoritmo de maximização). Para Turkman & Silva (2000) quando o modelo linear clássico não é adequado para explicar as situações de estudo, um alternativa é a utilização de modelos lineares generalizados (GLM).

O fato é que lidar com as falhas das premissas do RLM não é tarefa fácil e neste trabalho as falhas foram tratadas de diferentes formas e serão apresentadas no capítulo 4 junto com os modelos obtidos.

2.3 Modelos lineares generalizados

Os modelos de regressão linear clássicos dominaram a modelação estatística até meados do século XX, embora vários modelos não lineares também tenham sido desenvolvidos para explicar as situações que não atingiam os pressupostos do modelo clássico (Turkman & Silva, 2000). Os modelos lineares generalizados (GLM - Generalized linear models) introduzidos por Nelder e Wedderbun (1972) correspondem a uma síntese dos modelos que seguem uma distribuição exponencial (Turkman & Silva, 2000). E apresentam 3 componentes: Componente aleatória (componente do erro), componente sistemática (preditor linear - η) e função de ligação ($g(\cdot)$) (Agresti, 2015).

Agresti (2015) aponta que a componente aleatória do GLM consiste numa variável resposta Y com observações independentes $(y_1, \ldots, y_n)$ e distribuição com densidade ou massa de probabilidade de:

$$f(y_i; \theta, \phi) = exp\Big[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\Big]$$

θ_i é chamado de parâmetro natural ou de localização,

ϕ é chamado de parâmetro de dispersão,

$a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções reais conhecidas.

Os pressupostos do GLM diz que Y_i apresenta distribuição dentro da família exponencial e funções de ligação da média igualada ao preditor linear. Além de haver relação linear entre a variável resposta transformada pela função de ligação e a variável explicativa. E por fim os erros devem ser independentes

(mas não têm que ter uma distribuição normal).

Este tipo de modelo utiliza o método da máxima verossimilhança para estimar os parâmetros e tem a vantagem de não ter que manobrar (ou transformar) a variável resposta para ter a distribuição normal. Além disso a escolha da função de ligação é independente da componente aletória (o que oferece flexibilidade à modelação) e produz efeitos aditivos, e portanto, não é necessária a homogeneidade.

Para Turkman & Silva (2000) os modelos lineares generalizados são uma extensão do modelo linear clássico, o valor esperado $\mu_i = E[Y_i \mid \boldsymbol{X_i}], X_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ está relacionado com o preditor linear $\eta_i = Z_i^T \beta$ através da relação:

$$\mu_i = h(\eta_i) = h(Z_i^T \beta), \eta_i = g(\mu_i)$$

onde,

h é uma função monótona e diferenciável.

g é a função de ligação $(g = h^{-1})$.

$\boldsymbol{\beta}$ é um vetor de parâmetros de dimensão p $(K+1)(\beta = \beta_0, \beta_1, \ldots, \beta_k)$

Z é um vetor de especificação de dimensão p, em função do vetor de covariáveis $X = (X_1, X_2, \ldots, X_k)$.

A função de ligação depende do tipo de resposta e do estudo que se pretende fazer. Turkman & Silva (2020) dizem ainda que tem especial interesse a situação em que o preditor linear coincide com o parâmetro canónico $(\theta_i = \eta_i)$. Uma vantagem de usar função de ligação canónica , é que neste caso, desde que o parâmetro de escala seja conhecido, o vector parâmetro desconhecido da estrutura linear admite uma estatística suficiente mínima de dimensão fixa.

Para Agresti (2015) assim como os outros estimadores de modelos lineares, o GLM apresenta parâmetros consistentes e robustos à medida que a amostra aumenta. E além disso, mesmo que haja uma má especificação da distribuição de Y o GLM assume que a família exponencial apresenta certa propriedade de robustez. Portanto, se a função de ligação e o predictor linear estiverem corretamente especificado, então o parâmetro ainda é consistente. A função de ligação especifica a ligação entre a componente aletória e a sistemática, e também especifica como o valor esperado de Y, $E(Y) = \mu$, se relaciona com o preditor linear.

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

A inferência sobre a qualidade do ajustamentos dos parâmetros dum modelo linear generalizado apresentam 3 testes que são padrões: O teste da razão de verossimilhança, o teste de Wald e o teste score, para a seleção das covariáveis que formaram o melhor modelo. Para avaliar a bondade de ajustamento do modelo levou-se em conta o valor obtido pelo pseudo coeficiente de determinação (R^2), os valores de AIC e BIC.

Como mencionado anteriormente na regressão linear clássica, o R^2 descreve o quanto as variáveis que compõe o modelo, explicam a variável resposta. Para qualquer GLM, a correlação entre os valores ajustados $(\hat{\mu}_i)$ e a resposta observada (y_i) mede o poder preditivo. E também pode ser usado para comparar o ajustamento de diversos modelos de uma mesma base de dados (Agresti, 2015). Entretanto, os valores do pseudocoeficiente de determinação são usualmente baixos, quando comparado com os modelos de regressão linear clássica, o que pode causar certa confusão na interpretação do melhor modelo.

No caso do GLM a análise do pressuposto da linearidade das covariáveis e a análise dos resíduos

são formas de diagnóticos para avaliação do modelo, no caso em estudo, para a linearidade foi efetudo o método dos quartis, método de lowess e dos polinómios fracionários. Para os resíduos, fez-se a análise dos resíduos deviance que é um tipo de medida para observação da falta de ajustamento do modelo (Faraway, 2006). Foi feita também observação gráfica dos pontos influentes pela distância de Cook (generalizada) e pelo DfBeta.

Esta parte do trabalho teve como objetivo comparar um modelo da regressão clássica e sua ligação, com outras famílias da distribuição exponencial como por exemplo, a gama com função de ligação logarítmica, que segundo Turkman & Silva (2000) é usado na análise de dados contínuos com suporte positivo para a distribuição da variável resposta (como é o caso do custo por dia de produção). Outra distribuição de probabilidade testada foi a gaussiana inversa que é indicado para dados assimétricos e com valores positivos (Turkman & Silva, 2000).

2.4 Modelo de regressão logística

O modelo de regressão logística é um tipo de modelo linear generalizado que utiliza a função de ligação logit para a resposta que segue uma distribuição binomial (binária), é um modelo muito popular, provavelmente devido à simplicidade da sua implementação computacional (Turkman & Silva, 2000). Este tipo de modelo estima a probabilidade de uma característica estar presente (π) dado os valores da variável explicativa, de modo que $Y_i = 1$ apresenta probabilidade (π_i) de ocorrência do evento desejado em estudo. Dado o conjunto das k variáveis explicativas $X_1, X_2, \ldots, X_k$, e $\boldsymbol{X}$ o vetor por elas composto. Assim a função de ligação obtida é:

$$g(X) = logit(P(Y = 1 \mid X_1 = x_1, \ldots, X_k = x_k)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

E o modelo é dado por:

$$\pi(X) = \frac{exp(g(X))}{1 + exp(g(X))}$$

Para Turman & Silva (2000) o modelo de regressão logístico selecionado no procedimento adotado para a seleção de covariáveis sem interações tem a seguinte forma estrutural:

$$ln[\frac{\pi_i}{(1-\pi_i)}] = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j$$

Assim como a regressão linear múltipla este tipo de regressão também apresenta pressupostos que devem ser cumpridos:

- A variável resposta deve ser independente e seguir uma distribuição binomial (n_i, π_i),
- Os resíduos devem ser independentes mas não têm distribuição normal,
- As variâncias são heterogéneas.

Assim como descrito anteriormente para este método de modelação também se utiliza o teste de Wald para seleção dos parâmetros. Os pressupostos de linearidade também seguiram os mesmo procedimentos. O teste de bondade utilizado seguiu os passos da metodologia de Hosmer & Lemeshow (2013).

E também, foi realizado a análise da capacidade discriminativa do modelo que levou em consideração a curva característica de operação do receptor (curva ROC) que é usada para representar a relação entre a especificidade e a sensibilidade para os diferentes valores do ponto de corte.

O ponto de corte da curva ROC é o valor de probabilidade a partir do qual se considera que o modelo prediz a ocorrência do evento. A sensibilidade é a probabilidade da predição correta do evento dado que o evento ocorreu. E a especificidade é a predição correta do evento dado que o acontecimento não ocorreu. O que prentende-se rastrear com essas medidas é a ocorrência de falso positivo (predizer uma ocorrência incorretamente) ou falso negativo (predizer uma não ocorrência incorretamente). No caso em estudo, o interesse é obter a predição correta para a classificação do animal como abate DOP, dado que ao fim do ciclo produtivo ele realmente seja vendido como DOP.

3

Caracterização da base de dados

A base de dados inicial com que trabalhamos foi cedida pela ACBM à qual se juntou informações das características individuais dos animais provenientes da base de dados da GENPRO. A base de dados final incluía, entre outras, as seguintes variáveis: peso do animal à entrada e saída dos currais, idade do animal à entrada e saída dos currais, custo total de produção, custo com transporte, custo de funcionamento do curral, custo com alimentação, custo com profilaxia e outros custos de produção, além do peso da carcaça do animal, rendimento de carcaça, tipo de abate para qual o animal foi encaminhado. E também continha as informações genéticas como por exemplo: o valor genético (VG) da capacidade maternal, VG do intervalor entre partos, VG capacidade de crescimento, entre outros.

Pode-se analisar as características observadas, relativas aos dados em contexto do custo de produção no CTR. Os animais considerados neste estudo apresentavam no mínimo 6 meses à entrada e no máximo 15 meses à saída. Pode-se verificar que dos 716 animais, 54% foram para o abate que garante o selo DOP e 46% foram para o abate convencional (figura 3.1). Como já mencionado anteriormente os animais destinados ao abate convencional saem para o abate entre os 8 e os 12 meses de idade, enquanto que para o abate DOP os animais saem entre 10 e 15 meses de idade, permanecendo por mais tempo na engorda.

Os custos de produção foram divididos em custos fixos e custos variáveis de acordo com o tempo de permanência do animal na engorda. Assim, definiu-se como custos fixos de produção: Profilaxia (custo

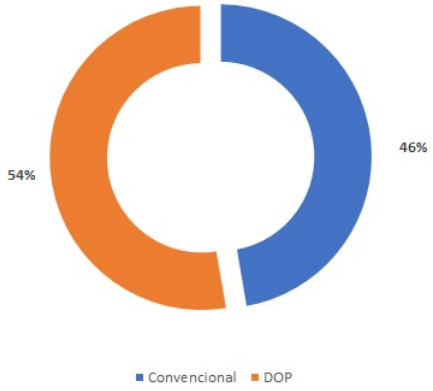


Figura 3.1: Percentagem de bovinos mertolengas da ACBM por destino de abate.

com veterinário), taxa Promert (Agrupamento de produtores de Bovinos Mertolengos S.A.) e o custo de transporte; E como os custos variáveis da produção: custo com alimentação (somou-se o custo com concentrado e o custo de palha), custo de funcionamento e os outros custos associados ao tempo de permanência do animal no CTR. No caso do custo de transporte, embora seja um valor calculado com base no peso da carcaça do animal, foi considerado um valor médio com base nos dados disponíveis e com este valor, esta variável compos o custo fixo.

A variável custo por dia na engorda foi construída a partir dos custos variáveis totais divididos pelo tempo que o animal permaneceu no curral. Pode-se verificar através do teste de Shapiro-Wilk (valor-p = 0,09) e confirmado pelo teste de Kolmogorov Smirnov com correção de Lilliefors (valor-p = 0,67) que a variável custo por dia de produção atende ao pressuposto da normalidade.

Pode-se verificar que o custo por dia de produção médio foi de 2,18 euros, com desvio padrão de 22 cêntimos de euro. Nota-se que para os animais DOP, observou-se custo médiol por dia de 2,24 euros com desvio padrão de 21 cêntimos de euro. Já para os animais que foram para abate convencional o custo médio por dia foi de 2,11 euros por dia com desvio padrão de 21 cêntimos de euros por dia. Como seria de se esperar, os animais que foram destinados ao abate DOP apresentaram maior custo por dia de produção e pelo teste t (valor-p $<$ 0,001) pode-se verificar que existe diferença estatística significativa nos valores médios do custo por dia de produção entre os grupos de abate.

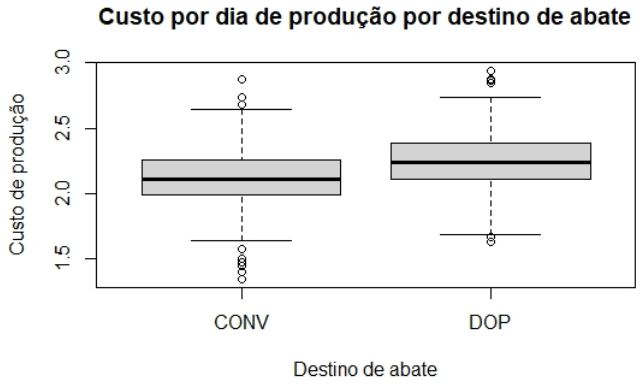


Figura 3.2: Boxplot do custo por dia de produção por destino de abate

É possível verificar na figura 3.2 a existência de alguns outliers, principalmente no grupo de animais que foram para abate convencional. Verificou-se pela análise do coeficiente do ponto ponto bisserial (-0,29) que as características custo por dia de produção e destino de abate apresentam associação fraca e inversa, ou seja, a medida que aumenta o custo por dia de produção, aumenta a tendência para que o animal seja encaminhado para o abate DOP. O que faz sentido uma vez que estes animais permanecem por mais tempo no CTR, e quanto mais velhos ficam, maior é o consumo alimentar.

Em relação ao rendimento da carcaça, é possível perceber que os animais que foram ao abate DOP apresentaram rendimento da carcaça de 52%, enquanto que os animais que foram ao abate convencional apresentaram rendimento da carcaça de 50%. Pelo teste t (valor-p $= 0{,}58$) pode-se verificar que não existe diferença significativa nas médias do rendimento de carcaça entre os grupos de abate.

Nota-se que o mais comum nos bovinos que foram enviados ao destino de abate que garante o selo DOP foi um registo de peso da carcaça acima de 180 kg (32,4%). Relativamente aos bovinos de abate convencional, observou-se mais frequentemente que as carcaças pesam entre 140 kg e menos que 160 ks (17,6%) (figura 3.3).

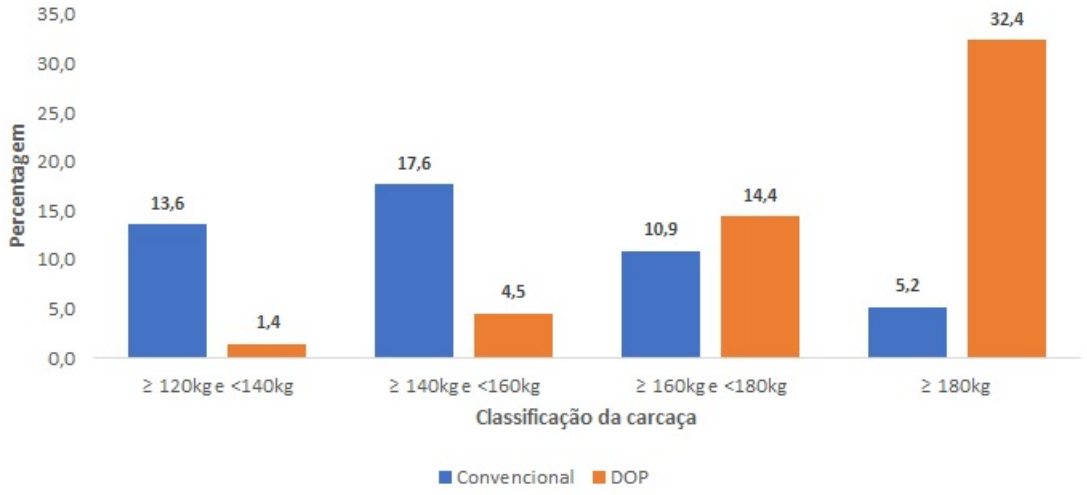


Figura 3.3: Percentagem de bovinos Mertolenga da ACBM por destino de abate e classificação do peso da carcaça

Nesta análise, também foi possível verificar que os animais que acabaram em vitelão mertolengo DOP entraram no CTR com idade superior (12,06%) e peso inferior (-4,90%). As variáveis idade e peso à saída e os dias na engorda apresentaram diferenças que segundo Pais et al. (2019), decorre sobretudo da idade limite superior dos dois tipos de produtos (15 meses no caso vitelão DOP e 12 meses no vitelão convencional). A duração da recria e acabamento é uma das variáveis que mais diferencia o vitelão DOP (em média 168 dias) e convencional (média de 108 dias), com 55,8% a favor do primeiro conforme pode ser verificado na tabela 3.1.

O ganho médio diário durante a produção é superior no vitelão DOP (1,14 contra 1,12 no convencional). Para o valor do preço por kg da carcaça, entre os dois grupos destaca-se a maior valorização do vitelão DOP (5,80%), com valor médio de 4,01 euros por kg de carcaça, enquanto que o animal encaminhado para abate convencional apresentou valor médio de 3,79 euros por kg de carcaça. Como consequência, o valor da venda da carcaça, também aponta a maior valorização do vitelão DOP (30,19%), com valor médio de 759,70 euros, enquanto que o convencional apresenta valor médio de venda de carcaça de 583,54 euros. Segundo Pais et al. (2019) o maior peso ao abate e melhor preço por kg de carcaça justificam o maior

valor médio de venda da carcaça do vitelão DOP.

Tabela 3.1: Dados técnico-económicos relativos aos animais que iniciaram a fase de recria nos anos de 2013 a 2017 no Centro de Testagem e Recria da ACBM.

Dados técnicos económicos	Vitelão DOP		
	Valor médio	Desvio padrão	Coef. variação
Idade entrada (meses)	8,37	1,46	17,45%
Peso entrada (kg)	170,56	39,37	23,08%
Idade saída (meses)	13,88	1,07	7,71%
Peso saída (kg)	360,19	42,35	11,76%
Número de dias na engorda	167,74	44,06	26,27%
Ganho médio diário (GMD)	1,14	0,20	17,54%
Custo total (euros)	435,67	105,92	24,31%
Rendimento de carcaça	0,52	0,02	3,85%
Preço por kg de carcaça (euros/kg)	4,01	0,15	3,74%
Valor da carcaça (euros)	759,70	119,90	15,78%
Dados técnicos económicos	Vitelão CONV		
	Valor médio	Desvio padrão	Coef. variação
Idade entrada (meses)	7,46	1,16	15,55%
Peso entrada (kg)	179,35	39,40	21,97%
Idade saída (meses)	11,00	1,15	10,45%
Peso saída (kg)	301,90	40,94	13,56%
Número de dias na engorda	107,62	42,95	39,91%
Ganho médio diário (GMD)	1,12	0,23	20,54%
Custo total (euros)	271,03	103,24	38,09%
Rendimento de carcaça	0,50	0,02	4,00%
Preço por kg de carcaça (euros/kg)	3,79	0,06	1,58%
Valor da carcaça (euros)	583,54	85,63	14,67%
Dados técnicos e económicos geral	Manada		
	Valor médio	Desvio padrão	Var. DOP / CONV *
Idade entrada (meses)	7,94	1,4	12,17%
Peso entrada (kg)	174,73	39,60	-4,90%
Idade saída (meses)	12,51	1,82	26,18%
Peso saída (kg)	332,51	50,83	19,31%
Número de dias na engorda	139,19	52,87	55,86%
Ganho médio diário (GMD)	1,13	0,21	1,79%
Custo total (euros)	357,49	133,06	60,75%
Preço por kg de carcaça (euros/kg)	3,91	0,16	5,80%
Valor da carcaça (euros)	676,05	136,99	30,19%

* Variação do valor do vitelão DOP relativamente ao vitelão convencional

Os produtores que estão interessados em melhorar uma determinada característica produtiva do seu rebanho, deverão identificar os animais que genéticamente são superiores e selecioná-los para que sejam os antecessores das futuras crias. Ao selecionar os animais genéticamente superiores é possível aumentar o termo médio da produção das gerações seguintes. E assim, ao longo das gerações, o valor produtivo dos animais irá aumentando e com ele os benefícios económicos associados (Cardoso, 2009).

Dentro da produção animal algumas características genéticas são muito visadas para o melhoramento e são denominados Valores Genéticos (VG). No caso dos animais em estudo, os valores genéticos

observados foram: peso aos 210 dias de idade (P210), capacidade maternal, capacidade de crescimento, intervalo entre partos, carcaça por dia de idade, ganho médio diário (GMD) em teste de estação, consumo alimentar residual, índice de conversão alimentar, longevidade produtiva e os resultados podem ser observados na tabela 3.2.

Tabela 3.2: Análise descritiva dos valores genéticos dos animais quanto ao seu grupo de destino de abate

Destino de abate (Convencional)				
Variáveis	Mínimo	Máximo	Média	Desvio padrão
P210	130,99	227,13	174,22	25,15
VG capacidade maternal	-18,30	24,80	-0,50	6,40
VG capacidade de crescimento	-37,70	23,00	-3,80	11,10
VG GMD estação	-20,70	59,80	8,90	12,50
VG carcaça dia idade	-44,00	72,60	9,80	20,60
VG intervalo entre partos	-39,30	31,20	-7,80	14,00
VG índice conversão	-0,50	0,50	0,00	0,20
VG longevidade produtiva	-19,90	4,30	-7,80	3,70
VG consumo alimento residual	-174,30	306,70	22,60	84,10
Destino de abate (DOP)				
Variáveis	Mínimo	Máximo	Média	Desvio padrão
P210	129,00	227,00	165,21	25,08
VG capacidade maternal	-17,80	23,20	-0,30	7,10
VG capacidade de crescimento	-37,90	28,60	-5,80	10,80
VG GMD estação	-21,70	62,30	7,80	13,60
VG carcaça dia idade	-38,80	94,10	4,60	23,80
VG intervalo entre partos	-37,00	33,90	-3,60	14,70
VG índice conversão	-0,60	0,60	0,00	0,20
VG longevidade produtiva	-19,20	6,90	-7,60	3,70
VG consumo alimento residual	-151,10	308,70	21,90	71,90

O valor genético para a capacidade maternal deverá ser o maior possível (mais positivo). Pretende-se que os reprodutores transmitam aos descendentes capacidade para desmamarem animais mais pesados (Carolino, 2016). A média geral para o valor genético da capacidade maternal foi de -0,37 unidade de medida (u.m.), a manada apresentou também desvio padrão de 6,75 (u.m). e o valor máximo dessa característica foi de 24,77 (u.m.). Para os animais DOP a média foi de -0,25 (u.m.), sendo portanto, um valor superior a média geral, esses animais apresentaram desvio padrão de 7,08 (u.m.). Para os animais que foram ao mercado convencional, verificou-se média de -0,50 (u.m.), e desvio padrão de 6,38 (u.m.).

Os valores genéticos para a capacidade de crescimento, o ganho médio diário durante o teste de performances, o peso de carcaça por dia de idade e peso aos 210 dias de idade (P210) também são tanto melhores, quanto maiores forem esses valores. Pretende-se que os reprodutores transmitam aos descendentes uma boa capacidade de crescimento até e após o desmame (descendentes mais pesados) (Carolino, 2016).

A média global para o valor genético da capacidade de crescimento foi de -4,86 (u.m.), a manada apresentou também desvio padrão de 10,97 (u.m.) e o valor máximo dessa característica foi de 28,57 (u.m.). Para os animais DOP a média foi de -5,75 (u.m.), esses animais apresentaram desvio padrão de 10,81 (u.m.). Para os animais que foram ao mercado convencional, verificou-se média de -3,89 (u.m.), a manada deste grupo apresentou desvio padrão de 11,09 (u.m.). Foi possível perceber (valor-p teste t $<$ 0,0001) que há evidência estatística para que se rejeite a igualdade das médias para o valor genético da capacidade de crescimento entre os grupos de abate. Portanto, os animais que foram para o abate convencional,

apresentaram média do valor genético da capacidade de crescimento superior quando comparado com o grupo de animais que foram ao abate DOP. Com esse resultado podemos talvez sugerir que, o peso e a capacidade que o animal tem de crescer, não sejam características que diferencie o animal DOP do convencional, sendo que os animais mais jovens mais pesados já vão direto para o abate, por já estarem com a preparação para tal.

O valor genético do GMD em teste de performance apresentou média geral de 8,29 (u.m.), a manada apresentou também desvio padrão de 13,07 (u.m.) e o valor máximo dessa característica foi de 62,25 (u.m.). Para os animais DOP a média foi de 7,79 (u.m.), esses animais apresentaram desvio padrão de 13,60 (u.m.). Para os animais que foram ao mercado convencional, verificou-se média de 8,84 (u.m.), a manada deste grupo apresentou desvio padrão de 12,47 (u.m.).

Para o valor genético do peso da carcaça por dia de idade a média foi de 7,20 (u.m.), a manada apresentou também desvio padrão de 22,58 (u.m.) e o valor máximo dessa característica foi de 94,09 (u.m.). Para os animais DOP a média foi de 4,59 (u.m.), esses animais apresentaram desvio padrão de 23,83 (u.m.). Para os animais que foram ao mercado convencional, verificou-se média de 10,08 (u.m.), e desvio padrão de 20,77(u.m.). Através da análise dos VG's que caracterizam o crescimento e ganho de peso do animal, foi possível notar que os animais que foram ao abate convencional apresentaram resultados melhores quando comparados com os animais que foram ao abate DOP, levando a uma possível indicação de que o peso e a capacidade de engorda do animal, não são necessariamente, características importantes para a classificação do animal como DOP.

Para o P210 o valor médio da manada foi de 170,65 kg e desvio padrão de 25,50 kg, o valor máximo dessa característica foi de 227 kg. Para os animais que foram ao abate DOP, a média de P210 foi de 165,21 kg com desvio padrão de 25,08 kg. Para os animais que foram ao mercado convencional a média de P210 foi de 174,22 kg e desvio padrão de 25,15 kg. Pelo valor-p do teste $t < 0{,}001$, foi possível perceber que existe diferença significativa entre os pesos médios aos 210 dias por grupo de abate. E que os animais que foram ao abate convencional, estavam mais pesados aos 210 dias de idade, quando comparado com os animais DOP.

O valor genético para o intervalo entre partos é tanto melhor, quanto menor for (mais negativo). Pretende-se que os reprodutores transmitam aos descendentes características genéticas que, no caso de serem fêmeas, lhes proporcionem intervalos entre partos mais reduzidos (ACBM, 2018). A média geral para o intervalo entre partos foi de -5,54 (u.m.), a manada apresentou também desvio padrão de 14,53 (u.m.) e o valor máximo dessa característica foi de 33,88 (u.m.). Para os animais DOP a média foi de -3,56 (u.m.), esses animais apresentaram desvio padrão de 14,73 (u.m.). Para os animais que foram ao mercado convencional, verificou-se média de -7,73 (u.m.), a manada deste grupo apresentou desvio padrão de 14,01 (u.m.).

O valor genético para o índice de conversão alimentar durante o teste de performance deverá ser o menor possível (mais negativo). Pretende-se que os reprodutores transmitam aos descendentes capacidade para consumirem menos alimento por cada quilograma de aumento de peso (ACBM, 2018). A média geral para o VG do índice de conversão foi de -0,03 (u.m.), a manada apresentou também desvio padrão de 0,16 (u.m.) e o valor máximo dessa característica foi de 0,56 (u.m.). Para os animais DOP a média foi de -0,03 (u.m.) e desvio padrão de 0,15 (u.m.). Para os animais que foram ao mercado convencional, verificou-se média de -0,04 (u.m.) e desvio padrão de 0,17 (u.m.). No apêndice A são apresentados alguns comandos do R e resultados adicionais referentes às análises apresentadas neste capítulo.

4

Modelos de regressão linear múltipla

Este capítulo é composto por três secções nas quais são apresentadas modelos de regressão linear múltipla para a variável custo por dia de produção, a primeira apresenta o modelo geral de regressão linear múltipla para para a manada que tentou explicar a variável custo por dia de acordo com as variáveis zootécnicas disponíveis. Também são apresentados o modelo de custo por dia de produção para os animais que foram ao abate DOP, e por fim, um modelo de custo de produção para os animais que foram ao abate convencional.

4.1 Modelo geral para o custo por dia de produção

Nesta fase procurou-se verificar as variáveis que compõe os custos por dia de produção para os dois tipos de abate: DOP e convencional. iniciou-se com a análise dos modelos simples para verificar quais variáveis, individualmente, apresentavam maior relação com o custo por dia de produção. Para chegar ao modelo de regressão múltipla, adicionou-se todas as variáveis de interesse da base de dados, e foram retiradas as que não apresentavam significância pela análise do valor-p do teste t, uma a uma (método backward) (tabela 4.1).

Tabela 4.1: Componentes do modelo inicial para o custo por dia de produção com estimativa dos coeficientes, erro padrão, estatística do teste t e o valor-p.

Variáveis	Coeficientes	Erro Padrão	Estatística	Valor-p
Intercepto	1,5063	0,1754	8,586	<0,001
Peso à entrada	0,0027	0,0009	2,930	0,003
Idade à entrada	0,0263	0,0223	1,182	0,238
P210	0,0003	0,0010	0,362	0,7177
VG GMD em teste estação	0,0002	0,0011	0,240	0,8107
VG longevidade produtiva	0,0015	0,0026	0,584	0,5598
VG consumo alimentar residual	-0,0004	0,0002	-1,629	0,1041
VG índice de conversão	0,2807	0,1404	1,999	0,0463
VG capacidade de crescimento	0,0013	0,0008	1,614	0,1073
VG capacidade maternal	0,0078	0,0016	4,833	<0,001
VG carcaça por dia de idade	0,0023	0,0004	4,770	<0,001
VG intervalo entre partos	0,0041	0,0006	6,254	<0,001

*Valor-p do teste t (Pr(>|t|) para significância dos coeficientes.
R^2= 0,42; AIC= -266,50

Por fim, o modelo obtido incluiu as variáveis: peso à entrada, VG da capacidade crescimento, VG da capacidade maternal, VG da carcaça por dia de idade e o VG do intervalo entre partos, com coeficiente de determinação de 45% e AIC de -518,83 (tabela 4.2).

$$CVdia.prod = 1,6145 + 0,0032PE + 0,0015CC + 0,0048CM + 0,0016CD + 0,0014IE$$

A variável destino de abate, usada para indicar se o animal foi a abate DOP ou não, no fim da produção, apresentou-se significativa, entretanto, como não é algo que se possa prever no início do processo produtivo, não foi adicionado ao modelo.

Tabela 4.2: Componentes do modelo geral para o custo de produção por dia com os valores dos coeficientes, indicação do erro padrão, estatística de teste t e valor-p

Modelo geral	Coeficientes	Erro Padrão	Estatística	Valor-p
Intercepto	1,6145	0,0304	53,103	<0,001
Peso à entrada (PE)	0,0032	0,0002	19,465	<0,001
VG capacidade de crescimento (CC)	0,0015	0,0006	2,558	0,011
VG capacidade maternal (CM)	0,0048	0,0009	4,940	<0,001
VG carcaça por dia de idade (CD)	0,0016	0,0003	5,359	<0,001
VG intervalo entre partos (IE)	0,0014	0,0004	3,190	0,001

R^2= 0,45; AIC= -518,83

Nenhuma interação se mostrou significativa neste modelo, entretanto, os resíduos também não atenderam aos pressupostos para a normalidade (nem simetria e curtose, valor- p Kolmogorov-Smirnov (K-S) $<$ 0,001, valor- p teste dAgostino $=$ 0,0002 e valor- p Anscombe $<$ 0,001) sendo perceptível também a ausencia da normalidade pela análise gráfica (Figura B.1). Para tentar contornar esse problema, sem assumir o TLC para a falha de normalidade, foram verificados os valores influentes e percebeu-se que os seguintes animais se portavam como outliers: 2, 3, 263, 264, 265, 266, 267, 268, 333, 337, 338, 348, 360, 363, 710. Destes animais, 4 foram ao abate DOP (348, 360, 363, 710) e 11 foram ao abate convencional

(2, 3, 263, 264, 265, 266, 267, 268, 333, 337, 338).

Os animais 2 e 3 ficaram apenas 29 dias na engorda, o que alterou diretamente o custo por dia na engorda. Em contra partida, os animais 333, 337, 338, e 710 ficaram mais de 200 dias na engorda, o que obviamente também afetou o custo. Os animais 2, 3, 263, 264, 265, 266, 267, 268, 348, 360 apresentaram custo por dia de produção menor ou igual a 1,69 euros, valor que está abaixo do limite inferior da média do custo por dia de produção do intervalo de 95% de confiança (Tabela 4.3).

Tabela 4.3: Média, limite inferior (LI) e superior (LS) do intervalo de 95% de confiança da média, e desvio padrão das variáveis dias na engorda, custo total de produção e custo por dia de produção.

Variáveis	Média	LI (95%)	LS(95%)	Std. Dev.
Dias na engorda	139,56	135,71	143,42	52,48
Custo total de produção	358,25	348,51	367,98	132,47
Custo de produção por dia	2,18	2,16	2,20	0,22

Foram removidas estas observações e ao testar os pressupostos novamente, notou-se que esse modelo, atende à curtose (Valor-p Anscombe = 0,17), à homoscedasticidade (Valor-p Breusch Pagan = 0,06) e à ausência de multicolinearidade (Valores da estatística do teste de fator de inflação da variância (vif) < 2), com pouca alteração no erro quadrático médio e aumento de 8% no coeficiente de determinação ($R^2 = 53\%$). A melhoria do comportamento dos resíduos pode ser verificado na figura B.4 e os parâmetros do modelo geral final estão disponíveis na tabela 4.4.

Retirar os outliers, conforme indicado por Pek (2018) e Agresti (2015) é uma manobra complexa, pois pode-se perder informações. Mas no caso da produção animal é um mal necessário, uma vez que alguns indivíduos apresentam características estremas por diversos fatores naturais: desde ao fato do animal vir a nascer dum parto precoce, a problemas de saúde que se agravam dentro do processo do próprio parto, ou ainda fatores genéticos até então desconhecidos. Portanto, esses indivíduos existem e é preciso ter um olhar diferenciado para os mesmos.

Tabela 4.4: Modelo geral final sem os outliers do custo de produção por dia na engorda com os valores das estimativas dos coeficientes, indicação do erro padrão, estatística do teste t e valor-p.

Modelo geral variáveis à entrada	Coeficientes	Erro padrão	Estatística	Valor-p
Intercepto	1,599	0,0269	59,408	<0,001
Peso à entrada	0,003	0,0001	22,682	<0,001
VG capacidade de crescimento	0,001	0,0005	2,973	0,003
VG capacidade maternal	0,003	0,0008	4,520	<0,001
VG carcaça dia idade	0,001	0,0003	6,179	<0,001
VG intervalo entre partos	0,001	0,0004	4,558	<0,001

R^2= 0,53; AIC= -713, 23

Através da tabela 4.4 é possível perceber que o peso à entrada e o VG capacidade maternal são os coeficientes com maior valor, e portanto, mantendo tudo constante, o aumento de 30 kg do peso à entrada do animal, aumenta o custo por dia de produção em 9 cêntimos de euros, a mesma análise pode ser feita para a capacidade maternal. No caso do peso, esse incremento no custo pode ser explicado pela adaptação da dieta no início da entrada dos currais (Arrigoni, M., 2017).

Para verificar as diferenças dos custos por dia de produção entre os grupos, construiu-se um modelo com as variáveis disponíveis na entrada dos currais para cada um. No apêndice B são apresentados alguns comandos do R e resultados adicionais referentes às análises apresentadas.

4.2 Modelo custo por dia de produção dos animais do abate DOP

A construção do modelo para os animais que foram destinados ao abate DOP seguiu a mesma metodologia do modelo geral, porém dividiu-se a base de dados e utilizou-se apenas os animais que foram ao abate DOP para construir este modelo. A ideia nesta fase foi de perceber se há variáveis significativas diferentes, considerando a variável resposta custo por dia na engorda para esse grupo específico. Pode-se perceber que as variáveis que exlicam o custo por dia dos animais DOP são: peso à entrada, idade à entrada,VG da capacidade maternal e VG da capacidade de crescimento (tabela 4.5). Nenhuma interação mostrou-se significativa, entretanto este modelo não passou no pressuposto da normalidade. Ao verificar os pontos influentes a 1% de nível de significância, foram identificados os seguintes indivíduos: 4, 295, 374, 375, 377.

Essas observações foram removidas e os pressupostos foram testados novamente, assim como feito no modelo geral. E pode-se perceber o aumento do R^2 (de 60% para 69%) e ao analisar os resíduos (figura B.6) pode-se verificar que eles seguem a normalidade, embora de forma marginal (valor-p K-S $=$ 0,02). Apresentam homoscedasticidade (valor-p Breusch Pagan $=$ 0,01) e também, não apresenta multicolinearidade (Valores da estatística do teste de fator de inflação da variância (vif) $<$ 2), o modelo obtido pode ser verificado na tabela 4.5. A dispersão gráfica dos resíduos do modelo obtido pode ser verificado na figura B.6, no qual pode-se destacar um bom comportamento linear do gráfico normal quantil-quantil.

Tabela 4.5: Modelo do custo por dia de produção dos animais DOP sem outliers, com os valores das estimativas dos coeficientes, indicação do erro parão, estatística do teste t e valor-p.

Modelo DOP variáveis à entrada	Coeficientes	Erro padrão	Estatística	Valor-p
Intercepto	1,895	0,0401	47,276	<0,001
Peso à entrada	0,004	0,0002	27,537	<0,001
Idade à entrada	-0,047	0,0044	-10,701	<0,001
VG capacidade maternal	0,002	0,0008	2,746	0,006
VG da capacidade de crescimento	0,002	0,0006	4,514	<0,001

R^2= 0,69; AIC= -538,93

É possível perceber que a idade tem relação inversa com o custo, e portanto, mantendo tudo constante, para o aumento da idade à entrada em um mês reduz em 4 cêntimos de euros o custo por dia de produção. Este resultado pode ser explicado pelo fato dos animais irem a abate DOP mais velhos (até 15 meses) quando comparado ao abate convencional (até 12 meses), e também, levando em consideração o tempo médio de permanência do animal na engorda, conforme pode ser verificado na análise descritiva efetuada no capítulo 3.

4.3 Modelo custo por dia de produção dos animais do abate convencional

A construção do modelo de regressão linear múltipla seguiu a mesma técnica utilizada para a construção dos modelos dos animais DOP e o modelo geral. Entretanto, para este caso, de acordo com a interpretação dos resultados da análise dos resíduos, optou-se pela transformação da variável resposta original para o seu logaritmo. Verificou-se a significância das seguintes variáveis a compor o modelo dos animais que foram ao abate convencional: Idade à entrada, peso aos 210 dias de idade (P210), VG capacidade maternal e VG da carcaça por dia de idade (tabela 4.6). O modelo atende à simetria (p-valor DAgostino $=$ 0,11), à homoscedasticidade (p-valor Breusch - Pagan $=$ 0,13) e à ausência de multicolinearidade (valores da estatística do teste de fator de inflação da variância (vif) $<$ 2) com $R^2 = 57\%$.

Na análise gráfica (figura B.7) é possível perceber que, mesmo com a transformação da variável resposta, ocorre falha da normalidade. Entretanto, essa manobra foi escolhida, para não se retirar os valores influentes, uma vez que a simples transformação melhorou o comportamento dos resíduos. Sendo, portanto, uma manobra diferente do que foi procedido com os dois modelos anteriores.

Tabela 4.6: Modelo do custo por dia de produção dos animais do abate convencional com os valores da estimativa dos coeficientes do modelo, da exponencial dos coeficientes, indicação do valor-p do teste t, erro padrão.

Modelo convencional variáveis à entrada	Coeficientes	Exp(coef)	Erro padrão	Estatística	Valor-p
Intercepto	0,002	1,002	0,0463	0,048	0,96
Idade à entrada	0,045	1,046	0,0047	9,477	<0,001
Peso aos 210 dias de idade	0,002	1,002	13,251	0,0002	<0,001
VG capacidade maternal	0,002	1,002	0,0007	2,881	0,004
VG carcaça dia de idade	0,001	1,000	0,0002	4,330	<0,001

R^2= 0,57; AIC= -592,39

Mantendo tudo constante, o aumento de 1 mês na idade a entrada aumenta o custo por dia de produção em 4,6%. Para Pais et al. (2019) a idade é a característica que limita na decisão do animal ir ao abate convencional, portanto, é de se esperar que o modelo para esses animais tenham as variáveis que apresentam essa carcaterística: idade e VG da carcaça por dia de idade. Já para os animais que vão a DOP o peso na entrada do processo de engorda é a característica mais importante.

Pode-se verificar que o VG da capacidade maternal apareceu em todos os modelos e ao verificar o relatório do teste de estação publicado em 2019 (ACBM, 2019), para a escolha dos animais como reprodutores, é dada atenção especial aos valores genéticos da VG capacidade maternal e VG do intervalo entre partos. Portanto é possível perceber o impacto do manejo genético desse grupo de animais, numa informação económica.

Por fim, é interessante verificar que mesmo se fosse utilizado o modelo da secção 4.1 com a variável destino de abate incluída, o conjunto de variáveis explicativas seria diferente do encontrado nos modelos em que apenas se usou o subconjunto dos animais de acordo com o seu grupo de abate.

5

Custo por dia de produção com modelos lineares generalizados

Este capítulo apresentará 4 secções e tem como objetivo verificar se as variáveis que compõe o modelo de regressão linear se adequam melhor a outras funções de distribuição da família exponencial, bem como se ocorre melhoria com outras funções de ligação. A linearidade das covariáveis que compõe o modelo foi verificada pelo método dos quartis, método de Lowess e pelos polinómios fracionários, com gráficos expostos no apêndice C. Todas as covariáveis atenderam a este pressuposto com os diferentes tipos de distribuição que foram testadas.

Para Agresti (2015) o modelo linear clássico corresponde ao GLM com distribuição de probabilidade gaussiana e função de ligação identidade. A primeira secção apresenta o modelo obtido com essas característica confirmando que coincide com o RLM obtido. Na segunda secção é apresentado o modelo GLM com distribuição de probabilidade gaussiana inversa e função de ligação inversa. Na terceira secção é apresentado os GLM com a distribuição de probabilidade gama e diferentes funções de ligação. E por fim, a última secção faz a comparação entre todos os modelos para a variável custo por dia de produção.

5.1 GLM com distribuição Gaussiana e função de ligação identidade

O modelo GLM obtido com a função de distribuição de probabilidade gaussiana e função de ligação identidade corresponde ao modelo geral com os valores influentes obtido pela técnica RLM. Os valores estimados para os coeficientes de regressão do modelo e os respectivos erros padrões encontram-se na tabela 5.1 abaixo. O coeficiente de determinação indica que este modelo explica 45% da variação total do custo por dia de produção, o que corresponde ao valor do coeficiente de determinação obtido pela técnica de regressão clássica, antes da remoção dos pontos influentes (tabela 4.2).

Tabela 5.1: Modelo GLM do custo por dia de produção com a f.d.p. gaussiana e função de ligação identidade.

Distribuição	Normal			
Link	Identidade	$Z = \eta + (y - \mu)$		
Modelo	Estimativa	Erro Padrão	Estatística	Valor-p
Intercepto	1,614	0,0304	53,103	<0,001
Peso à entrada	0,003	0,0001	19,465	<0,001
VG capacidade maternal	0,004	0,0009	4,940	<0,001
VG capacidade de crescimento	0,001	0,0005	2,558	0,011
VG carcaça dia de idade	0,001	0,0003	5,359	<0,001
VG intervalo entre partos	0,001	0,0004	3,190	0,001

R^2=0,45; AIC= -518,83; BIC = -486,83; Residual deviance = 19,82

Na análise de resíduos, pode-se verificar que os 15 indivíduos que foram retirados do modelo RLM por serem outliers, também foram influentes no GLM. Além desses animais, também foram acrescidos na lista de outliers para o modelo GLM os animais: 273, 274, 325, 331, 336, 351, 361, 362, 368, 403, 404, 446, 447, 459, 489, 550, 558, 612, 621, 626, 635, 703 e 711. Entretanto, pode-se verificar que emboram existam muitos outliers nenhum ultrapassa o valor indicado na distância de Cook (figura C.7) e por isso, optou-se por não retirar os outliers e comparar com o modelo geral que foi apresentado na tabela 4.2.

Com relação a adequabilidade da função de distribuição de probabilidade e da função de ligação, é possível perceber que os dados apresentam um padrão linear (figura 5.1).

5.2 GLM com distribuição Gaussiana Inversa e função de ligação inversa

A distribuição Gaussiana inversa é indicada para dados assimétricos positivos com presença de muitos outliers (Paula, 2013). Os valores estimados para os coeficientes de regressão do modelo e os respectivos erros padrões encontram-se na tabela 5.2. O pseudo coeficiente de determinação indica que este modelo explica 43,28% da variação total do custo por dia de produção, apresentando um valor inferior ao modelo com distribuição gaussiana e função de ligação identidade.

Com relação a adequabilidade da função de distribuição de probabilidade e da função de ligação, é possível perceber que os dados apresentam um padrão linear, porém, mais dispersa (Figura 5.2). E pelos resultados do AIC, BIC e R^2 verificou-se a superioridade do modelo clássico.

5.3 GLM com Distribuição Gama

Para Turkman & Silva (2000) a distribuição de probabilidade Gama é indicada para dados que assumem valores positivos, assimétricos e é uma distribuição muito utilizada para modelação de dados

Figura 5.1: Adequação da família de distribuição Gaussiana e da função de ligação identidade ao modelo geral para o custo por dia de produção.

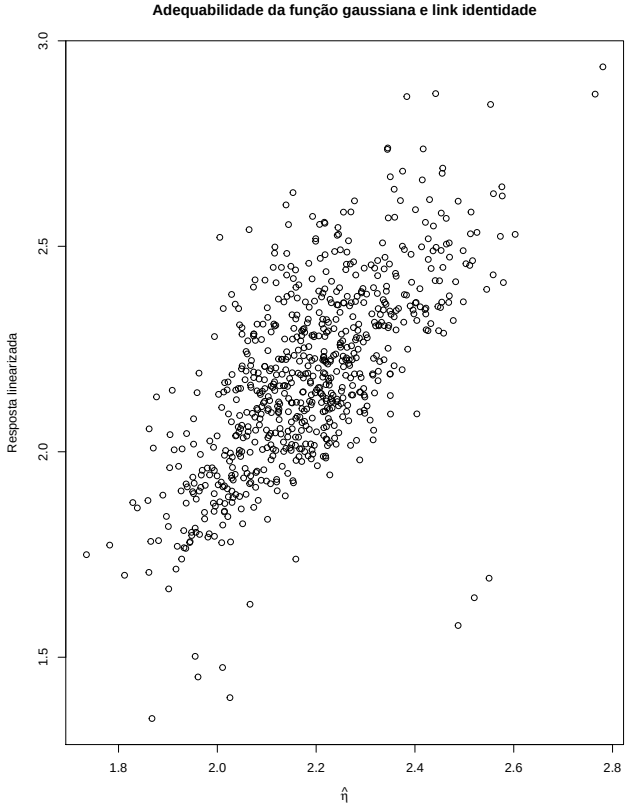


Tabela 5.2: Modelo GLM do custo por dia de produção com a f.d.p. gaussiana inversa e função de ligação inversa.

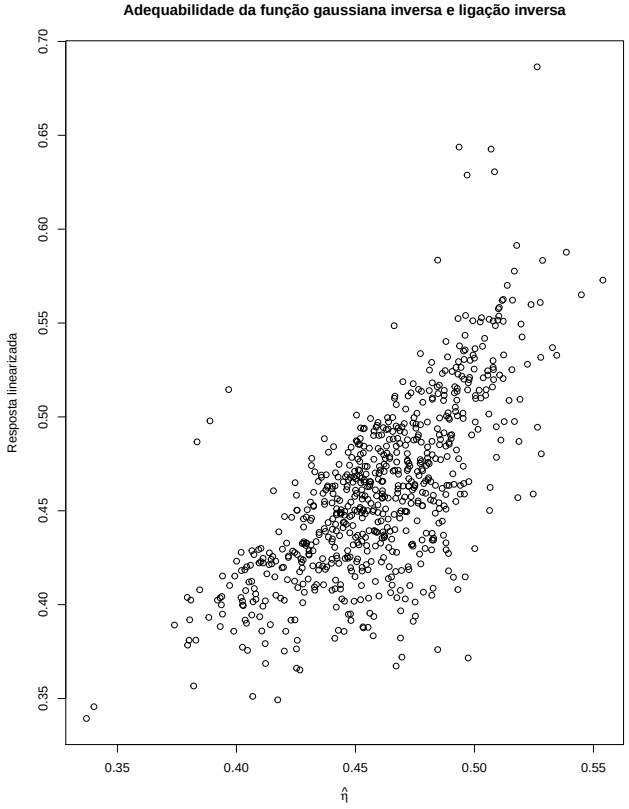
Distribuição	Normal inversa			
Link	Inversa $Z = \eta - \frac{2(y-\mu)}{(\mu^3)}$			
Modelo	Estimativa	Erro Padrão	Estatística	Valor-p
Intercepto	0,5790	0,0064	89,817	<0,001
Peso à entrada	0,0006	0,0000	-19,520	<0,001
VG capacidade maternal	0,0011	0,0002	-5,092	<0,001
VG capacidade de crescimento	0,0003	0,0001	-2,672	0,007
VG carcaça dia de idade	0,0003	0,0001	-4,828	<0,001
VG intervalo entre partos	0,0003	0,0001	-3,249	0,001

R^2=0,43; AIC= -471,44; BIC = -439,44; Residual deviance = 2,0653

econométricos. Na distribuição gama a variância é proporcional ao quadrado da média e esta propriedade sugere que este tipo de modelo pode ser útil em situações onde a variância dos dados não é constante, mas proporcional ao quadrado da média. No caso em estudo a variância se manteve constante, mas dada as caracteríticas da distribuição gama decidiu-se testar para a base de dados em estudo, com 3 funções de ligação diferentes: a identidade, a logaritmica e a inversa.

A função de ligação identidade foi a primeira a ser testada por ser a mais simples e ser a função de ligação utilizada no modelo linear (Agresti, 2015). Os valores estimados para os coeficientes de regressão do modelo e os respectivos erros padrões encontram-se na tabela 5.3. O pseudo coeficiente de determinação

Figura 5.2: Adequação da família de distribuição Gaussiana inversa e a função de ligação inversa ao modelo geral para o custo por dia de produção.



indica que este modelo explica 45% da variação total do custo por dia de produção, sendo um valor muito próximo do obtido pelo modelo de regressão múltipla, entretanto, os valores de AIC e BIC foram menores do que quando comparados com a distribuição gaussiana.

Tabela 5.3: Modelo linear generalizado do custo por dia de produção com a f.d.p. gama e função de ligação identidade

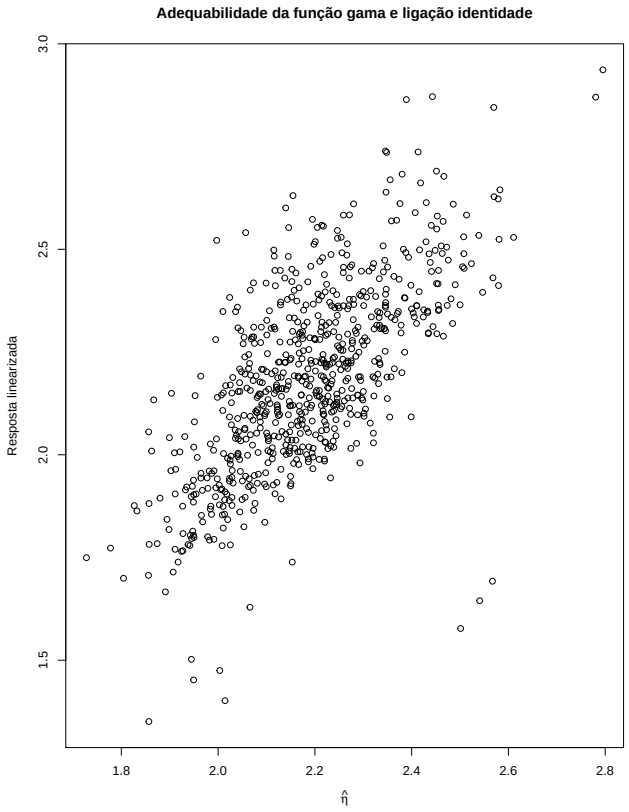
Distribuição	Gama			
Link	Identidade	$Z = \eta + (y - \mu)$		
Modelo	Estimativa	Erro Padrão	Estatística	Valor-p
Intercepto	1,592	0,0301	52,836	<0,0001
Peso à entrada	0,003	0,0001	19,990	<0,0001
VG capacidade maternal	0,004	0,0009	5,097	<0,0001
VG capacidade de crescimento	0,001	0,0006	2,250	0,0247
VG carcaça dia de idade	0,001	0,0003	4,817	<0,0001
VG intervalo entre partos	0,001	0,0004	3,616	0,0003

R^2=0,45; AIC= -506,78; BIC = -474,78; Residual deviance = 4,26

Com relação a adequabilidade da função de distribuição de probabilidade e da função de ligação, é possível perceber que os dados apresentam um padrão linear (Figura 5.3).

A ligação inversa, que é a ligação canónica da distribuição gama, também foi testada e os valores estimados para os coeficientes de regressão do modelo gama com ligação inversa são apresentados na tabela 5.4 e a adequabilidade da função de distribuição e de ligação está na figura 5.4. Pode-se notar que embora

Figura 5.3: Adequação da família de distribuição Gama e a função de ligação identidade ao modelo geral para o custo por dia de produção.



a função de ligação de ligação inversa se adeque a uma reta, a ligação identidade produziu melhores valores de AIC, BIC e R^2 ($R^2 = 43\%$).

Tabela 5.4: Modelo linear generalizado do custo por dia de produção com distribuição gama e função de ligação inversa

Distribuição	Gama			
Link	Inversa	$Z = \eta - \frac{(y-\mu)}{(\mu^2)}$		
Modelo	Estimativa	Erro Padrão	Estatística	p-valor
Intercepto	0,5763	0,0064	90,054	<0,001
Peso à entrada	0,0006	0,0000	-19,333	<0,001
VG capacidade maternal	0,0010	0,0002	-5,033	<0,001
VG capacidade de crescimento	0,0003	0,0001	-2,815	0,005
VG carcaça dia de idade	0,0003	0,0001	-5,081	<0,001
VG intervalo entre partos	0,0002	0,0001	-3,021	0,002

R^2=0,43; AIC= -488,11; BIC = -456,11; Residual deviance = 4,37

Por fim, foi testado a distribuição gama com a função de ligação logarítimica, os valores estimados para os coeficientes de regressão do modelo e os respectivos erros padrões encontram-se na tabela abaixo. A adequabilidade da função de distribuição de probabilidade e desta função de ligação, também apresentaram um padrão linear, porém, como no caso anterior, o modelo clássico produziu melhores valores de AIC, BIC e R^2 ($R^2 = 44\%$).

Figura 5.4: Adequação da família de distribuição Gama e a função de ligação inversa ao modelo geral para o custo por dia de produção.

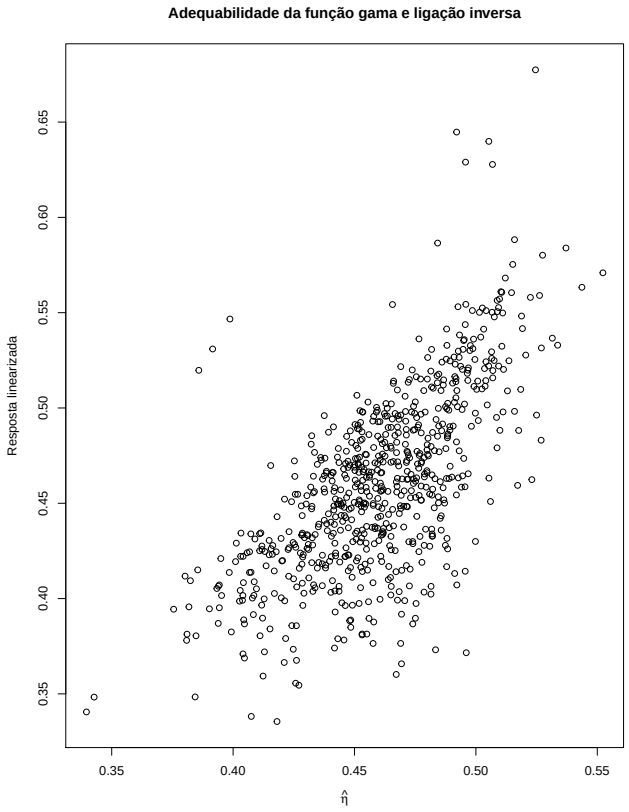


Tabela 5.5: Modelo linear generalizado do custo por dia de produção com distribuição gama e função de ligação log.

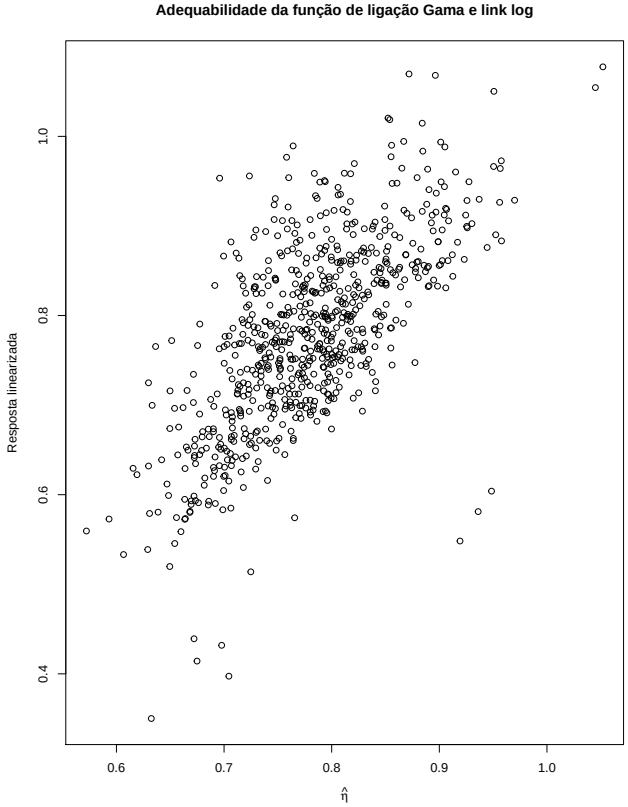
Distribuição	Gama			
Link	$\log$	$Z = \eta + \frac{(y-\mu)}{\mu}$		
Modelo	Estimativa	Erro Padrão	Estatística	Valor-p
Intercepto	0,5147	0,01398	36,817	<0,001
Peso a entrada	0,0015	0,0001	19,658	<0,001
VG capacidade maternal	0,0023	0,0004	5,071	<0,001
VG capacidade de crescimento	0,0007	0,0002	2,550	0,011
VG carcaça dia de idade	0,0007	0,0001	4,976	<0,001
VG intervalo entre partos	0,0007	0,0002	3,340	<0,001

R^2=0,44; AIC= -498,49; BIC = -466,49; Residual deviance = 4,31

5.4 Comparação entre os modelos

Como pôde ser verificado nas análises gráficas de adequabilidade da distribuição de probabilidade e função de ligação do modelo, a ligação identidade apresentou melhor ajustamento à uma reta, do que as outras funções de ligação. Além disso, na tabela 5.6 é possível perceber que o modelo com distribuição gaussiana e função de ligação identidade foi o que apresentou melhores resultados para AIC, BIC e R^2, seguido da distribuição gama. Portanto, os dados se adequam bem a esse tipo de distribuição e ligação. O ajuste realizado no modelo clássico geral retirando os 15 outliers, além de aumentar o coeficiente de

Figura 5.5: Adequação da família de distribuição Gama e a função de ligação logaritmica ao modelo geral para o custo por dia de produção.



determinação, também melhorou a distribuição dos resíduos, e apresentou melhores resultado para inferência (tabela 4.4).

Tabela 5.6: Resumo dos resultados do tipo de distribuição de probabilidade, função de ligação, AIC, BIC Deviance, R^2 e se o modelo se adequa aos dados.

Dist. de probabilidade	Func. ligação	AIC	BIC	Deviance	R^2	Adequa
Gaussiana	Identidade	-518,83	-486,83	19,82	45,47	sim
Gaussiana inversa	Inversa	-471,44	-439,44	2,06	43,28	sim
Gama	Identidade	-506,78	-474,78	4,26	45,20	sim
Gama	Inversa	-488,11	-456,11	4,37	43,75	sim
Gama	Log	-498,49	-466,49	4,31	44,56	sim

6

Modelação do destino de abate do animal

A regressão logística (binária) é um tipo de modelo linear generalizado que apresenta como variável resposta uma variável binária. A regressão logística estima a probabilidade de uma característica estar presente, dado os valores das variáveis explicativas (Hosmer & Lemeshow, 2013).

Este capítulo consiste numa secção de apresentação do modelo que tem como objetivo encontrar os fatores que explicam a ida do animal ao tipo de abate que garante o selo DOP através das variáveis de caracterização zootécnica e dos valores genéticos disponíveis na base de dados.

A primeira etapa começa com a seleção das covariáveis que irão compor o modelo múltiplo. Utilizou-se a metodologia backward e, como nos casos anteriores, o grau de importância duma covariável foi medida pelo seu valor-p no teste de Wald (Turkman & Silva, 2000), utilizou-se a metodologia de Hosmer & Lemeshow (2013).

6.1 Modelo logístico obtido

Considerando o sucesso como os animais que foram destinados ao DOP (categoria 1 = 376 animais) e o insucesso os animais que foram destinados ao mercado convencional (categoria 0 = 338), utilizando a razão de chances (odd ratio) em que:

$$OR = \frac{Sucesso}{Insucesso}$$

tem-se que a razão de chances é de 1,11. Este valor é igual ao obtido pela exponencial de 0,106 (valor obtido para o modelo nulo). O modelo logístico para as chances do animal ir ao abate DOP encontra-se na tabela 6.1.

Pode-se verificar que as duas variáveis genéticas que atuam na escolha do reprodutor (Carolino, 2016), apareceram na equação de classificação para o animal DOP. E também é possível notar que o peso contribui de forma negativa para esta classificação, portanto, quanto maior o peso à entrada (em kg) as possibilidades do animal ser classificado para o abate DOP diminuem em 1%, talvez pelo fato do animal já estar muito próximo do preparo necessário para ser vendido ao destino de abate convencional. O animal entra pesado, e se não possuir as características genéticas necessárias para a reprodução, é encaminhado prontamente ao abate convencional. Em contrapartida, para cada mês a mais na idade do animal na entrada dos currais, as possibilidades do animal ser classificado como DOP aumentam em duas vezes.

$$DestinoDOP = \frac{1}{\exp(-2,631 - 0,016PE + 0,718IA + 0,036CM + 0,022IE)}$$

Tabela 6.1: Modelo logístico para classificação do animal DOP com os valores dos coeficientes do modelo, exponencial dos coeficientes (OR), indicação da estatística z do teste de Wald e valor-p.

Modelo logístico	Coeficientes	OR	Estatística Z	$Pr(\mid z \mid)$
Intercepto	-2,631	0,07	-4,919	<0,001
Peso à entrada (PE)	-0,016	0,98	-6,550	<0,001
Idade à entrada (IA)	0,718	2,05	9,724	<0,001
VG capacidade maternal (CM)	0,036	1,04	2,729	0,006
VG intervalo entre partos (IE)	0,022	1,02	3,743	<0,001

No estudo dos resíduos do modelo pode-se verificar que 8 indivíduos se apresentaram como possíveis valores influentes (indivíduos: 273, 319, 335, 337, 344, 348, 489 e 711), entretanto, nenhum contribui com alterações significativas nos coeficientes. E pela observação na distância de Cooks apenas 5 desses indivíduos se destacaram, mas desses 5, nenhum ultrapassou o limite de 0,5 indicado, e portanto, os indivíduos não foram retirados do modelo.

Os resíduos do modelo obtido atende aos pressupostos e não apresentam problemas de multicolinearidade, uma vez que no teste VIF a estatística obtida para as variáveis apresentavam valores menores que 2. Através da verificação do valor-p no teste de Hosmer-Lemeshow (valor-p = 0,06), há evidência estatística para sugerir que o modelo indicado apresenta bondade de ajustamento ao nível de significância de 5%. Assim, as probabilidades preditas não se desviam das probabilidades observadas. É possível perceber que o modelo apresenta um coeficiente de determinação (R^2) de 24%, mas com uma acurácia aceitável de 70% e área abaixo da curva ROC também aceitável (75%) essa informação pode ser verificada na tabela 6.2 junto com os valores do critério de informação de AKAIKE (AIC), a área do valor em baixo da curva ROC e valor-p para o teste de Hosmer e Lemeshow.

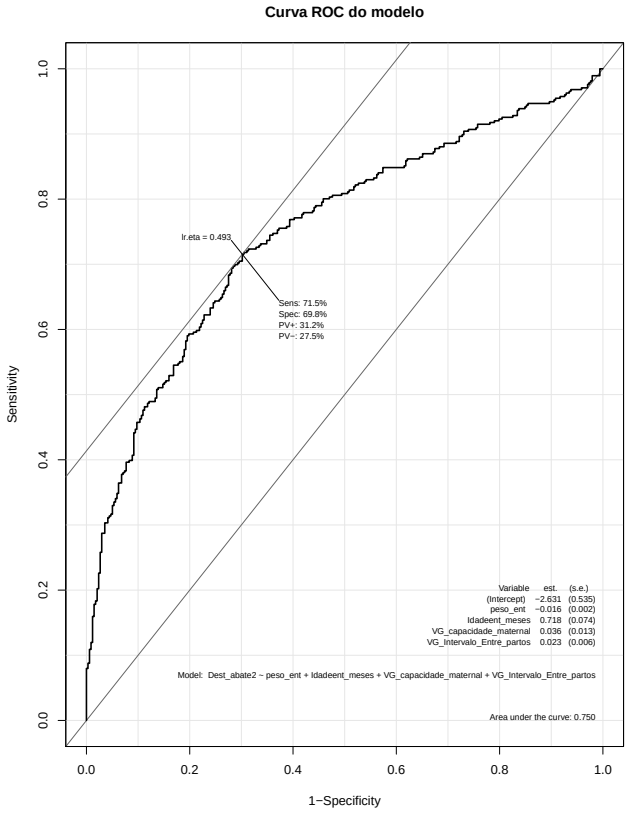
Como pode ser verificado, o modelo sugerido também apresenta uma capacidade discriminativa

Tabela 6.2: Teste de ajustamento, medidas explicativas e capacidade discriminativa do modelo logístico

Ajustamento	Teste de Hosmer	Valor-p = 0,06
	Teste de Cessie	Valor-p = 0,62
Medidas explicativas	R^2	0,2431
	Racio de verossimilhança	143,64
	AUC	0,75
	Sensibilidade	71,50%
	Especificidade	69,80%
	Percentagem de acerto	70,50%
	SQE observada	144,61
	SQE esperada	144,84

aceitável com 75% de área abaixo da curva ROC (Figura 6.1). E apresenta sensibilidade de 71,5% e especificidade de 69,8%. Portanto, o modelo sugerido é válido para objetivo de classificação.

Figura 6.1: Curva ROC do modelo obtido.



Ao testar o modelo considerando um animal com as características médias para peso à entrada, idade à entrada e os valores genéticos que compõem o modelo, fez-se a verificação da probabilidade do indivíduo ir ao abate que garante o selo DOP.

Levando em consideração um animal com as características médias desta base de dados, a probabilidade desse indivíduo ser classificado como DOP é de 40,98%. Entretanto, um indivíduo 2 meses mais velho (com idade à entrada de 9 meses), apresenta probabilidade de 74,48% de ser classificado como DOP (tabela 6.3).

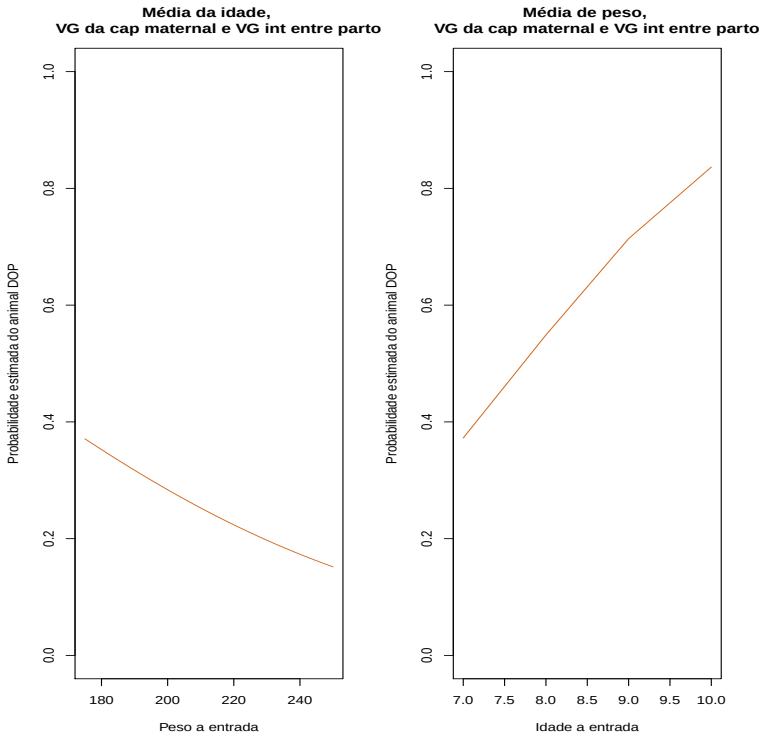
Ao considerar que o tempo médio na engorda é de aproximadamente 3 ou 4 meses, os animais com idade a entrada de 9 meses, ao fim do processo nos currais, apresentarão de 12 a 13 meses de idade, o que corresponde a idade exigida para o abate DOP. Já os animais com 7 meses de idade, finalizaram a engorda em torno dos 10 meses de idade, sendo, portanto, animais jovens para o abate DOP. E prolongar a permanência desses animais na engorda, iria despender de mais recursos, provocando o aumento do custo de produção do indivíduo.

Ao considerar que o animal irá entrar mais pesado nos currais (peso a entrada = 200 kg), pode-se verificar que a probabilidade desse indivíduo ser classificado como DOP é de 32,30%. O que pode significar que para o criador, entregar um animal mais pesado para a engorda, não necessariamente implica que esse animal será vendido como animal de selo DOP. Portanto, a sugestão que se pode fazer com base nos resultados obtidos é que, o produtor talvez terá maior possibilidade de obter animais com selo DOP, se investir na idade do animal na entrada dos currais para a engorda (figura 6.2). Entretanto, isso corresponde a um aumento do custo do bezerro.

Tabela 6.3: Probabilidades do animal ser classificado como DOP com base no modelo logístico obtido (coeficientes) e apresentando: características médias, animal mais velho e animal mais pesado.

Variáveis do modelo	coeficientes	Valores médios	Média alt. idade	Média alt. peso
Intercepto	-2,631	-	-	-
Peso à entrada	-0,015	175,00	175,00	200,00
Idade à entrada	0,718	7,00	9,00	7,00
VG capac maternal	0,036	-0,38	-0,38	-0,38
VG int entre partos	0,022	-5,55	-5,55	-5,55
PROBABILIDADES	-	0,4098	0,7448	0,323

Figura 6.2: Comportamento da probabilidade do animal ser classificado como DOP com o peso a alterar e com a idade a alterar.



7

Conclusão

Os produtos DOP são os que têm ligações mais fortes com o local em que são fabricados (Comissão Europeia, 2018) e para Gama et al. (2004) as raças autóctones apresentam uma capacidade única de tirar partido das condições ambientais muitas vezes adversas e restritivas, em que as raças exóticas não conseguem produzir, ou mesmo sobreviver. A FAO (2014) sugere que investir nas raças adaptadas pode ser uma forma de buscar a produção animal sustentável. A raça mertolenga é uma raça autóctone e sua localização está maioritariamente na região do Alentejo (ACBM, 2018).

Durante o processo de desenvolvimento deste trabalho, ao perceber a diferença entre os dois produtos, DOP e convencional, e ao constatar a diferença no tempo de permanência entre os grupos de abate no CTR, uma vez que, os animais destinados ao abate convencional saem para o abate entre os 8 e os 12 meses de idade, enquanto que para o abate DOP os animais saem entre 10 e 15 meses de idade, permanecendo por mais tempo na engorda. Identificou-se a necessidade de caracterizar os animais que vão para o abate DOP e os animais que vão para o abate convencional, e a análise estatística e modelação foi também realizada por tipo de abate. A base de dados continha a informação de 716 animais machos, dos quais 54% foram para o abate que garante o selo DOP e os restantes foram para o abate convencional. Dispúnhamos de dados referentes à estrutura de custos de produção dos animais desde a entrada no CTR até o abate (custos com alimentação, profilaxia, entre outros) e a características individuais de cada animal (peso à entrada, idade à entrada, os valores genéticos, etc.).

Pretendendo-se modelar o custo diário de produção recorreu-se a modelos de regressão linear múltipla, modelos de regressão linear generalizados e, em particular, modelos de regressão logística. Para cada uma destas técnicas de modelação foi obtido o melhor modelo e validados os seus pressupostos. Relativamente a regressão linear múltipla, o melhor modelo geral obtido para o custo por dia de produção foi composto pelas variáveis peso à entrada, VG da capacidade de crescimento, VG da capacidade maternal, VG da carcaça por dia de idade e VG do intervalo entre partos. Quanto ao custo por dia de produção dos animais que vão para o abate DOP, o melhor modelo mostrou como significativas as variáveis peso à entrada, idade à entrada, VG da capacidade maternal e VG da capacidade de crescimento. Para o custo de produção dos animais que vão para o abate convencional, as variáveis que se mostraram significativas para explicar o custo por dia de produção foram idade à entrada, P210, VG da capacidade maternal e VG da carcaça por dia de idade. Por fim, também foi interessante verificar que mesmo se fosse utilizado o modelo geral acrescido da variável destino de abate, o conjunto de variáveis explicativas seria diferente do encontrado nos modelos dos grupos, em que apenas se usou o subconjunto dos animais de acordo com seu grupo de abate.

Aplicação dos modelos lineares generalizados vieram confirmar que o modelo de regressão linear múltipla era mais adequado para explicar o custo por dia de produção. Uma vez que foram considerados os modelos de distribuição gaussiana, gaussiana inversa e gama, com funções de ligação identidade, inversa, logarítmica, constatou-se que o modelo que melhor se adequava aos dados correspondia a distribuição gaussiana e função de ligação identidade. Que coincide com o modelo de regressão linear múltipla.

No que diz respeito a regressão logística, sendo que é um tipo de modelo linear generalizado que apresenta como variável resposta uma variável binária, no nosso caso o objetivo era encontrar os fatores que explicavam a ida do animal ao tipo de abate que garante o selo DOP. Concluiu-se que a idade à entrada, o peso à entrada, o VG da capacidade maternal e o VG do intervalo entre partos são os fatores que influenciam a classificação do animal como DOP. Pode-se perceber que a idade é uma variável que tem grande impacto nas chances do animal ir para o abate DOP, mas que em contra partida, aumenta os custos de produção para o criador, uma vez que o animal fica por mais tempo no CTR. Em relação aos valores genéticos que foram identificados neste modelo, de acordo com o catálogo de reprodutores da ACBM (Carolino, 2016) estes vêm sendo trabalhados nos reprodutores mertolengos, por serem considerados dos mais relevantes da raça.

Com este trabalho, através do estudo detalhado do custo por dia de produção, pretende-se contribuir com informação que possa ser útil para os criadores e à associação para tomada de decisão, oferecendo maior conhecimento da dinâmica dos custos de produção.

No futuro será interessante prosseguir com o estudo do lucro da produção de vitelão mertolengo. Embora o retorno verificado pelo peso da carcaça e pelo valor por quilo da carcaça seja maior, dado ao custo por dia de produção elevado, não se sabe até que ponto é mais lucrativo à produção de vitelão com destino de abate DOP quando comparado ao vitelão de abate convencional.

Bibliografia

[1] ACBM (2018). Associação de Criadores de Bovinos Mertolengos. Consultado em 10 nov. 2018. Disponível emhttp://www.mertolenga.com

[2] Afonso, A. & Nunes, C. (2019). Estatística e Probabilidades: Aplicações e Soluções em SPSS. Versão revista e aumentada. Universidade de Évora.

[3] Agresti, A. (2015). Foundations of linear and Generalized linear models. 1. Ed. John Wiley & Sons, inc.

[4] Aupy, G.; Robert, Y. & Vivien, F. (2017). "Assuming failure independence: Are we right to be wrong?". IEEE International Conference on Cluster Computing (CLUSTER), pp. 709-716.

[5] Araújo, M. J. (2008). Fundamentos de Agronegócios. 2. Ed. 3. Reimpr. São Paulo: Atlas.

[6] Arrigoni, M. (2017). Adaptação ao confinamento evita perda de até 7% do peso do animal. Consultado em 01 jan. 2020. Disponível em https://www.girodoboi.com.br/capa/adaptacao-ao-confinamento-evita-perda-de-ate-7-do-peso-do-animal/

[7] Ayres, M. (2012). Elementos de Bioestatística. 2a edição, Sociedade Civil Mamirauá, Belém, Brasil.

[8] Berenson, M.; Levine, D. M.; Szabat, K. A.; Watson, J.; Jayne, N. & OBrien, M. (2012). Basic Business Statistics: Concepts and applications. United States: Person Education.

[9] Braumann, C.A. (2005). Introdução às equações diferenciais Estocásticas e Aplicações. Ericeira: SPE.

[10] Braumann, C. A. (2019). Introduction to Stochastic Differential Equations with Applications to Modelling in Biology and Finance. Hoboken: Wiley.

[11] Bussab, W. O. & Morettin, P. A. (2012). Estatística Básica. 7ł edição, São Paulo: Editora Saraiva.

[12] Cardoso, F. F. (2009). Ferramentas e estatégias para o melhoramento genético de bovinos de corte. Embapa Pecuária Sul: Bagé, Brasil. Consultado em 10 out. 2019. Disponível em https://www.infoteca.cnptia.embrapa.br/bitstream/doc/657470/1/DT83.pdf

[13] Carolino, N.; Pais, J.; Henriques, N. & Rodrigues, S. (2016). Catálogo de touros 2016: Avaliação genética da raça bovina mertolenga. Consultado em 01 set. 2020. Disponível em http://www.mertolenga.com/Catalogo_2016.pdf

[14] Crepaldi, S. A. (2005). Contabilidade Rural: uma abordagem decisorial. 3. ed. revista, atualizada e ampliada São Paulo: Atlas

[15] Comissão Européia (2018). Os regimes de qualidade explicados. Consultado em 08 ago. 2020. Disponível em https://ec.europa.eu/info/food-farming-fisheries/food-safety-and-quality/certification/quality-labels/quality-schemes-explained_pt

[16] Costa, T. M. A (2015). Explorações de bovinos de carne em modo extensivo e semi-intensivo no Alentejo: Uma análise técnico-económica. Dissertação de mestrado, Faculdade de Medicina Veterinária, Universidade de Lisboa, Lisboa, Portugal. Consultado em 15 nov. 2018. Disponível em http://hdl.handle.net/10400.5/9146

[17] Direção Geral de Agricultura e Desenvolvimento Rural: Produtos tradicionais portugueses Carne Mertolenga DOP. Consultado em 08 ago. 2020. Disponível em https://tradicional.dgadr.gov.pt/pt/cat/carne/carne-de-bovino/233-carne-mertolenga-dop

[18] Efron, B. (1979) Bootstrap methods: another look at the jackknife. Ann. Statist. 7, 1-26.

[19] Emiliano, P.C. (2009). Fundamentos e aplicações dos critérios de informação: AKAIKE e Bayesiano. Consultado em 28 mar. 2020. Disponível em http://repositorio.ufla.br/jspui/handle/1/3636

[20] Faraway, J. J. (2006). Extending the Linear Model with R. Chapman & Hall.

[21] FAO (2014). Livestock and Animal Production. Consultado em 15 nov. 2018. Disponível em http://www.fao.org/ag/againfo/themes/en/animal_production.html

[22] FAO (2018). Península Ibérica e Itália. Consultado em 15 nov. 2018. Disponível em http://www.fao.org/docrep/015/an473s/an473s04.pdf

[23] FAO (2019). Animal production and health: Meat & meat products. Disponível em 10 out. 2019. Disponível em http://www.fao.org/ag/againfo/themes/en/meat/home.html

[24] Filipe, A. P.; Braumann, C.A. & Roquete, C.J. (2007). Modelos de crescimento de animais em ambiente aleatório. Actas do XIV Congresso Anual da SPE. Sociedade Portuguesa de Estatística. Covilhã (Portugal): 401-440

[25] Gama, L. T.; Carolino, N.; Costa, M. S. & Matos, C. P. (2004). Recursos genéticos animais em Portugal. Consultado em 17 dez. 2018. Disponível em http://www.fao.org/3/a1250e/annexes/CountryReports/Portugal.pdf

[26] Hosmer,D.W & Lemeshow, S. (2013). Applied Logistic Regression, 4rd edition, John Wiley, New York.

[27] INE (2017). Estatística Agrícolas - 2017. Consultado em 29 de nov. 2018. Disponível em https://www.ine.pt/xurl/pub/320461359.ISSN0079-4139.ISBN978-989-25-0445-2

[28] INE (2020). Consumo de carne per capita por tipos de carne; anual. Consultado em 18 jan. 2020. Disponível em https://www.ine.

[29] IFAP (2019). Animais residentes por região e raça. Consultado em 08 ago. 2020. Disponível em https://www.ifap.pt/estatisticas-animais

[30] Konishi, S. & Kitagawa, G. (2008) Information criteria and statistical modeling. New York: Springer.

[31] Levine, D.M., Berenson, M.L. & Stephan, D. (1996). Basic Business Statistics: Concepts and Applicatios. Upper Saddle River, NJ: Prentice Hall, 6.ed.

[32] Liu, Y. (2007). On goodness of fit of logistic regression model. Consultado em 13 ago. 2020. Disponível em `https://core.ac.uk/download/pdf/5164624.pdf`

[33] Mazerolle, M.J. (2004). Mouvements et reproduction des amphibiens en tourbières perturbées. 78p. Tese (Doutorado em ciências Florestais) - Université Laval, Québec.

[34] Meyer, P.L. (1983). Probabilidade: aplicações a estatística. 2.ed. Rio de Janeiro: LTC. 421

[35] Miloca, S. A., & Conejo, P. D. (2013). Multicolinearidade em modelos de regressão. XXII Semana Acadêmica da Matemática, Maringá.

[36] Murteira, B. J. F (1993). Análise Exploratória de dados Estatística descritiva. Lisboa: Editora Mc Graw-Hill de Portugal. 325 páginas

[37] Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models. Journal of the Royal Statistical Society, A 135, 370-384

[38] Pais, J.; Fernandes, L. & Minhoto, M. (2019). Análise técnico-económica da produção de vitelão Mertolengo DOP no centro de testagem e recria da Associação de criadores de bovinos Mertolengos. Consultado em 01 set. 2019. Disponível em `https://dspace.uevora.pt/rdpc/bitstream/10174/25726/1/AICA_Vol13_Trabajo013.pdf`

[39] Paula, A. G. (2013). Modelos de regressão com apoio computacional. Instituto de Matemática e Estatística: Universidade de São Paulo.

[40] Pek, J.; Wong, O. & Wong A. C. M. (2018). How to address non-normality: A taxonomy of approaches, reviewed and illustrated. Consultado em 29 jan. 2020. Disponível em `https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02104/full`

[41] Pino, F. A. (2014) A questão da não normalidade: Uma revisão. Revista de economia agrícola, São Paulo. p. 17 - 33.

[42] Rodrigues, S.C.A. (2012). Modelos de regressão linear e suas aplicações: Relatório de estágio para obtenção do grau de mestre. Consultado em 18 mar. 2020. Disponível em `http://hdl.handle.net/10400.6/1869`

[43] Ruralbit (2020). ruralbit: Tecnologia ao serviço do mundo rural. Consultado em 26 fev. 2019. Disponível em `https://www.ruralbit.pt/`

[44] Sharma, S. (1996). Applied multivariate techniques. Hoboken: John Wiley & amp; Sons.

[45] Turkman, M. A. A. & Silva, G. L. (2000). Modelos Lineares Generalizados: da teoria à prática. Lisboa: FCT PRAXIS XXI FEDER.

[46] Verbeek, M. (2017). A guide to modern econometrics. Hoboken: John Wiley & amp; Sons.

[47] Wooldridge, J.M. (2010). Introdução à econometria: Uma abordagem moderna. São Paulo: Cengage Learning. p. 280 315.

A

Apêndice 1

Neste apêndice são apresentados os principais comandos utilizados no software R para a análise descritiva das principais variáveis identificadas neste estudo. Carrega-se o pacote necessário para trabalhar com a base de dados que está organizada em Excel:
library(readxl)
A base de dados composta por todos os animais foi nomeada como dados.

```
dados=read.table("bov_mais_RECRIAalterado.csv", header=T, sep=";", dec=",")
```

A base de dados composta apenas por animais que foram ao destino de abate convencional foi nomeada como dadosCONV e a base de dados composta apenas por animais que foram ao destino de abate DOP foi nomeada de dadosDOP.

```
dadosCONV=read.table("bov_mais_CONV.csv", header=T, sep=";", dec=",")
dadosDOP=read.table("bov_mais_DOP.csv", header=T, sep=";", dec=",")
```

Foi carregado o pacote necessário para análise exploratória de dados e realizada a verificação dos animais por destino de abate conforme disponibilizado na tabela A.1:

```
library(fBasics)
table(dados$Dest_abate)
```

Tabela A.1: Verificação dos animais por destino de abate.

	DOP	CONV
N. animais	376	338

Foi então transformada a variável categórica destino de abate numa variável dummy, com o destino DOP como 1 (tabela A.2), que mais à frente será usada na regressão logística como o sucesso.

```
dados$Dest_abate2<-as.numeric(factor(dados$Dest_abate))-1
table(dados$Dest_abate2)
```

Tabela A.2: Confirmação da transformação da variável Destino de abate numa variável dummy, com o 1 sendo o destino de abate DOP.

	1	0
N. animais	376	338

Iniciou-se a análise descritiva com a variável idade à entrada, através dos seguintes comandos:

```
basicStats(dados$Idadeent_meses)
basicStats(dadosDOP$Idadeent_meses)
basicStats(dadosCONV$Idadeent_meses)
```

O uso destes comandos nas variáveis deram origem às informações que foram utilizadas na tabela 3.1, também foi possível verificar que esta variável não atende a normalidade (valor-p do teste de Kolmogorov Smirnov $< 0{,}001$ e histograma na figura A.1).

```
grafico1<-hist(dados$Idadeent_meses, main = "Idade na entrada dos currais",
xlab = "Idade (em meses)", ylab = "Frequencias")
min(dados$Idadeent_meses)
seq(1,5,length=22)
xajust<-seq(min(dados$Idadeent_meses, na.rm=T),
max(dados$Idadeent_meses, na.rm=T), length=50)
yajust<-dnorm(xajust, mean=mean(dados$Idadeent_meses,na.rm=T),
sd=sd(dados$Idadeent_meses,na.rm=T))

hist(dados$Idadeent_meses, freq = F, main = "Idade à entrada",
xlab = "Idade (em meses)", ylab = "densidade")
lines(xajust, yajust, col="red", lwd=2)

lillieTest(dados$Idadeent_meses)
```

Realizou-se então a comparação entre as medianas pelo teste Mann-Whitney U para verificar se a localização dos pontos medianos entre os grupos é igual a 0, entretanto pode-se perceber que existe diferenças estatísticas entre os grupos.

```
wilcox.test(dados$Idadeent_meses~dados$Dest_abate2)
Wilcoxon rank sum test with continuity correction
data:  dados$Idadeent_meses by dados$Dest_abate2
W = 40038, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Ignorando a não normalidade e assumindo o TLC, realizou-se também o teste t para verificar se existe diferença estatística significativa entre as médias das idades dos grupos de animais pelo comando:

```
t.test(dados$Idadeent_meses, dados$Dest_abate2)
Welch Two Sample t-test
data:  dados$Idadeent_meses and dados$Dest_abate2
t = 132.93, df = 890.69, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
7.305092 7.524039
sample estimates:
mean of x mean of y
7.9411765 0.5266106
```

E pode-se verificar que, a idade à entrada em meses dos animais DOP é superior quando comparada com os animais que vão ao abate convencional (p-value $<$ 2.2e-16), o que é obvio quando se conhece a dinâmica de produção.

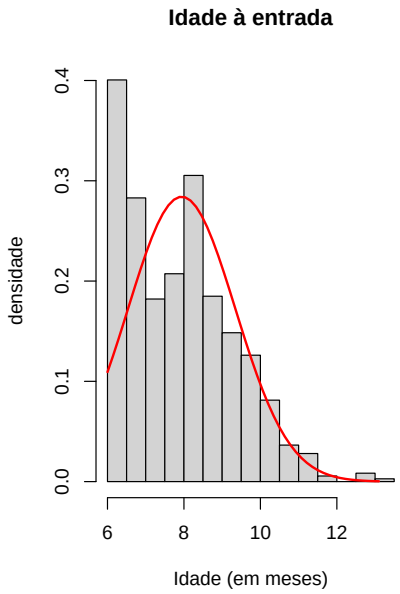


Figura A.1: Histograma da variável idade à entrada

Este mesmo procedimento foi utilizado para peso à entrada, idade à saída, peso à saída, número de dias na engorda, custo com profilaxia, taxa promert, custo individual com transporte, custo de funcionamento, custo com alimentação, custo total, lucro líquido entre outros.

Já para a análise do peso da carcaça, para além da análise anterior, também realizou-se a categorização da carcaça por classes, conforme o preço por kg de carcaça informado pela ACBM.

```
basicStats(dados$pesodacarcaca)
basicStats(dadosDOP$pesodacarcaca)
basicStats(dadosCONV$pesodacarcaca)
```

A categorização das faixas de peso da carcaça foi feita de acordo com o informado pelos produtores seguiu o script abaixo.

```
summary(dados$pesodacarcaca)
dados$pesocarccat<-cut(dados$pesodacarcaca, breaks=c(min(dados$pesodacarcaca),
120,140,160, 180, max(dados$pesodacarcaca)), include.lowest=T)
table(dados$pesocarccat)
```

Tabela A.3: Tabela de frequência observada por faixa de peso de carcaça.

[99.3,120[	[120,140[	[140,160[	[160,180[	[180,274]
11	105	153	177	268

Também foi possível observar, a frequência dos indivíduos por destino de abate.

```
table(dados$pesocarccat, dados$Dest_abate)
```

Tabela A.4: Número de observações da faixa de peso de carcaça por destino de abate.

	0 (CONV)	1 (DOP)
[99,3; 120[	8	3
[120; 140[	97	8
[140; 160[	121	32
[160; 180[	75	102
[180; 274]	37	231

```
chisq.test(dados$pesocarccat, dados$Dest_abate2)
Pearson's Chi-squared test
data:  dados$pesocarccat and dados$Dest_abate2
X-squared = 272.78, df = 4, p-value < 2.2e-16
```

É possível observar que existe associação entre o peso da carcaça e o destino de abate (p-value $<$ 2.2e-16). Sendo que as carcaças mais pesadas estão mais associadas ao destino de abate DOP, foi então feito a verificação da variável sem a categorização através do comando abaixo que deu origem a tabela A.5.

Tabela A.5: Peso da carcaça por destinos de abate

$CONV					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
102,9	137,9	150,1	154,2	165,1	252,9
$DOP					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
99,3	171,6	188,3	188,7	203,2	273,7

```
tapply(dados$pesodacarcaca, dados$Dest_abate, summary)
```

Também foi realizado o estudo da normalidade para verificação do peso da carcaça pelos grupos de abate.

```
lillieTest(dados$pesodacarcaca)
Title:
Lilliefors (KS) Normality Test
Test Results:
  STATISTIC:
  D: 0.041
  P VALUE:
  0.006259
```

```
wilcox.test(dados$pesodacarcaca, dados$Dest_abate2)
Wilcoxon rank sum test with continuity correction
data:  dados$pesodacarcaca and dados$Dest_abate2
W = 509796, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Ignorando a não normalidade e apelando para o TLC, foi obtido o mesmo resultado.

```
t.test (dados$pesodacarcaca, dados$Dest_abate2)
Welch Two Sample t-test
data:  dados$pesodacarcaca and dados$Dest_abate2
t = 152.68, df = 713.39, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
169.6162 174.0350
sample estimates:
mean of x   mean of y
172.3522126   0.5266106
```

Portanto, é possível perceber que existe diferença estatística significativa (p-value $<$ 2.2e-16) no valor médio do peso da carcaça entre os grupos de abate. E o animal DOP apresenta carcaça com peso maior (peso médio de carcaça de 188 kg) quando comparado com o animal convencional (peso médio de carcaça de 154 kg).

No caso do custo total de produção a análise ainda foi mais detalhada para poder perceber as características dessa variável entre os grupos de produção

```
basicStats(dados$total_custos)
basicStats(dadosDOP$total_custos)
basicStats(dadosCONV$total_custos)

tapply(dados$total_custos, dados$Dest_abate, summary)
```

Através do último comando, foi possível perceber que a média do custo total de produção dos animais DOP é superior quando comparada com os animais convencionais (tabela A.6).

Tabela A.6: Custo total de produção pelos grupos de abate

Animais Convencionais					
Min	1St Qu.	Median	Men	3rd Qu.	Max.
79,33	193,68	268,49	272,13	323,78	779,06
Animais DOP					
Min	1St Qu.	Median	Men	3rd Qu.	Max.
140,8	372,6	429,8	435,7	489,4	755,2

Pode-se verificar também que a variável aparentemente não segue a normalidade (figura A.2) e confirmado pelo teste de Kolmogorov Smirnov (valor-p = 0,001).

```
grafico11<-hist(dados$total_custos, main = "Custo total de producao",
xlab= "Custo total de producao, (euro)", ylab = "Frequencias")
min(dados$total_custos)
seq(1,5,length=22)
xajust<-seq(min(dados$total_custos, na.rm=T),
        max(dados$total_custos, na.rm=T),
        length=50)
yajust<-dnorm(xajust,
        mean=mean(dados$total_custos,na.rm=T),
        sd=sd(dados$total_custos,na.rm=T))

hist(dados$total_custos, freq = F, main = "Histograma do custo total de producao",
xlab = "Custo total de produção (euro)", ylab = "densidade")
lines(xajust, yajust, col="red", lwd=2)



lillieTest(dados$total_custos)
Title:
Lilliefors (KS) Normality Test
Test Results:
  STATISTIC:
  D: 0.0462
  P VALUE:
  0.001003
```

Levando em consideração a não normalidade e ignorando o fato de poder assumir o TLC, foi verificado a diferença entre as medianas através do teste de Mann-Whitney U.

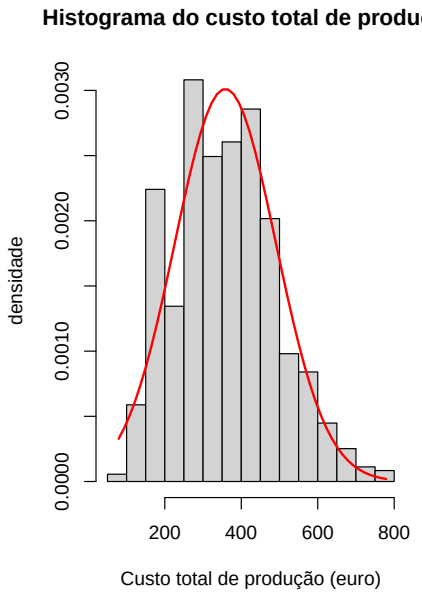


Figura A.2: Histograma da variável custo total de produção

```
wilcox.test(dados$custototal~dados$Dest_abate2)
Wilcoxon rank sum test with continuity correction
data:  dados$custototal by dados$Dest_abate2
W = 41190, p-value = 4.531e-16
alternative hypothesis: true location shift is not equal to 0
```

Assumindo o TLC também foi verificado pelo teste t, apenas por questão de curiosidade.

```
t.test(dados$total_custos,dados$Dest_abate2)
Welch Two Sample t-test
data:  dados$total_custos and dados$Dest_abate2
t = 72.154, df = 713.02, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
347.9922 367.4595
sample estimates:
mean of x   mean of y
358.2524790   0.5266106
```

Portanto, existe diferença estatistica significativa (p-value $<$ 2.2e-16) na média do custo total de produção entre os grupos de destino de abate, com o animal DOP sendo o mais oneroso para a produção quando comparado com o animal convencional. Para além destes análises, também se fez a análise de correlação do ponto bisserial para verificar se a variável custo total de produção variável tinha associação com o destino de abate.

```
Library(ltm) #Pacote necessário para fazer a análise da correlação do ponto bisserial,
que é o mais indicado para o caso em estudo, que se trata de um dado dicotômico
```

```
(Destino de abate) e um dado quantitativo (Custo de produção).

biserial.cor(dados$total_custos, dados$Dest_abate)
[1] -0.616814
```

O custo de produção total e o destino de abate apresentam associação moderada e negativa. Para entender a direção dessa relação, foi feita uma verificação de qual categoria o software estava assumindo como 1.

```
table(dados$total_custos, dados$Dest_abate)
```

Pode-se perceber que para esta análise, o R assumiu o DOP como 0 e o abate convencional 1, portanto, à medida que aumenta o custo total de produção indica que aumenta a tendência para que o animal vá para o abate DOP e isso explica o sinal negativo dessa correlação. Verificamos também a composição dos custos fixos e variáveis do custo total (Custo total = custo fixo + custo variável).

```
dados$CF<-(dados$profilaxia+dados$servico_Promert+dados$custotransporte.kgcarcaca)
summary(dados$CF)
```

Tabela A.7: Valor do custo fixo da produção de bovinos mertolengas

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
28,15	28,18	38,18	37,31	38,22	53,64

E os custos variáveis que posteriormente compuseram o custo por dia de produção são: Custo alimentação, Custo funcionamento e outros custos.

Depois do estudo detalhado do custo total de produção, realizou-se o estudo do custo por dia de produção.

```
basicStats(dados$Custototal_dia)
basicStats(dadosDOP$Custototal_dia)
basicStats(dadosCONV$Custototal_dia)
```

Já com interesse na modelação, realizou-se a verificação da normalidade desta variável.

```
lillieTest(dados$custototal.dia)
Test Results:
  STATISTIC:
  D: 0.0204
  P VALUE:
  0.6713
```

É possível perceber que o custo por dia de produção apresenta o gráfico com formato de sino (figura A.3) e através do valor-p de 0,67 do teste de Kolmogorv Smirnov é possível concluir que a variável custo por dia de produção atende à normalidade.

```
t.test(dados$custototal.dia, dados$Dest_abate2)
Welch Two Sample t-test
data:  dados$custototal.dia and dados$Dest_abate2
t = 80.753, df = 994.37, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
1.617705 1.698285
sample estimates:
mean of x mean of y
2.1846057 0.5266106
```

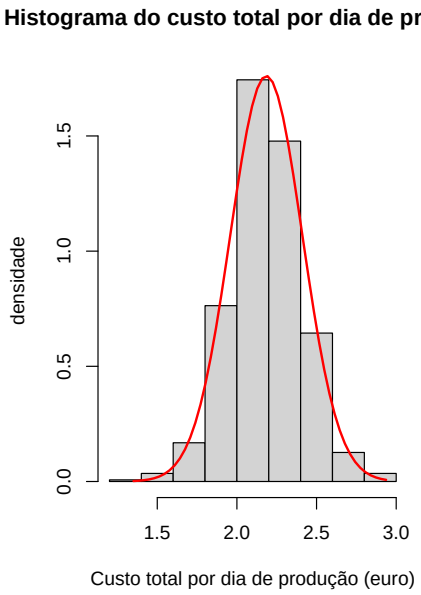


Figura A.3: Histograma da variável custo por dia de produção

Já para verificar a associação pelo ponto bisserial, é possível perceber que existe uma relação mais fraca entre o custo por dia de produção e o destino de abate, quando comparada com o custo total de produção.

```
biserial.cor(dados$custototal.dia, dados$Dest_abate)
[1] -0.2957118
```

Pensando em trabalhos futuros, o valor do lucro líquido da recria também foi estudado com mais atenção.

```
basicStats(dados$liquido_recria)
basicStats(dadosDOP$liquido_recria)
basicStats(dadosCONV$liquido_recria)
summary(dados$liquido_recria)
```

Pode-se perceber que essa variável atende ao pressuposto da normalidade a 1% de nível de significância e pelo teste t é possível perceber que existe diferença estatistica significativa (p-value $<$ 2.2e-16)

Tabela A.8: Valores do lucro líquido da recria da produção de vitelão mertolenga

Manada					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
39,29	264,09	323,43	318,50	373,41	680,08
CONV					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
131,20	269,10	320,20	312,30	357,40	463,50
DOP					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
39,29	252,36	328,14	324,04	389,71	680,08

entre o valor médio do lucro dos animais que foram ao abate DOP quando comparado ao valor médio dos animais que foram ao abate convencional. Um estudo posterior relacionando custo e lucro poderia ser de interesse para o complemento de informações estratégicas para a associação de bovinos mertolengas.

```
lillieTest(dados$liquido_recria)
Title:
Lilliefors (KS) Normality Test
Test Results:
  STATISTIC:
  D: 0.0363
  P VALUE:
  0.02657

t.test(dados$liquido_recria, dados$Dest_abate2)
One Sample t-test
data:  dados$liquido_recria
t = 102.36, df = 713, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
312.3934 324.6108
sample estimates:
mean of x
318.5021
```

Quanto ao Preço/kg de carcaça, além da análise descritiva também fez-se a categorização de faixa de preço, por faixa de preço conforme indicado na tabela A.9.

```
basicStats(dados$preco_kg_carcaca)
basicStats(dadosDOP$preco_kg_carcaca)
basicStats(dadosCONV$preco_kg_carcaca)

t.test( dados$preco_kg_carcaca, dados$Dest_abate2)
One Sample t-test
data:  dados$preco_kg_carcaca
t = 629.43, df = 713, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
3.897849 3.922241
sample estimates:
mean of x
3.910045
```

É possível verificar estatisticamente (p-value $<$ 2.2e-16) que os animais que foram ao abate DOP apresentaram maior valorização do valor do kg/carcaça, quando comparado com os animais que foram ao abate convencional.

```
(dados$precokgcarccat<-factor(dados$pesocarccat,
labels = c("3.60", "3.60","3.85","4.00","4.15")))
round (prop.table(table(dados$precokgcarccat, dados$Dest_abate))*100, digits = 2)
```

Tabela A.9: Percentagem do preço por kg de carcaça por destino de abate.

	CONV	DOP
Até 3,60 euros	14,71	1,54
de 3,60 a 3,85 euros	16,95	4,48
de 3,85 a 4,00 euros	10,5	14,29
de 4,00 a 4,15 euros	5,18	32,35

Para ter maior noção da associação entre as variáveis fez-se uma correlação de Pearson entre as principais variáveis estudadas (figuras A.4 e A.5), que por se tratar da relação entre variáveis quantitativas é o mais indicado.

```
Varinteresse <- data.frame (dados$Idadeent_meses, dados$peso_ent,
dados$Rendimento_carcaca, dados$preco_kg_carcaca, dados$pesodacarcaca,
dados$Diasengorda, dados$custototal.dia, dados$liquido_recria)
Varinteresse <- na.omit(varinteresse)
Corvarinteresse <- round (cor (varinteresse, method = "pearson"), 2)
Str (varinteresse)
install.packages ("ggcorrplot")
library (ggcorrplot)

ggcorrplot(Corvarinteresse, hc.order = T,
          type = "full",show.legend = TRUE,
          lab = TRUE, show.diag = T,
          lab_size = 3, lab_col = "black",
          method="square", outline.color = "black",
          colors = c("red", "white", "blue"), tl.cex = 10,
          title ="", legend.title = "Correlação",
          ggtheme=theme_bw, digits = 2)
```

É possível perceber um valor positivo e moderado de associação entre os dias na engorda e o peso da carcaça, e o peso da carcaça e o custo total por dia de produção (figura A.4).

Fez-se também a mesma análise descritiva para as variáveis genéticas e os valores obtidos deu origem a tabela 3.2, e também foi possível perceber que existe uma relação forte e direta entre a capacidade de

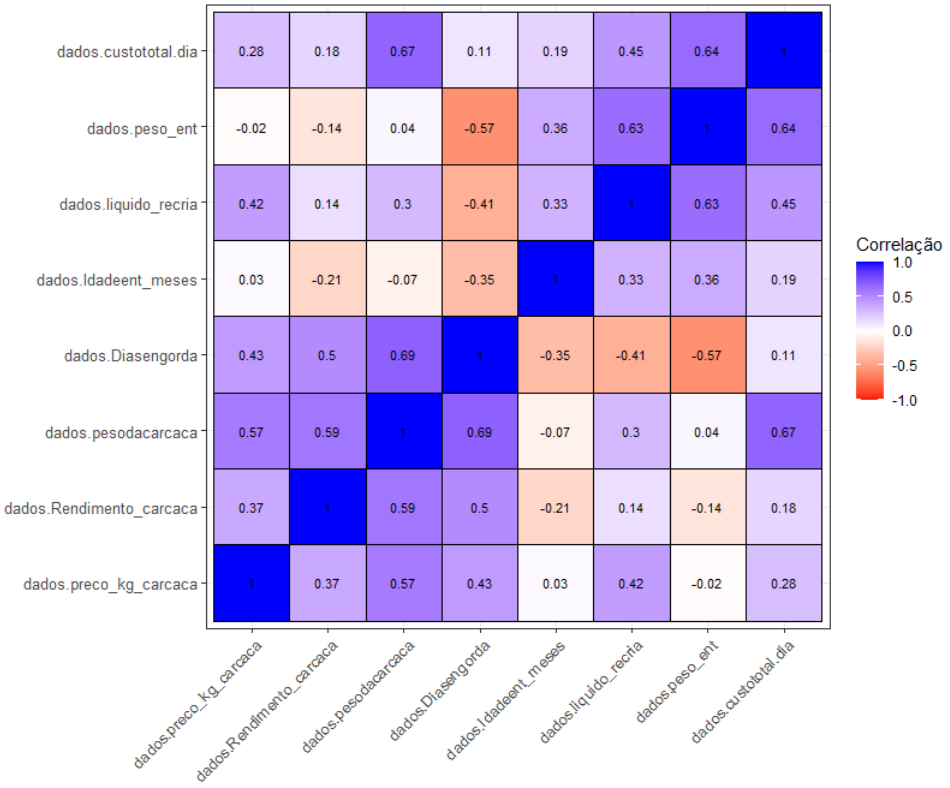


Figura A.4: Quadro de correlação das principais variáveis zootécnicas.

crescimento e o VG do GMD em teste de estação. Em contra partida, há uma correlação forte e inversa entre a capacidade de crescimento e o índice de conversão e a capacidade de crescimento e a carcaça por dia de idade (figura A.5).

Os valores genéticos também foram estudados conforme apresentado, entretanto, apenas serão mostrados neste apêndice os resultados dos valores genéticos da capacidade maternal e o intervalo entre partos, uma vez que foi identificado a sua importância para a produção da raça mertolenga.

```
summary(dados$VG_capacidade_maternal)
tapply(dados$VG_capacidade_maternal, dados$Dest_abate, summary)
```

Tabela A.10: Resumo do Valor genético da capacidade maternal, valor da manada e dos grupos de abate.

Manada					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-18,28	-4,96	-0,98	-0,38	3,66	24,77
CONV					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-18,28	-4,41	-0,78	-0,52	3,14	24,77
DOP					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-17,75	5,62	-1,17	-0,25	4,59	23,24

É possível perceber que os animais DOP apresentaram maior capacidade maternal, sendo que para essa variável é melhor quanto maior for. Foi possível perceber também que a variável não atende à

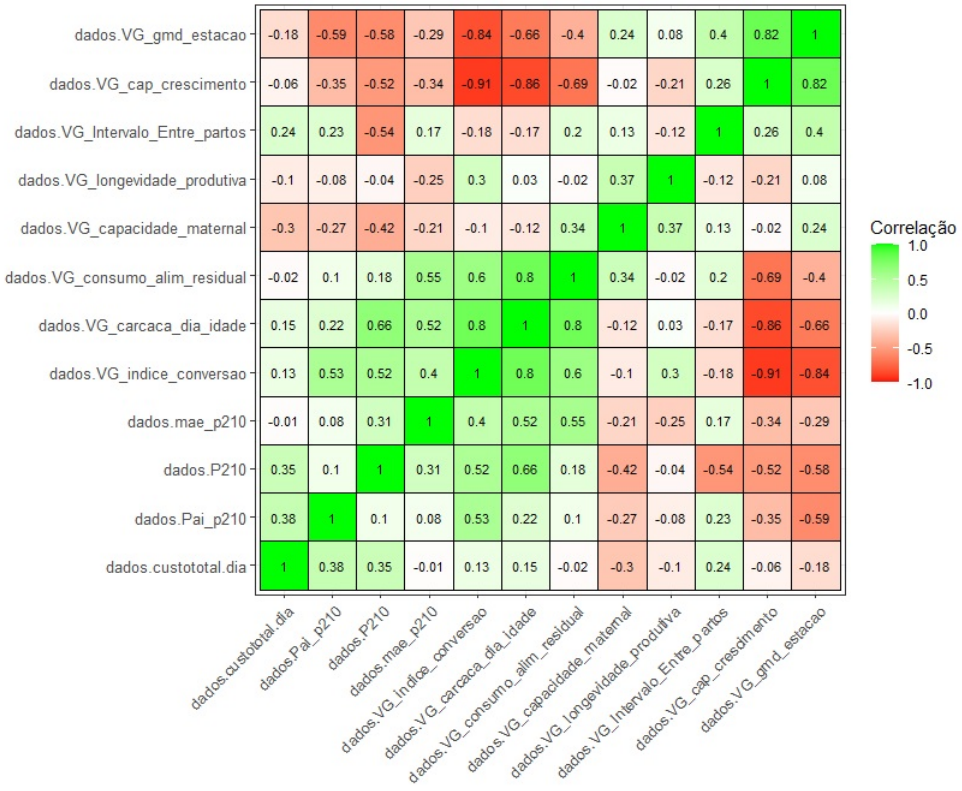


Figura A.5: Quadro de correlação entre as variáveis genéticas e algumas variáveis zootécnicas.

normalidade.

```
lillieTest(dados$VG_capacidade_maternal)
Title:
Lilliefors (KS) Normality Test
Test Results:
  STATISTIC:
  D: 0.0476
  P VALUE:
  0.0005908
```

```
wilcox.test(dados$VG_capacidade_maternal)
Wilcoxon signed rank test with continuity correction
data:  dados$VG_capacidade_maternal
V = 112536, p-value = 0.006197
alternative hypothesis: true location is not equal to 0
```

Pela análise do teste de Mann-Whitney rejeitamos hipótese nula de valores iguais de mediana entre os grupos, portanto existe diferença entre eles. Entretanto, ao análisar o teste t (assumindo o TLC) e realizando a análise gráfica, a sugestão é de que não há diferença entre os grupos.

```
t.test(dados$VG_capacidade_maternal, dados$Dest_abate2)
One Sample t-test
```

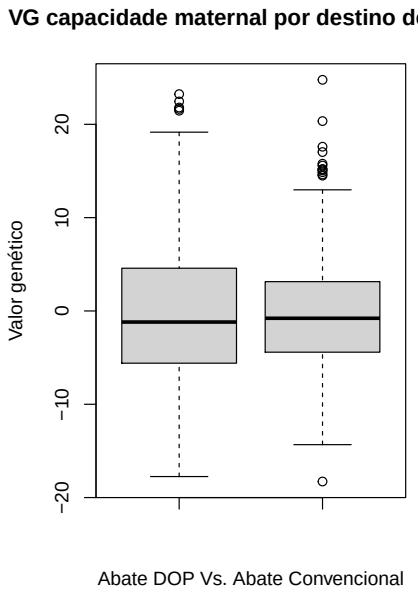


Figura A.6: Boxplot do valor genético da capacidade maternal pelos grupos de abate da produção de vitelão mertolengo.

```
data:  dados$VG_capacidade_maternal
t = -1.5065, df = 713, p-value = 0.1324
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.8787117  0.1156865
sample estimates:
 mean of x
-0.3815126
```

Quanto a variável intervalo entre partos podemos verificar uma maior amplitude de variação (-39 a 33) da unidade de medida, com os animais DOP apresentando uma média maior deste valor genético quando comparado com o animal convencional. Sendo que a variável intervalo entre partos é melhor quanto menor for, pois indica que as vacas estão voltando para o ciclo da reprodução mais rápido.

```
summary(dados$VG_Intervalo_Entre_partos)
tapply(dados$VG_Intervalo_Entre_partos, dados$Dest_abate, summary)
```

```
lillieTest(dados$VG_Intervalo_Entre_partos)
Title:
Lilliefors (KS) Normality Test
Test Results:
 STATISTIC:
 D: 0.0484
 P VALUE:
 0.0004272
```

Tabela A.11: Resumo do valor genético do intervalo entre partos da produção de vitelão mertolengo e seus grupos de abate

Manada					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-39,30	-16,25	-6,03	-5,55	5,56	33,88
CONV					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-39,30	-16,90	-8,51	-7,77	1,77	31,23
DOP					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-36,99	-14,96	-3,85	-3,56	9,13	33,88

```
wilcox.test(dados$VG_Intervalo_Entre_partos)
Wilcoxon signed rank test with continuity correction
data:  dados$VG_Intervalo_Entre_partos
V = 76524, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
```

Para esta variável foi possível verificar o mesmo resultado, tanto no teste paramétrico (assumindo o TLC), quanto no não paramétrico. Portanto, podemos afirmar que existe diferença estatisticamente significativa para o valor genético do intervalo entre partos, entre os grupos de abate, sendo que os animais convencionais apresentam menor valor desta característica, que é o desejável na produção de vitelão.

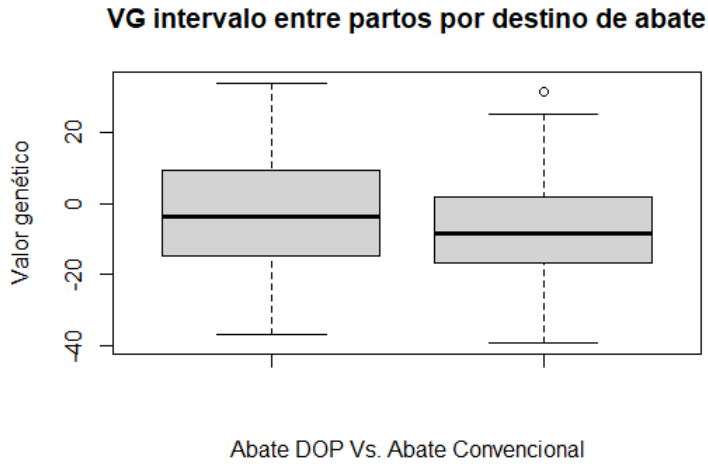


Figura A.7: Boxplot do valor genético do intervalo entre partos pelos grupos de abate da produção de vitelão mertolengo.

B

Apêndice 2

Nesta secção são apresentados alguns comandos e análises realizados para obter os resultados do modelo geral apresentado no capítulo 4, que tinha como objetivo tentar obter um modelo do custo por dia de produção com as informações disponíveis na entrada do animal no CTR. Iniciou-se a modelação com o modelo simples, sendo a primeira variável peso à entrada.

```
modcusto1<-lm(dados$custototal.dia~dados$peso_ent)
summary(modcusto1)
Call:
lm(formula = dados$custototal.dia ~ dados$peso_ent)
Residuals:
     Min       1Q   Median       3Q      Max
-0.91658 -0.11453 -0.01645  0.10093  0.53894

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)    1.5448335  0.0295126   52.34   <2e-16 ***
dados$peso_ent 0.0036641  0.0001648   22.23   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1742 on 712 degrees of freedom
Multiple R-squared:  0.4096,Adjusted R-squared:  0.4088
F-statistic:   494 on 1 and 712 DF,  p-value: < 2.2e-16
```

E seguiu o mesmo molde para as demais variáveis.

```
#Custo por dia de produção em função da idade à entrada
modcusto2<-lm(dados$custototal.dia~dados$Idadeent_meses)
summary(modcusto2)
```

E assim por diante, com todas as variáveis para entender a relação individual com a variável de interesse. Em seguida fez-se o modelo de regressão linear múltiplo pelo método backward com as informações disponíveis à entrada dos animais no CTR.

```
modcustodiaM0<- lm(dados$custototal.dia~dados$peso_ent+dados$Idadeent_meses+
dados$VG_gmd_estacao+dados$VG_longevidade_produtiva+
dados$VG_consumo_alim_residual+dados$VG_indice_conversao+VG_cap_crescimento+
VG_capacidade_maternal+VG_carcaca_dia_idade+VG_Intervalo_Entre_partos,  data = dados)
summary(modcustodiaM0)
```

As variáveis foram retiradas à medida que não se apresentaram significativas no teste t (significância do coeficiente) com confiança de 99%. Foram retiradas as variáveis: VG_gmd_estacao, VG_longevidade_produtiva, VG_consumo_alim_residual e VG_indice_conversao. Que, por fim, geraram o modelo geral apresentado na tabela 4.1 que foi nomeado de modcustodiaM.

Realizou-se a análise dos resíduos no modelo obtido, a iniciar pelo teste de Kolmogorov-Smirnov com correção Lilliefors para verificação da normalidade dos resíduos.

```
resid<-resid(modcustodiaM)  #resíduos do modelo múltiplo
pred<-fitted(modcustodiaM)  #Valores ajustados pelo modelo
resid.std <- rstandard(modcustodiaM)

library(nortest)
lillie.test(resid)
Lilliefors (Kolmogorov-Smirnov) normality test
data:  resid
D = 0.057284, p-value = 8.417e-06
```

Pode-se verificar que não se admite a normalidade. Tentou-se avaliar o achatamento e a curtose como manobra para a ausência da normalidade (figura B.1).

```
library(moments)
anscombe.test(resid)
Anscombe-Glynn kurtosis test
data:  resid
kurt = 6.2291, z = 7.7096, p-value = 1.262e-14
alternative hypothesis: kurtosis is not equal to 3

agostino.test(resid)
D'Agostino skewness test
data:  resid
skew = -0.34919, z = -3.74274, p-value = 0.000182
alternative hypothesis: data have a skewness
```

Contudo, o resultado não se alterou. Os resíduos do modelo não atenderam à curtose, nem à simetria. Verificou-se então a independência dos resíduos:

```
library(car)
durbinWatsonTest(modcustodiaM)
 lag Autocorrelation D-W Statistic p-value
   1       0.5609262     0.8713225       0
 Alternative hypothesis: rho != 0
```

Pode-se verificar que os resíduos não atendem à independencia. Fez-se a verificação da homocedasticidade e os resíduos atenderam a este pressuposto.

```
library(lmtest)
bptest(modcustodiaM)
studentized Breusch-Pagan test
data:  modcustodiaM
BP = 7.8553, df = 5, p-value = 0.1644
```

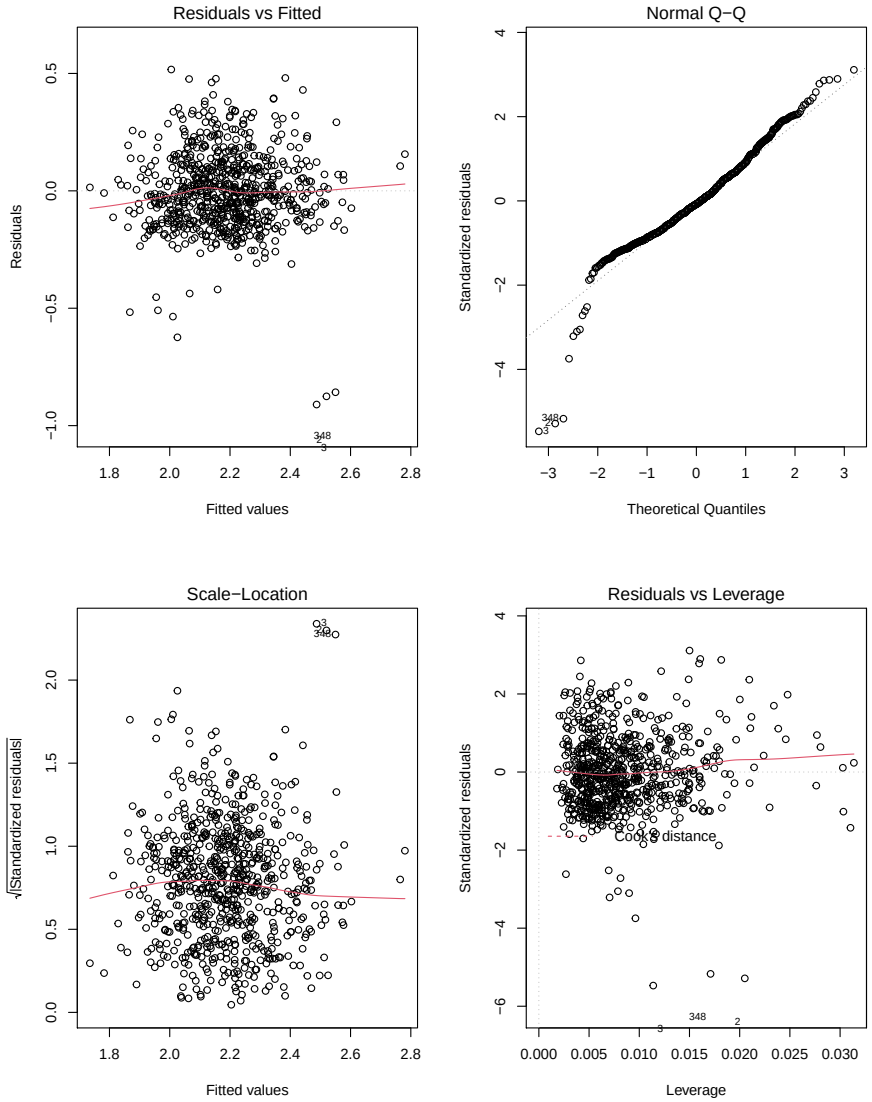
Para o teste de multicolinearidade, pode-se perceber que os resíduos também atendiam a este pressuposto.

```
vif(modcustodiaM) # variance inflation factors
                 peso_ent        VG_cap_crescimento    VG_capacidade_maternal
                 1.144825                  1.098327                  1.122326
     VG_carcaca_dia_idade VG_Intervalo_Entre_partos
                 1.202472                  1.090351
```

De forma geral, o modelo apresentou baixa capacidade explicativa e pela análise gráfica (figura B.1) é possível perceber o mal comportamento dos resíduos.

Fez-se então a verificação das observações que são consideradas outliers ao nível de significância de 1% (tabela B.1).

Figura B.1: Dispersão gráfica dos resíduos do modelo geral com os outliers presentes.




```
pvalue<-2*(1-pt(abs(resid.std),length(resid.std)))
names(pvalue)<-1:length(resid.std)
pvalue[pvalue<0.01]
```

Tabela B.1: Valores p do teste para identificar os indivíduos influentes do modelo obtido

Observação	2	3	263	264	265	266	267	268
Valor-p	<0.001	<0.001	0.005	0.001	0.001	0.001	0.002	0.001
Observação	333	337	338	348	360	363	710	
Valor-p	0.009	0.004	0.001	<0.001	0.009	0.003	0.004	

Os pontos extremos das variáveis explicativas num modelo regressão são detectados por meio da matriz dos valores esperados e como são muitos valores influentes fez-se a análise gráfica (figuras B.2 e B.3).

```
lev<-hatvalues(modcustodiaM)
lev
names(lev)<-1:length(lev)
plot(pred,lev,abline(h=0.2))
plot(pred,lev,ylim=c(0,0.2),abline(h=0.2))4
```

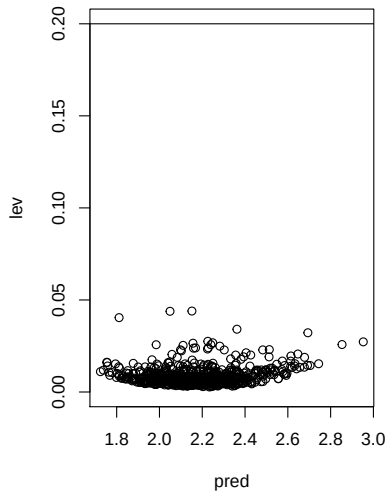


Figura B.2: Análise dos valores influentes do modelo geral para o custo por dia de produção.

Nesta análise nenhuma observação ultrapassou o limite indicado, portanto, realizou-se outro meio gráfico de observação de valores influentes que é a análise da distância de Cook (figura B.3), onde é possível perceber que nenhum indivíduo ultrapassa o limite sugerido.

```
dcook<-cooks.distance(modcustodiaM)
names(dcook)<-1:length(dcook)
plot(dcook,ylim=c(0,1.1))
abline(h=1.0,col="red")
identify(dcook)
```

Decidiu-se por retirar os animais que se apresentaram outliers no teste-t da base de dados para obtenção do modelo sem assumir o TLC.

```
dados2<-dados[-c(2,3,263,264,265,266,267,268,333,337,338,348,360,363,710),]
```

E verificou-se novamente os pressupostos do modelo sem estes indivíduos, que gerou o modelo apresentado na tabela 4.4. E portanto, o modelo atende curtose (e normalidade via TLC), homocedasticidade e ausência de multicolinearidade, além de explicar 53% do custo por dia de produção. O comando utilizado para chegar ao modelo apresentado foi:

```
modcustodiaMB2<-lm(custototal.dia~peso_ent+VG_cap_crescimento+VG_capacidade_maternal+VG_carcac
VG_Intervalo_Entre_partos, data = dados2)
```

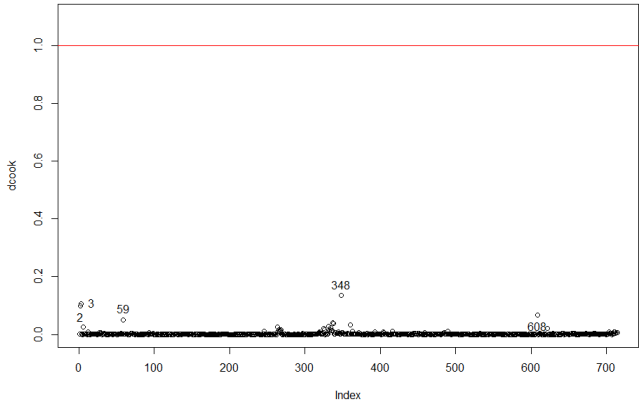
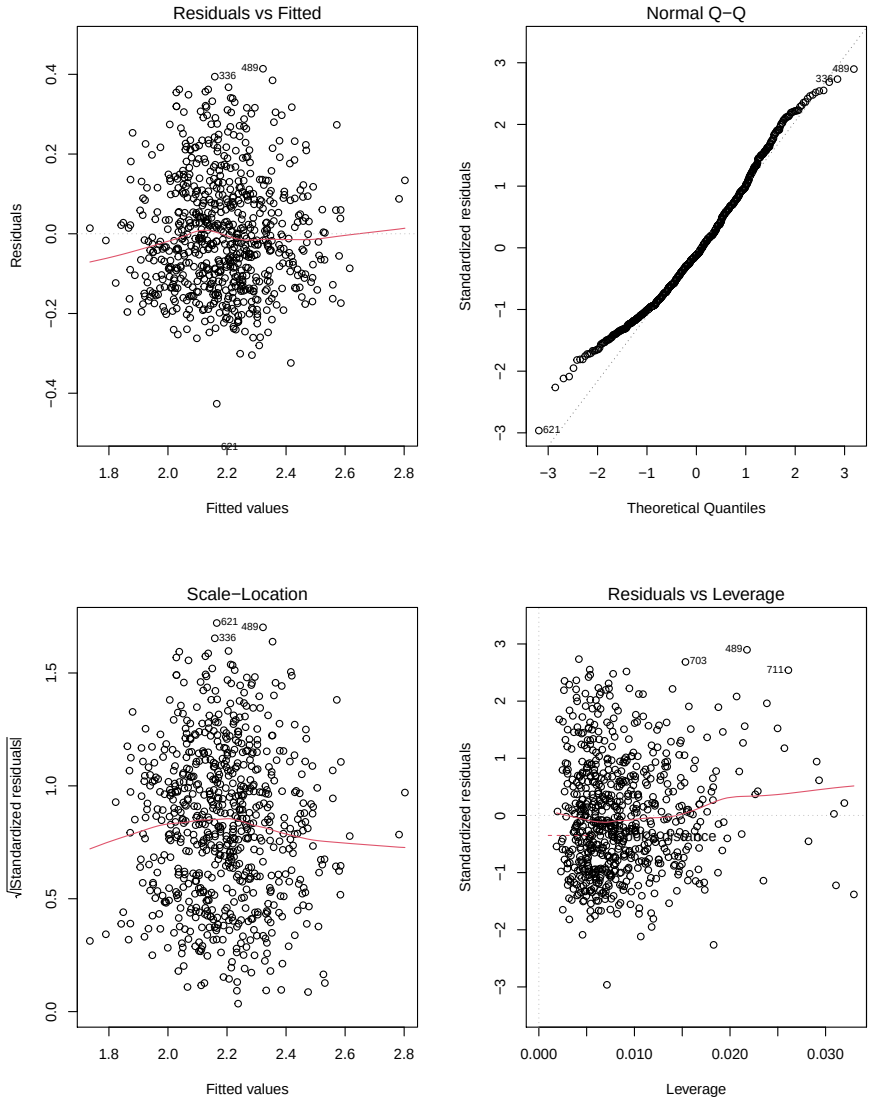


Figura B.3: Distância de Cook.

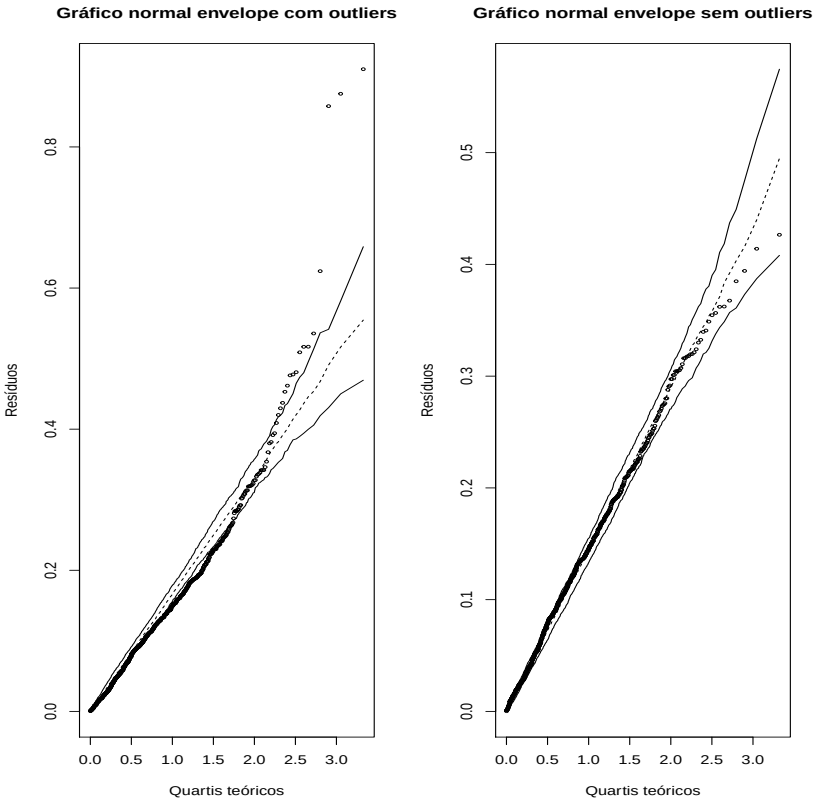
Figura B.4: Dispersão gráfica dos resíduos do modelo geral sem os outliers presentes.



Retirar os outliers é uma técnica que exige atenção, pois pode perder informações importantes. Para o caso em estudo, além de melhorar a capacidade explicativa do modelo (com o aumento do coeficiente de determinação), também é possível notar um melhor comportamento dos resíduos na análise gráfica (figura B.4).

Essa mesma verificação gráfica foi realizada com os demais modelos por grupos de abate, sendo que no caso dos animais que seguiram para o abate convencional, não foi necessário retirar os outliers uma vez que a variável resposta foi transformada.

Figura B.5: Avaliação dos quartis dos resíduos dos modelos obtidos pela técnica de regressão linear múltipla com e sem outliers.



Os gráficos apresentados foram obtidos através do comando:

```
par(mfrow = c(2,2))
plot(modcustodiaM32)
```

Figura B.6: Dispersão dos resíduos do modelo do custo de produção por dia na engorda dos animais DOP sem os outliers.

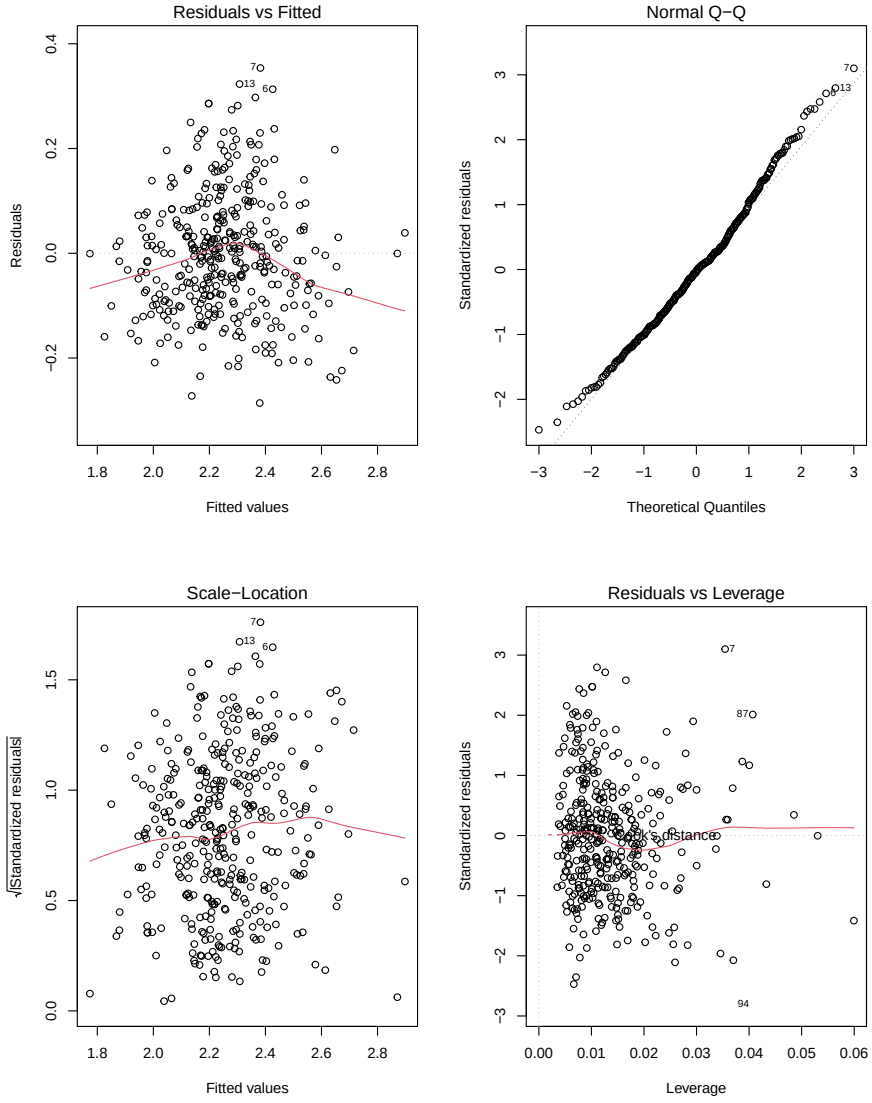
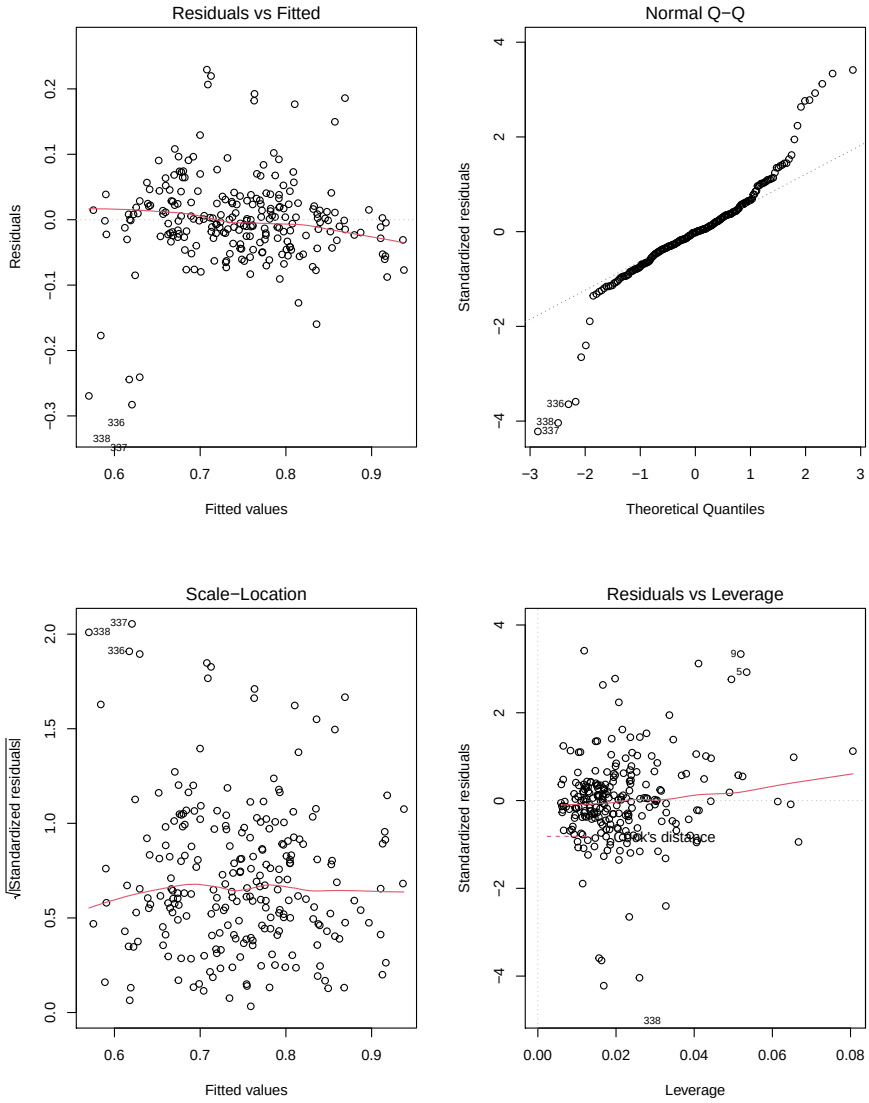


Figura B.7: Dispersão dos resíduos do modelo final do custo de produção por dia na engorda dos animais que foram ao abate convencional.



B.1 Modelo Misto

Após essa fase da modelação decidimos verificar se o produtor poderia ter algum efeito sobre o custo por dia de produção. Portanto, seguiu-se por fazer um modelo misto no qual o produtor seria a variável aleatória. Foi possível perceber que nessa base de dados existiam 47 produtores e que alguns produziam poucos animais, sendo que, alguns produtores produziam apenas 1 animal. Foi possível verificar que alguns produtores dividiam-se na base, como por exemplo, o produtor 164 e 164a. Mas como não obtivemos muitas informações acerca dos produtores e nem como estes poderiam ser agrupados, manteve-se a formação original de acordo com a tabela B.2.

Foi necessário carregar o pacote abaixo para ajustar os modelos lineares e lineares generalizados de efeitos mistos.

Tabela B.2: Produção de animais por produtor.

Produtor	105	107	122	129	164	164a	181	187	19	201	232	233
N. animais	4	18	9	17	5	1	3	23	12	8	12	17
Produtor	234	267	283	299	315	315b	316	322	346	353	367	372
N. animais	2	11	1	9	1	5	9	2	2	6	36	32
Produtor	373	396	400	416	426	439	446	448	454	454a	456	458
N. animais	65	7	29	10	19	69	1	3	66	32	10	8
Produtor	459	465a	467	473	474	478	478a	65	79	80	94	
N. animais	2	7	3	4	1	11	44	21	23	32	27	

```
library(lme4)
mixed.lmer <- lmer(custototal.dia ~ peso_ent+VG_cap_crescimento+VG_capacidade_maternal+
VG_carcaca_dia_idade + VG_Intervalo_Entre_partos+ (1|criador_origem), data = dados2)
summary(mixed.lmer)
Linear mixed model fit by REML ['lmerMod']
Formula: custototal.dia ~ peso_ent + VG_cap_crescimento + VG_capacidade_maternal +
   VG_carcaca_dia_idade + VG_Intervalo_Entre_partos + (1 | criador_origem)
   Data: dados2

REML criterion at convergence: -743.9

Scaled residuals:
     Min       1Q   Median       3Q      Max
-3.07168 -0.70263 -0.07171  0.67680  2.83693

Random effects:
 Groups         Name        Variance Std.Dev.
 criador_origem (Intercept) 0.006349 0.07968
 Residual                   0.016518 0.12852
Number of obs: 699, groups:  criador_origem, 48

Fixed effects:
                           Estimate Std. Error t value
(Intercept)               1.5683245  0.0311439  50.357
peso_ent                  0.0035215  0.0001563  22.526
VG_cap_crescimento        0.0013617  0.0007765   1.754
VG_capacidade_maternal    0.0014162  0.0012524   1.131
VG_carcaca_dia_idade      0.0023895  0.0003715   6.432
VG_Intervalo_Entre_partos 0.0010387  0.0005198   1.998

Correlation of Fixed Effects:
            (Intr) pes_nt VG_cp_c VG_cpc_ VG_c__
peso_ent    -0.886
VG_cp_crscm  0.400 -0.366
VG_cpcdd_mt  0.201 -0.176  0.233
VG_crcc_d_d  0.033 -0.110 -0.189  -0.249
VG_Intrv_E_  0.133 -0.021  0.060   0.142  -0.035
```

Ao observar a variância quase nula dos efeitos aleatórios e a percentagem de explicação que tem, foi possível concluir que o efeito do criador é nulo e portanto, não o criador não apresenta grande efeito sobre o custos por dia de produção

C

Apêndice 3

Neste apêndice serão apresentados os comando utilizados para chegar aos resultados que deram origem ao capítulo 5. Os pacotes utilizados nesta fase foram: library(fBasics), library(mfp), library(car), library(rms).

A primeira família de distribuição testada foi a gaussiana com ligação identidade que deve equivaler ao modelo obtido pela RLM. E como de costume na modelação de GLM, primeiro é obtido o modelo nulo para este tipo de família.

```
fit0 <- glm(dados$custototal.dia ~ 1,  data=dados,  family=gaussian(link = "identity"))
summary(fit0)
```

Verificação do modelo múltiplo gaussiano com ligação identidade.

```
MULTGAUS<-glm(custototal.dia~peso_ent+VG_capacidade_maternal+VG_cap_crescimento+
VG_carcaca_dia_idade+VG_Intervalo_Entre_partos, data=dados,
family=gaussian(link = "identity"), na.action = na.exclude)
summary(MULTGAUS)
```

Para obter a variação explicada do modelo com resposta contínua, utilizou-se o comando:

```
1-((deviance(MULTGAUS)/(df.residual(MULTGAUS))))/
(MULTGAUS$null.deviance/MULTGAUS$df.null)
```

Para a análise da linearidade para covariáveis contínuas utilizou-se o método dos quartis, o método de lowess e o método dos polinômios fracionários. Abaixo tem-se um exemplo do script utilizado para o método dos quartis e os gráficos obtidos (de C.1 a C.5) para o modelo selecionado estão na sequência. Verificação da linearidade da variável peso à entrada pelo método dos quartis:

```
Qis <- as.numeric(quantile(dados$peso_ent, probs=seq(0, 1, 0.25)))
# Calcula os quartis da variável quantitativa (inclui max e min)
# Categorizar a variável quantitativa com base nos quartis
dados$pesoCAT<- cut(dados$peso_ent,   # variável a categorizar
                    breaks=Qis,           # pontos de corte das classes (nos quartis)
                    right=FALSE,         # classes abertas a direita
                    include.lowest=TRUE)   # a última classe e fechada a direita
table(dados$pesoCAT)

# Ajustar modelo com a variável quantitativa categorizada
MULTGAUS1a <-glm(custototal.dia~pesoCAT+VG_capacidade_maternal+
VG_cap_crescimento+ VG_carcaca_dia_idade+VG_Intervalo_Entre_partos,
             na.action = na.exclude,
             family=gaussian("identity"),
             data=dados)
# Coeficientes estimados
summary(MULTGAUS1a)
# Gráfico com os betas da variável quantitativa categorizada, sendo o 1 beta=0
k <- 5
x <- (Qis[1:(k-1)]+Qis[2:k])/2       # pontos médios das classes
y <- c(0, as.numeric(MULTGAUS1a$coef[2:4]))
plot(x, y,
     main="Linearidade de peso categorizada")
lines(lowess(x,y))
```

Para a verificação da linearidade pelo método de lowess, utilizou-se o seguinte comando (figura C.4):

```
plot(lowess(predict(MULTGAUS)~dados$VG_carcaca_dia_idade), type="l",
main="Linearidade pelo método de Lowess",xlab="Valor genético da carcaça por dia de idade",
ylab="logOdds")
```

Portanto, foi possível verificar que o pressuposto da linearidade foi atendido para o modelo obtido com a distribuição gaussiana e ligação identidade. No GLM também se realizou a análise dos resíduos via indivíduos para verificar se os indivíduos que se apresentam como outliers são os mesmos. Iniciou-se pela observação gráfica dos resíduos deviance (figura C.6) e distância de cook (figura C.7).

Os indivíduos que se destacaram na distância de Cook também foram outlier no modelo de regressão apresentado no capítulo 4. Para identificar as possíveis observações influentes utilizou-se o comando:

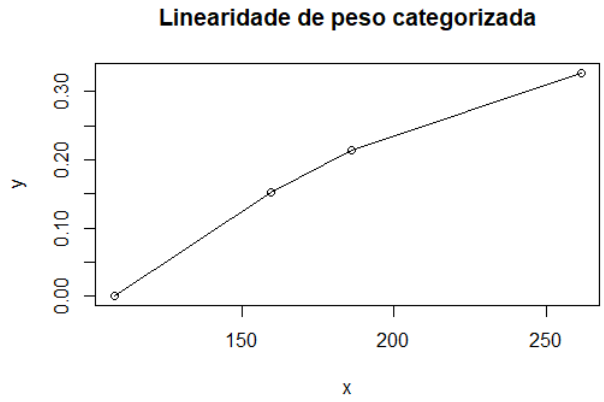


Figura C.1: Linearidade do peso à entrada categorizada para a distribuição gaussiana e ligação identidade.

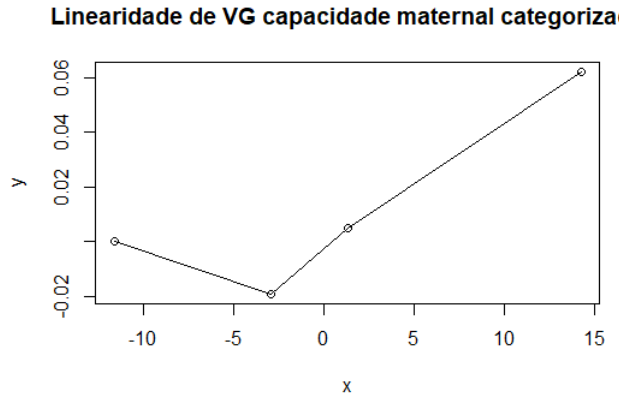


Figura C.2: Linearidade do valor genético da capacidade maternal categorizada para a distribuição gaussiana e ligação identidade.

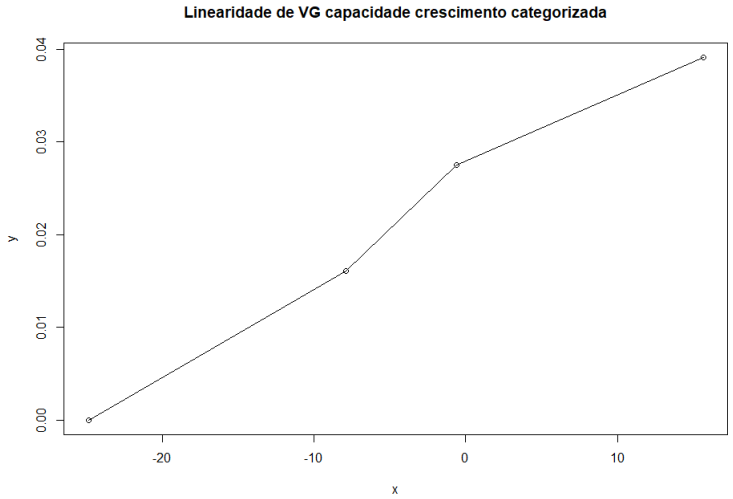


Figura C.3: Linearidade do valor genético da capacidade de crescimento categorizada para a distribuição gaussiana e ligação identidade.

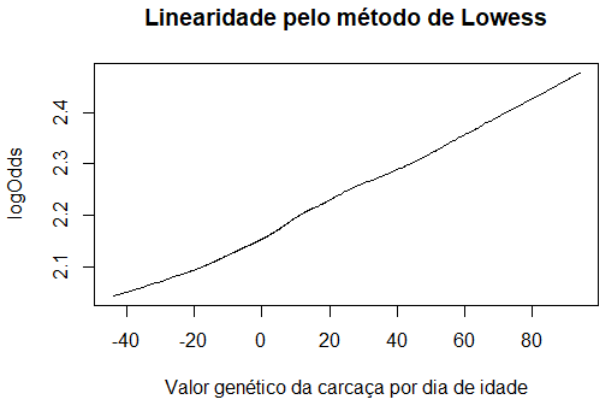


Figura C.4: Linearidade do valor genético da carcaça por dia de idade para a distribuição gaussiana e ligação identidade.

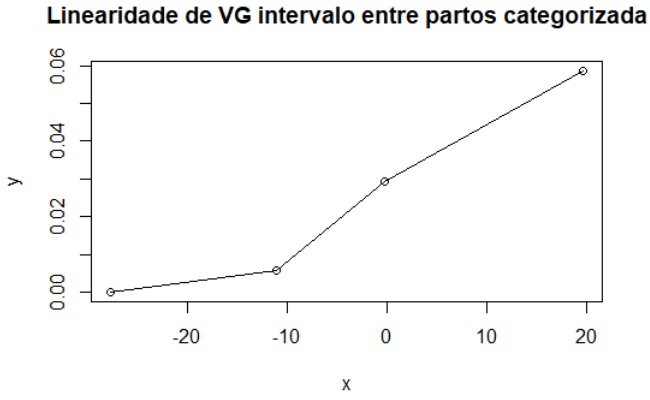


Figura C.5: Linearidade do valor genético intervalo entre partos categorizada para a distribuição gaussiana e ligação identidade.

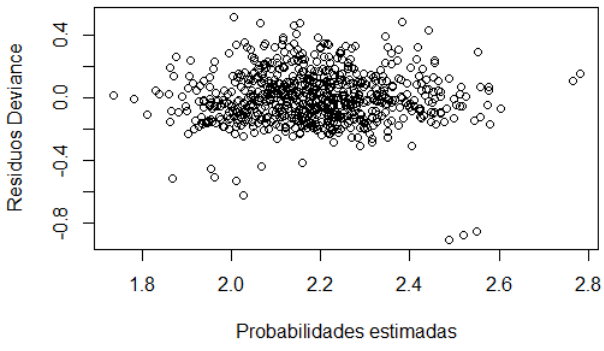


Figura C.6: Análise gráfica dos resíduos do GLM com distribuição gaussiana e ligação identidade.

```
temp<-influence.measures(MULTGAUS)
(lista <- which(apply(temp\$is.inf, 1, any)))
```

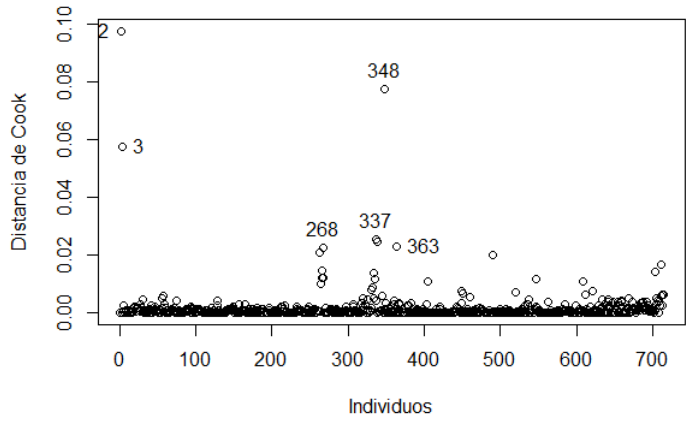


Figura C.7: Distância de cook dos resíduos do GLM com distribuição gaussiana e ligação identidade.

```
# lista as candidatas a observações influentes
summary(temp)
2, 3, 263, 264, 265, 266, 267, 268, 273, 274, 325, 331, 333, 336, 337, 338, 348,
351, 360, 361, 362, 363, 368, 403, 404, 446, 447, 459, 489, 550, 588, 612, 621,
626, 635, 703, 710, 711.
```

O modelo GLM apresentou, portanto, mais pontos influentes quando comparado com o modelo RLM, entretanto, nenhuma passou os limites nas análises gráficas. E também, pode-se perceber que o modelo se adequou bem a família gaussiana e a ligação identidade (tabela 5.6).

UNIVERSIDADE DE ÉVORA
INSTITUTO DE INVESTIGAÇÃO
E FORMAÇÃO AVANÇADA

Contactos:
Universidade de Évora
Instituto de Investigação e Formação Avançada — IIFA
Palácio do Vimioso | Largo Marquês de Marialva, Apart. 94
7002 - 554 Évora | Portugal
Tel: (+351) 266 706 581
Fax: (+351) 266 744 677
email: iifa@uevora.pt